Epistemic Representation Beyond Models: Thought Experiments, Specimens, and Pictures

Lorenzo Sartori

A thesis submitted for the degree of
Doctor of Philosophy

Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science

London (UK)
July 2024

Abstract

Scientists often make use of epistemic representations in order to perform investigations about the real world. So far, philosophers of science interested in epistemic representation of this sort have mostly focused on scientific models. In this thesis, I argue that there are other interesting instances of representation besides models: thought experiments, experimental organisms, and mechanically-produced pictures. These represent portions of the world in the same way as models do, if the concept of epistemic representation is properly understood. In Chapter 1, I introduce and develop the idea that a system functions as an epistemic representation of a designated target system insofar as it is interpreted as a symbol that both possesses and highlights some properties that are then to be imputed to that target system via a proper de-idealising function. In the main body of the thesis, I analyse the above-mentioned types of representations and argue that they function as representations in the same way as scientific models. Chapter 2 discusses the use of thought experiments in physics, with particular focus on Galileo's thought experiment that illustrates the principle of inertia. Chapter 3 focuses on experimental organism research in biology, with specific attention to the *Drosophila melanogaster*. Finally, Chapter 4 is about the epistemic use of mechanically produced pictures with the picture of a black hole as the primary case study. In each chapter, I further show that by studying these types of representation through the lens of the account developed in Chapter 1, one can (dis)solve specific issues about thought experiments, model organisms, and pictures respectively. In Chapter 5, I further defend my proposal from scepticism about the concept of representation and its application to these different cases. Chapter 6 concludes the thesis by illustrating the most general implications of my analysis and proposing routes for future enquiry.

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 99,549 words.

Kunst gibt nicht das Sichtbare wieder, sondern macht sichtbar.

(Art does not reflect the visible, rather it makes visible.)

— PAUL KLEE, 1920, p. 28

Contents

List of Figures

Acknowledgements

When people describe their own experience of pursuing a PhD, especially in philosophy, one of the most common themes is loneliness. For many reasons, writing this thesis has often felt like quite a solitary endeavour. However, this solitude is just one side of the coin, and focusing only on that aspect would be, *ça va sans dire*, an overt misrepresentation. Nothing of what I have achieved so far has benefited more from the intellectual dialogue and human interaction with other people than this dissertation. It is a long list of acknowledgements, but for once, I don't feel too guilty if I verbally expatiate: I have reached this point mostly thanks to all of you.

First and foremost, I am immensely grateful to my supervisors. All these years, Roman Frigg has been a true model of how a supervisor should be. He has always been dedicated, enthusiastic, rigorous, generous, extremely patient, and fruitfully pragmatic – the last virtue being tremendously useful if you are doing a PhD, especially on highly theoretical and abstract subjects. In all honesty, I have never ended a conversation with him without immediately feeling more confident, intellectually stimulated, and full of energy, always with a smile on my face. Being supervised by Jonathan Birch has also been a true pleasure: his perspicacity and ability to see the big picture have crucially affected my work, and his ceaseless kindness and encouragements have definitely helped me survive the stress and fear characterising the PhD years. I am especially grateful to both of you for showing me a general method and spirit to do philosophy and research: precision, clarity, simplicity, depth, and critical attitude are only some of the virtues that I will try to bring with me from your teachings and example.

Many people at LSE have constantly showed me kindness and support, and they inspired me as both outstanding researchers and wonderful human beings. A special thought goes to Lewis Ross, Liam Kofi Bright, Ella Whiteley, Harriet Fagerberg, and Jingyi Wu, who have repeatedly listened to my existential questioning and nevertheless saw something valuable in me. They all helped me find a way to navigate life, both in its academic dimension and more generally. In this respect, a special, colossal thank you goes to Giacomo Giannini, a precious gem of friendship who has never lost the opportunity to show me his unconditional love by indefatigably making fun of me when I most deserved it.

Among the formative figures, it would be impossible to leave out my ex-supervisor

from Bologna, Raffaella Campaner, a unique exemplar of straightforward human kindness, who has never ceased to keep an eye on me and support me through my academic career, often by generously reprimanding me when I was excessively doubting myself.

The list of people who helped me find my own path in philosophy, and that affected the intellectual investigations of the last four years, would be really too long to be summarised here. The junctures of this philosophical journey are spread across time and space: from my very first steps in analytic philosophy under the unmatchable supervision of Achille Varzi – my first true philosophical "maestro" – my studies made me travel a lot besides my "home" universities for exchanges (Vienna, New York), summer schools (Vienna, again) and conferences – Erlangen, Lisbon, Buenos Aires, Malaga, Weimar, and Paris, as well as many quick trips to the CAMPoS seminars in Cambridge. In all these places, I met wonderful, enthusiastic scholars, who enriched me so much with their ideas and insights. Among the people encountered in these peregrinations of the body and the mind, a special thank you goes to Chiara d'Ambrosio, Anna Alexandrova, Hasok Chang, Dominic McIver Lopes, and Mauricio Suárez, who have always showed an intellectual respect that I definitely don't deserve.

In this technological era and particularly during the pandemic, my travels got a bit more ethereal, and of all places, I found on Zoom an amazing group of philosophers of science, who immensely contributed to the development of the ideas presented in this thesis: James Nguyen, Mike Stuart, Joe Roussos, Michal Hladky, Rawad El Skaf, and all the other participants to the Working Models reading group.

So far, I primarily focused on the academic and formative dimension of the journey. But if I had to choose a word that characterises these four years more accurately, that word would be "friendship". Going back to LSE, nothing would have made sense or worth without its amazing group of PhD students, true companions of adventures, thanks to whom I learnt some of the most important things about myself, life, and relations. When I first tried to write this section, I realised that each PhD student deserved their own paragraph. I trust that all of you know that even though I refer to you as a group, I love each of you in a unique, unrepeatable way. Among the large group of PhDs, a special mention cannot but go to Margherita, Fabian, Adam, Léa, Eva, Talita, Bele, Shira, Somayeh, Dmitri, Paloma, Sophie, and Cecily. I still don't know why I have been so lucky to have met you, my friends. *Vi voglio bene.*

Despite spending most of my time closed within the walls of the office on the fifth floor of the Lakatos Building, I have also encountered many wonderful people in London (and surroundings), who made this cold, expensive and frenetic city into what I can well call my home. Some amazing friends that I cannot avoid mentioning are Chiara, Didem and Robbie, Oshy, Vic, Tom, Lukas, Lorena, Julia, Sam, Facundo, Maria Teresa, Lara and all the "Max Pezzali" circle, Rosa and all the Cardozo's gang. Of course, I want to also give a huge hug to my flatmates and outstanding friends Michele, Carla (and Carlo) and Alan, who not only endured cohabitation with me

(which is already something to be proud of), but also gave me so much love in these years.

Since I left my little hometown in Italy, I have also collected friends all around Europe and the world, and some of them have never stopped being close to me, even if geographically distant. Of course, the list must start with my oldest and dearest friends Francis, Ale, Pier, Carra and Matilde. An open-hearted hug goes also to my old Karate friends, with whom I never lose the opportunity for a beer at the Irish pub in Pistoia. Then my thoughts fly to Trento, especially to my dear friends Ste, Gio, Fazion and Francesca. Next, I always keep with me my friends from Bologna, those beautiful, funny souls of Steve, Ale, Leti, Pile, Ire, Angelo, Zambo, and Tummi. Last but not least, I hold tight Matteo and all the "VIP" group from my LSE's Masters: my "buddy" Mason, Johanna, Jakob, Karl and Eoin.

Every one of us has also some old life mentors, who even when not physically present in our lives, have shaped them so deeply that will always be actively affecting the way in which think and behave. I have already dedicated a mention to some of these "maestri" in these acknowledgements, but I want to add a few more: Alberto Bacchi, unmatchable master of martial arts and truly practical philosophy; Andrea Vaccaro, my first and most exemplary model of Socratic intellectual; Marco Pancani, brilliant teacher of mathematics and connoisseur of both philosophy and poetry, who made me discover the beauty of logic and the philosophy of mathematics; Vincenzo di Giacomo, heroic teacher of history and philosophy; and Stefano Bindi, true lover of knowledge, who made me an incurable "continental" (despite all my attempts to become an analytic) by teaching me the love for literature.

It is now strange to say that from friends I move to family, given that the people mentioned so far are the family that I have chosen myself. Still, there is a biological family that couldn't really choose to have me around, but has nevertheless been there for me, always. There are no words to thank my siblings Andrea and Camilla and my parents Cesare and Cristina, or to express all the love I have for them, even though I could call them more often, admittedly. You have heroically been by my side and loved me when I most needed you and beyond. Non ci può essere una famiglia disfunzionale, diffusa, e come direbbe Andri, semplicemente strana, migliore di voi.

Preface

This PhD thesis investigates the philosophical issues regarding surrogative reasoning in the sciences through the conceptual lens provided by the concept of epistemic representation.

After a first presentation of the philosophical questions regarding representation, Chapter 1 introduces and further develops and clarifies the details of the so-called DEKI account of epistemic representation, which I will then originally apply to new instances of surrogative reasoning in the rest of the thesis. In this Chapter, I also clarify the relation between DEKI and its philosophical "ancestors", namely the work done on representation by Nelson Goodman and Catherine Elgin. Throughout the Chapter, while selecting only the relevant information for the purpose of the arguments unfolding later in the thesis, I also develop the original ideas of DEKI's authors and raise further questions and new directions to develop it. Finally, I specify what distinguishes DEKI from other alternative accounts put forward by the participants to the debate.

Chapter 2 is a polished and extended version of an article published in *Synthese* (Sartori 2023).[1] The Chapter focuses on thought experiments and their epistemic role in science. In this Chapter, I reconstruct the current philosophical debate on scientific thought experiments around the question on whether thought experiments can and do provide new knowledge about the empirical world. I argue that the debate started on the wrong foot and still lacks a shared conceptual framework. I provide such a common ground by emphasising the methodological distinction between internal and external validity of an experiment. Furthermore, I provide an analysis of both types of validity in the context of thought experiments: internal validity is conceptualised in terms of games of make-believe, while external validity is defined as accurate representation of a thought experiment relative to a designated target system. Finally, I show that this diarchic characterisation allows us to escape the initial impasses of the current debate and that it provides a neat understanding of the epistemic role of thought experiments in science.

Chapter 3 is a developed version of an article that has been accepted for publication by *The British Journal for the Philosophy of Science* in October 2023 (Sartori in press).

[1] Reproduced with permission from Springer Nature.

This Chapter addresses questions about a further representational medium, namely model organisms as systems for surrogative reasoning in biology. The main question is whether these organisms function like scientific models by representing their target systems. I develop an already existing representational account of model organisms proposed by Ankeny and Leonelli (2020) and provide details on how model organisms function as epistemic representations of other organisms in the sense expressed by the DEKI account. Specifically, I emphasise the necessary interpretive activity involved in the inferences from model organisms to their target system and illustrate what kind of de-idealisation functions are at play when we export results obtained from a model organism to a different form of life. The second part of the Chapter is devoted to criticising two non-representational accounts of model organisms. First, I critically assess Currie and Levy's (2015) account of model organisms and show that it remains wanting. Then, I take into consideration Weber's (2004) view on model organisms and argue that it is in fact not in conflict with my account, and the two perspectives can further enrich each other in a positive synergy.

A concise version of Chapter 4 has been selected for publication in a special issue of the journal *Philosophy of Science* together with a subset of the PSA 2024 Contributed Papers. This Chapter tackles visual representation and specifically mechanically-produced pictures and their scientific use. First, assuming Meynell's (2013) account as a critical point of reference, I show that an approach to representation based on similarity, such as hers, does not seem able to provide a plausible account of the semantic and epistemic features of scientific pictures. To vividly illustrate my criticism, I focus on the case study of the picture of the black hole M87* recently produced by the Harvard Black Hole Initiative. Then, I analyse that picture with the help of the DEKI account's conceptual apparatus and illustrate its most relevant semantic and epistemic aspects. I further suggest that the justificatory strategy typical of mechanically produced pictures differs from the ones typically employed with scientific models. In the former case, both the interpretation of the picture and its justification crucially depend on the history and mode of production of the representation system, and thus with the causal processes relating it to the original target system. This, I suggest, is not the case with scientific models, where justification for our inferences lie completely outside the single representational system and depends on direct tests on the target or, in the lack of that, on theoretical or empirical support of our assumptions and results. I thus draw a distinction between measurement representations and model representations, distinct in the justificatory strategy we employ to support our inferences from them to the target system.

Chapter 5 addresses the most common attacks to a representational understanding of models and of surrogative reasoning in general. I divide this anti-representationalist camp in three main families of views: an antirealist, success-first understanding of models (Isaac 2013); the artifactualist views of models, with specific focus on the accounts proposed by Knuuttila (2011, **Knuuttila:2021**) and Sanches de Oliveira

(2021, 2022); and the pragmatist-inferentialist views, among which I specifically concentrate on a recent proposal put forward by Khalifa, Millson and Risjord (2022). For all three views, I both show that DEKI is immune to their criticisms and argue that their positive non-representational alternatives do not survive major objections. This is done by recourse to the results obtained in the previous chapters and substantial analysis of the details of the main non-representationalist arguments.

Finally, Chapter 6 draws a few general reflections on epistemic representation as conceptualised in this work and lays down promising routes for future philosophical investigation.

Overall, this thesis should be read as a monograph on representation that originated from a collection of articles. Because of the general bottom-up approach that characterised its production, most of the chapters can also be read autonomously. Particularly, Chapters 2, 3 and 4 apply the notion of representation to specific instances of systems employed for surrogative reasoning in the sciences – thought experiments, model organisms, and pictures. While scientific representation has been a fundamental topic in the philosophy of science for at least the last five decades, little has been done to apply the general philosophical accounts of representation to instances of surrogative reasoning beyond models. These three applicative chapters then constitute an original contribution to the current debate in that they fill this lacuna.

A few notable results follow from this endeavour. First, I show how thought experiments, experimental specimens such as model organisms, and mechanically produced pictures can all be philosophically understood as epistemic representations in the same sense that models are. Second, I individuate the deep similarities and differences between these different forms of representational practices. Third, being more local enquiries, these chapters also provide the inductive basis for more general reflections on representation overall. Fourth, and most importantly, I show how my representational analysis provides an answer to specific issues originating within the specialised literature about thought experiments, model organisms, and scientific pictures respectively.

While these central chapters can be mostly understood without reference to each other, they mutually support each other's conclusions and insights, and along my argumentation, I will try to draw connections and parallels between them. Additionally, even though they have been devised and developed as self-sufficient streams of enquiry, their reading would be clarified and considerably improved by Chapter 1, which sets the ground on many details of the general account of representation that I endorse, develop, and defend throughout the thesis. Similarly, Chapter 5 and its attack to the anti-representationalist camp would lose much of its argumentative strength if it were not supported by the "inductive basis" of the earlier chapters, as well by the general conceptual framework developed throughout the entire thesis. While part of my criticism against the various anti-representationalist views simply arises from

the individuation of serious issues in their own arguments, most of the problems of these accounts of surrogative reasoning depend on an inaccurate understanding both of representation as a philosophical concept and of scientific practices of surrogative reasoning – not only models but also thought experiments, experimental specimens, and pictures. The work carried out in the previous chapters will then allow me to clarify this misunderstanding and support my argument with a generalised account of epistemic representation in science, well-informed by a rich and detailed analysis of more specific case-studies.

Chapter 1

Point of departure: models and epistemic representation

1.1 Epistemic representation in science

The central topic of this work is epistemic representation in science and the use of surrogate systems in order to produce descriptive, predictive, or explanatory hypotheses about some portion of the world these surrogate systems are meant to represent. In the philosophical jargon, this is usually called "surrogative reasoning" – a term coined by Swoyer (1991) in his seminal paper on structural representation in physics.

While the discussion of epistemic representation in contemporary analytic philosophy of science is relatively young in comparison with other topics, such as prediction, causation, and realism, the theoretical roots of the debate can be detected in earlier work in the philosophy of language (Goodman 1976), aesthetics and image theory, as well as the philosophical study of analogical reasoning and metaphors in the sciences (Black 1962, Hesse 1963).

However, interest of philosophers of science for representation, at least in the terms of the current debate, originated in the study of models, their epistemic role as tools for investigating phenomena, and their relationship to scientific theories. This in turn was driven by attempts to move away from what is now called a syntactic view of scientific theories and towards what is famously known as the semantic view, which defined theories as families of models, understood as mathematical structures.[2] In this way, the philosophical discussion of models took a life of its own, separate from the model-theoretical framework in which it originated.[3]

However, this view of theories assumes a concept of a model as a mathematical

[2]The Model-Theoretical view of scientific theories was originally proposed by Suppes (1960, 1967, 1970).

[3]The literature on scientific models is now vast and cannot be summarised here. For a broad overview, the reader can consult the relevant Stanford Encyclopedia entry (Frigg and Hartmann 2018) and the literature referenced therein.

structure, which seems an inadequate description of a number of paradigm examples of scientific models,[4] such as Bohr's model of the hydrogen atom, Watson and Crick's model of DNA, the Phillips-Newlyn model of a national economy, or the Lotka-Volterra model of a prey-predator population system. What is noteworthy about each of these models is that, while they are all objects, with their own properties, structure, and internal dynamics, we usually do not focus just on what is true about them: we use them and what we know about them to make hypotheses about the portions of the real world of which these models are meant to be a representation. Bohr's model "represents" hydrogen atoms, Watson and Crick's model represents DNA, and so on. From now on, I will call the model the *representation system*, or simply representation, and the portion of reality that the model represents the *target system*, or target for short.

Once the (at least partial) independence of models from theories was acknowledged, it became ever clearer that a philosophical analysis of the specific nature of the relation between models and the world was required. Now, all participants in the debate would of course agree that representation is only one of the many purposes that models can serve, and in the course of the thesis I will show how these different function can interact with each other. However, the present Chapter will focus on a specific use of models, namely as representations of portions of the world.

In order to clearly define the scope of the discussion, it will first be necessary to distinguish the senses of both "representation" and "models" which will constitute the object of our enquiry. Let us begin with representation. Some opt for the expression *scientific* representation in order to distinguish scientific models from, say, the aesthetic representational functions which we encounter in the arts. However, it is not necessarily the case that the type of epistemic representation with which we are concerned here is confined to the scientific realm. *Prima facie*, at least, some works of art do seem to have an epistemically representational function in the general sense intended above. From the caricatures of public figures we find in newspapers, to historical novels narrating events that actually occurred, to portraits of people who in fact existed in the past, it is fair to concede that artistic representations do at least sometimes aim to represent something in the world. In the terms of this debate, such artworks not only refer to some target systems in the world (in a very minimal sense of "this-stands-for-that"), but they also prompt us to reason about these target systems: caricatures invite us to make inferences about the person being portrayed, and historical novels in some cases allow us better to know and understand the past.[5] Given that we are not for now interested in the reliability or truthfulness of these representations so much as in their representational aims,[6] it is worth taking

[4] See Frigg (2022, pp. 153-184) and references therein for a discussion.

[5] Art seems to allow much more than that, of course. The reader can find interesting insights on the epistemic value of art in Davies (2007), Young (2003), Neill and Ridley (1995), and Gadamer (1960/2013).

[6] The explicit formulation of this distinction as a shared assumption for the current philosophical

artistic representations into consideration, and not excluding them a priori from the realm of epistemic representation as more broadly understood, that is, as surrogate systems which allow us to make inferences and formulate hypotheses about real target systems. As such, I will henceforth adopt the more general expression "epistemic representation" to refer to all representations which share this aim.[7]

The epistemic sense of representation can be kept distinct from other senses of representation used in contemporary analytic philosophy, three of which are worth mentioning here. First, political theorists are interested in issues of *political representation*, which describes how the preferences of individual citizens are (or can fairly be) represented by their elected organs, and in the associated decision-making processes. Second, philosophers of mind wonder about the nature of *mental representations*, such as beliefs, desires, and imaginings, and their relationship with the external world. Third, in both formal epistemology and in economics, philosophers evoke *representation theorems* to express the relation between a certain preference ordering and the utility function used to represent it mathematically.

While there are of course shared semantic links between all of these senses of "representation" and the epistemic sense of the word on which I will focus, I want to suggest that none of the above three senses necessarily imply the epistemic use of a surrogate system to investigate a portion of the real world. A political body, for example, represents a certain portion of a population by expressing that population's interests and preferences in the best way it can, but it would be conceptually incorrect to interpret the body as being an *epistemic representation* of the populace, that is, that the political body is used to draw inferences about the represented population. Of course, there is nothing inherently preventing such inferences: sometimes we may learn something about the electorate by looking at its political representation. But the concept of political representation does not require such reasoning *per se*.

A mental representation, on the other hand, is often taken to be *about* something, often a portion of reality,[8] and in this it retains a referential relation analogous to cases of epistemic representation.[9] And one could in principle study the mental

discussion can be traced back to Suárez (2010). As we will see, this distinction between a representation and that representation's accuracy or reliability may not be welcomed by some participants in the debate – cf. Chapter 5. However, it is largely agreed by philosophers that a relationship of representation does not in itself entail representational accuracy or success of any sort.

[7]Some may be suspicious that this move represents a shifting of the definition of representation to make it more convenient for my own account. In fact, the absence of a strong demarcation between scientific representations and other types of surrogative reasoning will not play the role of an assumption, but rather a consequence of the account of representation I endorse. For now, focusing on epistemic representation instead of simply on scientific representation should be taken simply as an *ex-hypothesi* statement, which will be confirmed by what follows.

[8]The contemporary notion of the *intentionality* of mental representations can be tracked back to the work of Franz Brentano (1874) – for a discussion, see Jacquette (2004).

[9]The literature on mental representation is too vast to be summarised here, but the interested reader can find a good introduction to these topics in Pitt (2022). Mental representation has also been theorised in terms of modelling. For an entry point to the topic of mental models, see Held *et al.* (2013).

representations of certain portions of the world in order to investigate those target systems, in a manner analogous to epistemic representations such as models and pictures. However, we do not generally study people's mental representations in order to acquire information about the target systems of those representations, and when we do, we are usually interested in some form of conceptual analysis – seeking knowledge about our concepts, beliefs, and attitudes regarding the world, rather than about the world independent of those attitudes. The difference is not simply one of private vs. public representations: some mental representations can be and are in fact shared by large communities. What seem to distinguish the two groups of representations are their different semantic and epistemic features. With mental representations, we seem to deal with direct conceptualisation of portions of the world, without the use of surrogative systems that characterise epistemic representation.

This line of thought opens a full set of epistemological and metaphysical problems about the mind and its relationship to the world. Particularly, many of these issues strongly depend on one's views about scientific realism. For example, one may want to endorse some more or less radical form of constructivism or idealism about scientific knowledge. In that case, most of what we call "reality" would still be some form of mental representation.

I do not intend to take a stance on these issues here but simply develop an account of epistemic representation that is compatible with different philosophical positions about both the mind and scientific realism. What seems uncontroversial in any case, is that the degree to which mental representations inside the brain resemble the sort of representations constructed in science or art remains an open question.

Nevertheless, any meaningful comparison will require a detailed account of both phenomena. My analysis here, then, will be helpful in that it clarifies what it takes for something to be an epistemic representation. This will in turn be useful for any enquiry about the analogy between epistemic and mental representations. Simply, what the reader should *not* expect from my analysis is an account of representation that also applies to the mental realm.

Finally, representation theorems are simply mathematical descriptions of an order of preferences, understood in a very minimal sense of option A has a greater value than option B. In this sense, representation theorems are mathematical functions where variables are interpreted as alternative lines of action. Here, there is no surrogate system involved in making inferences about preference ordering: the "representation" is just a specific way of describing, defining, or spelling out those preferences in a mathematical language. Certainly, the resulting mathematical function can then be interpreted as an epistemic representation of allegedly actual preferences of real people and used to model them. However, this involves the further step of taking a mathematical description of a certain preference ordering to *also* represent an actual preference ordering in people's minds. Unless this further representational step is taken, the function may not represent anything in the world: it would just be an

interpretation of an abstract mathematical structure as a set of preferences. We will see shortly that scientific models often carry out a similar, purely theoretical function. However, I will argue that this case is a use of models distinct from their use as epistemic representations.

In what follows, beyond a few clarifications on mental representation as it relates to scientific thought experiments in Chapter 2, we will not be further concerned with these political, mental, and decision-theoretical senses of representation, and the reader should not expect my treatment of representation to be applicable to them. Nonetheless, some of the conclusions drawn regarding the concept of epistemic representation may turn out to be helpful in understanding certain aspects of these other senses of representation.

Let us move on to clarify what we mean by "models". We are not, or not principally, interested in clearly non-scientific types of models. For example, a "model" can designate a person who wears clothes in front of cameras. Someone can also be a model in the sense that they inspire similar behaviour in other people. It is true that this latter sense of model has also found a scientific application in the context of cultural evolution, where some people are "models" for particular skills. While again there is some semantic connection with the sense of "model" in which I am interested, the reader should not expect my reflections to hinge upon these specific meanings of "model", nor to enlighten us about their use in everyday language.[10]

In a scientific context, it may be useful to distinguish three senses of "model": models of phenomena, models of data, and models of theories (Frigg and Hartmann 2018). Models of phenomena are epistemic representations of portions of reality (phenomena) and stand in for those phenomena for certain epistemic purposes. As I have already anticipated above, this means that models allow for surrogative reasoning: we are able to produce inferences (sometimes, even correct ones) about the designated target system on the basis of the observation, manipulation, and development of the model. Watson and Crick's double-helix model of DNA, Bohr's model of the hydrogen atom, Newton's model of the Solar System, and the Lotka-Volterra model of prey-predator population system are all examples of models of phenomena.

In what follows, I use the expressions "portions of the world", "portions of reality", "states of affairs", and "phenomena" interchangeably to indicate the set of things that are the target of a given representation. In so doing, I do not imply that models are not also part of the real world: they are, but they also function as representations. As such, they have a symbolic function that the target systems do not necessarily have.

It is also important to notice that the target systems can be sorts of representa-

[10] However, the second sense of model mentioned, understood in the very general sense of a system that is used to make normative inferences (what one ought to or should do), has important affinities with the concept of an epistemic model which is under investigation here. On this point, cf. sections 3.4.1 and 3.4.2.

tions in a loose sense. By saying this, I want to clarify that I am not making any realist assumption as to the target systems of our epistemic representations. The target system of a representation may be more or less theory-laden and partially a product of our conceptual schema and frameworks. Further reflections on this topic can be found in section 1.4. For the moment, what matters is that, in order to have a meaningful discourse about a representation relation, we need a first conceptual distinction between that which represents and that which is represented, and to express this relation so that it is asymmetrical: a model represents its target and this does not imply that the target also represents its model. Neither the distinction between representation and target, nor the directionality of representation, should be controversial, as they are compatible with any position in the realism/antirealism debate, where all participants generally agree that there is at the very least a distinction to be made, in any specific context, between a representation (a model, a picture, a map) and what that representation represents (a national economy, a supernova, a territory). Surely, then, nothing rules out a target system that is in turn a mental or social construct. In such a case, however, there will still be a distinction to be made between the representation and its target system, as well as a representational direction which flows from the former to the latter and not vice versa, plus the generally agreed fact that we tend to use representations as epistemic surrogates of their targets.

Furthermore, what matters here is not whether a system actually represents an existing target, but whether it is that kind of thing that should be understood to represent an existing target.[11] Drawings of unicorns and centaurs, ballets representing evil wizards and princesses, and crime novels with fictional detectives are all representations, but they seem to fail to represent anything existing in the world. Similarly, scientists created in the past models of the ether, of the caloric fluid, and of Galen's humours. Even if it was subsequently discovered that these things did not exist, their models are still representations: they have to be understood as representing something. It is just that those intended target systems do not exist. So, the fact that the target actually exists or not does not in principle change the representational function of a system. What changes is what we can achieve with that system. An account of representation should be able to clarify in what sense we still talk about representations in these cases, where the intended target does not exist.

The representational function of models of phenomena can be conceptually distinguished from two other important functions that models can serve, namely to the modelling of data or the instantiation of a theory. A model of data is "a corrected, rectified, regimented, and in many instances idealized version of the data we gain from immediate observation, the so-called raw data" (Frigg and Hartmann 2018). As Bogen and Woodward (1988) have argued, phenomena cannot be reduced to data, and paradigm examples of scientific models should be understood as representations

[11] This point is made by Elgin (2010, p. 2).

of the former and not the latter (even though they can be tested against models of data for confirmatory purposes). This, as Bogen and Woodward point out, is clear from the fact that data are finite, spatially and temporally located, and contextual observations of a more general and complex phenomenon, which is the actual target of our representation and of our surrogative inferences. Accordingly, Watson and Crick's model is meant to represent the DNA, and not the individual data points obtained by our punctual, spatially and temporally located measurements of DNA properties. Data about DNA may of course contribute to the description of the phenomenon, but the latter doesn't conceptually reduce to the former.

This thought also applies to other instances of epistemic representation: the picture of the black hole M87*, about which I will talk about in Chapter 4, is a representation of the black hole in question, even though it is constructed also on the basis of interferometric data collected by the telescopes. These raw data are then interpolated, organised, and simplified in models of data that are of course part of the process of creating the picture. But the representational relation still goes from the picture to the black hole itself, not to the data. Of course, nothing rules out a model being both a representation of a certain phenomenon and a systematisation of a number of data points about that phenomenon.

Finally, a model of a theory Θ is any tuple (O, R) – where O is a set of objects and R a set of relations among those objects – that makes Θ true about that structure. The object and relations can be defined in purely mathematical terms, or they can be endowed with some interpretation. However, they are not representations in the same sense as models of phenomena, unless the structure is also endowed with a physical interpretation that makes it a representation of some phenomenon in the world.

As an example, consider Newton's mathematical model of the Solar System. It is a representation of a phenomenon, namely the actual Sun and planets orbiting around it. But it is also a model of a theory: classical mechanics. This is the case because the mathematical structure of the model makes the axioms (and the derived theorems) of classical mechanics true.

While in the past philosophers of science thought that models represented phenomena in virtue of also being models of a theory, such as in the case of Newton's model, this is often not the case. Models such as the Lotka-Volterra model, which are paradigmatically models of phenomena (in this case, of prey-predator population systems) were born without any overarching general theory supporting them On the other hand, one can find models that are not intended to represent anything in the world, but only to instantiate the basic tenets of a theory. As a paradigm examples, consider Weisberg's (2013, § 7) models of four-sex organism populations. In Chapter 2, we will see how this also occurs in the case of scientific thought experiments, particularly with the case of Galileo's famous thought experiment of falling bodies, aiming at illustrating the inconsistency of Aristotle's theory of fall.

The meaning of "model of a theory" is thus conceptually separable from the concept of a model as representation of phenomena, which is my primary focus here. Of course, a model of a theory can be used to make inferences and support reasoning about that theory: for example, it can be used to illustrate and clarify certain tenets of the theory, or even show contradictions in or problems with that theory. However, such a theory will not be taken here to be here as a portion of reality (that is, a system that has to be represented), but rather as a description of certain aspects of reality.

The two uses of a model as a model of phenomena and a model of a theory, respectively, can interact in interesting ways, and later in the work we will see how this can occur. Nevertheless, my focus remains on the former use.

As I anticipated above, when I say that the function of models is to represent, I am not saying that it is their *only* function. Because they can be models of theories, they can also serve intra-theoretical purposes: they can be used to test, develop, amend, and possibly reject these theories. Moreover, by representing both phenomena and theories, they can be used as inter-theoretical tools: for instance, by demonstrating the compatibility of two theories, showing the superiority of one theory over another, or offering a basis on which to build new theories from scratch (see e.g. Hartmann 1995).

Finally, there is a further function of models that is not representational, namely when a model is employed to develop know-how (techniques, methodologies) that we can then import into other contexts. A paradigmatic case of the sort is provided by some uses of model organisms, to which I will return to in greater detail in section 3.4. There, I will also discuss the relation between this form of surrogative know-how inferences and representation, and how the former often fundamentally relies on the latter.

In practice, the representational function is not isolated from the rest of the functions of a model just mentioned: the fact that a model functions as an epistemic representation of a portion of reality is important in order to its use in theory building, or to generate new know-how and methodologies to be imported into new contexts. Still, these functions concern conceptually distinct practices and purposes, so they are independent of each other.

While most philosophers acknowledge the importance of the representational function of models, there are accounts that deny models a representational function altogether, or at least do not emphasise it relative to other functions. Some examples are Knuuttila (2011, **Knuuttila:2021**), Isaac (2013), Khalifa, Millson and Risjord (2022), and, concerning model organisms, Weber (2004). I will postpone a detailed discussion of Weber's concerns to section 3.4, and I will systematically evaluate the rest of these positions in Chapter 5. For now, let us clarify more precisely the specific view on representation that I endorse. This specification will also prepare the ground for my later analysis of thought experiments, experimental specimens and pictures in

science, as well as for my answer to the above anti-representationalists about models.

In order to introduce the account that I endorse in this thesis, it will be helpful to place it within the various accounts of scientific representation put forward in the recent years. These accounts can be organised in at least six great families: Griceanism, similarity accounts, structuralist accounts, inferentialist accounts, direct fictionalism, and representation-as accounts. Let us briefly look at them one by one.

Gricean views, such as the one expressed by Callender and Cohen (2006) entail taking a minimalist position on representation, arguing that it is usually just a matter of stipulation and convention whether one system represents another.

Accounts based on similarity (see Giere 2004, 2010; Weisberg 2013, Chapter 8) take similarity between two systems, in the sense of that sharing of properties, as the fundamental grounds for epistemic representation.

Structuralism, elaborated for example by Da Costa and French (1990, 2000) and French and Ladyman (1999), chooses to focus on the sharing of one specific type of properties, namely the mathematical structures of the two systems,[12] such that the ground for representation is identified with a specific mathematical morphism between the mathematical structures in question (with isomorphism, partial isomorphism, and homomorphism being the most common candidates).

Inferentialist accounts take the surrogative reasoning condition above as a fundamental bedrock the concept of scientific representation, meaning that the possibility of using a system to draw inferences about another system is what makes it a representation. This family of views is diverse, ranging from more deflationary accounts like that of Suárez' (2024), to accounts that inflate the inferentialist account with some form of interpretive activity (Contessa 2007, Díez 2020).

Direct fictionalism, such as the one put forward by Toon (2012) and Levy (2012, 2015) take models to be fictional direct descriptions of their target systems, connecting the debate about representation to a broader philosophical analysis of fictional entities.

Finally, representation-as accounts conceptualise representation not as a two-place relation, but as a three-place one: the general intuition here is that a model represents its target *as* something else, where this something else is not an object but rather a type of representation. Since the account I endorse, the DEKI account, belongs to this family of accounts and builds upon contributions by Goodman (1976) and Elgin (2010), I will postpone a more detailed discussion of it to the next section.

Reviewing each individual contribution to this debate would be a book-length enterprise and cannot be carried out within the scope of the present work. The reader can find a good summary and critical evaluation of these accounts, together with relevant bibliographical references, in Frigg and Nguyen (2020). In the following sections, I will illustrate the DEKI account of epistemic representation. DEKI stands for "denotation, exemplification, keying-up, and imputation", and has been proposed

[12] Several different notions of structure are discussed in the literature. For a review, see Thomson-Jones (2011).

as an account of epistemic representation by Frigg and Nguyen (2020). In the rest of this Chapter, I illustrate in greater detail the general tenets of the DEKI account using paradigm examples of scientific models. I will further clarify the interaction between the four constituents of representation and develop some new implications and features of the account.

1.2 Representation-of, representation-as

Watson and Crick's model is a representation of the DNA, Bohr's model is a representation of hydrogen atoms, and the Philips-Newlyn machine is a representation of a national economy. In all of these examples, *representation-of* is meant to express a two-place referential relation: the model is "about" a target system, meaning that the model refers to the target in the same way a symbol refers to the thing that the symbol stands for. Following Goodman (1976) and particularly the later developments of his theory provided by Elgin (1983, p. 19 and ff.), the DEKI account recognises the relation of representation-of as a denotative one, i.e. the relation connecting a name to its bearer, or a term to its class.[13] In this Chapter and in the work as a whole, I will follow Elgin (1983) in using "reference" as a very broad concept, generally indicating any relation between a symbol and an object or between a symbol and another symbol. Denotation is a special case of reference, namely a referential relation of a symbol to an object.[14]

As can be seen, Goodman and Elgin's concept of denotation is quite unorthodox, even "slightly tendentious" (Elgin 2010, p. 2), in that it applies not only to names but also to non-linguistic objects, such as material models, maps, and pictures. This amounts to an intuition that the core of the relation between, say, a portrait and its subject is the same as the relation between a name and its bearer.

It is important to clarify from the start, though, that this claim, shared by Goodman and Elgin and imported in the DEKI account, does not amount to saying that epistemic representation *is*, or *can be reduced* to, denotation. We will see in a moment that denotation alone is not sufficient to explain how we use models as sources for epistemic surrogative reasoning. The point is just that denotation seems necessarily involved when it comes to a system representing another. For the present purposes, the intuition is fairly reasonable: the referential relation between a symbol and what the symbol stands for seems to apply both to linguistic names and to

[13] Another author who acknowledges denotation as a basic ingredient of model representation is Hughes (1997).

[14] Later in the Chapter, we will encounter another type of referential relation, namely exemplification, which is the referential relation which flows from an object to a property – or, with a more nominalist flavour, from an object to a label to which that object complies. Because names are simply a type of labels in this sense, one can understand why Goodman (1976, pp. 52 and 66) talks about exemplification as the referential relation that goes in the "opposite direction" to that of denotation.

non-linguistic representations.[15]

Now, denotation can simply come to be by fiat or arbitrary stipulation, and this is one of the central complaints from Callender and Cohen (2006) concerning the philosophical works on scientific representation. There is not much new, they argue, about scientific representation: representation of models and pictures is just a matter of stipulation, with the former representing something else just because we decide it does. The problem, they suggest, lies elsewhere: in the nature of background mental representations, which are responsible for the epistemic use of a certain object as an epistemic surrogate for another.

I grant to Callender and Cohen, like most of the participants in the debate, that epistemic representation depends at least in part on some form of mental representations on the part of the users. Even among rival theories of representation, such as the similarity views (Giere 2010) or structuralism (van Fraassen 2008), it is well accepted that representation requires an agent with intentions and some epistemic purposes. Likewise, in the representation-as accounts like the one proposed by Goodman and Elgin, there is nothing in an object *per se* that makes it a symbol of something else; rather, it is some form of interpretation on the part of an epistemic community that makes that object a symbol of a target system. But the point here is exactly to understand how epistemic representations like models and pictures relate to mental ones. In other words, it is reasonable to ask how we characterise the interpretive activities required by our use of a system as an epistemic surrogate for another, and how the properties of certain representational systems combine with our mental activities in the production of such surrogative reasoning. These philosophical questions will be answered in more detail later, and my analysis will show that this task is worth the philosophical effort.

What is too radical, then, is Callender and Cohen's claim that representation is a *mere* matter of stipulation. This does not follow from their reductionism of epistemic representation to mental representation: even granting this stance, it does not then follow that epistemic representation is always a mere result of arbitrary stipulation. For stipulation does not seem to be enough to characterise the mental representations involved in epistemic representation. Watson and Crick's model, for example, does not simply denote the real DNA, in the sense that we limit ourselves to saying that one object stands for the other. The model represents the DNA *as* a double-helix structure constituted of two spiral chains of nucleotides, horizontally connected pairwise by couples of nitrogenous bases. In other words, the model gives epistemic us access to certain specific characteristics of the target on which we are

[15] A further complication is that, even among linguistic objects, denotation is usually restricted to naming and does not apply to predication. Elgin (1983, pp. 29-35) proposes a general extension of the concept of denotation and suggests that predicates, such as "(being) human", denote the members of their extensions (each individual human being), while the class of those members, if you believe in the existence of classes, is denoted by the corresponding abstract term (e.g. "humanity").

supposed to focus, in contrast to other possible features of the DNA. It is this intuition that motivates the rest of the theoretical development of this Chapter. Let us look at this idea more in detail.

Following Goodman's (1976, p. 27) lesson, I have already implicitly recognised that the term "representation" is ambiguous, and it could either mean representation-of or representation-as. Representation-of is a referential relation connecting a symbol with what the symbol stands for. Representation-as, or Z-representation, on the other hand, is not a relation, but rather it is a monadic predicate which indicates, in the terms employed by Elgin (2010, p. 3), the "genre" of a representation. In general terms, we can say that the Z is a type of representation, that is, a way to categorise a representation. A landscape painting, for example, is not necessarily a representation *of* a real landscape. It is simply a landscape-painting, and we can categorise it as such in our archives, essays, and museums holdings. A caricature of Winston Churchill as a bulldog, is a representation-of relation between the drawing and the target Winston Churchill, and it is also a bulldog-representation – there is no actual bulldog the caricature denotes.

In science, we usually do not have recognisable genres like those in the arts, but rather tend to have theoretical domains and fields of enquiry. In the example of the model of DNA, the theoretical domain could be described as molecular biology, so Watson and Crick's model would be a molecular-biology-representation of DNA. However, I would like to add to the reflections by Goodman, Elgin, and even the authors of DEKI the crucial fact that the level of details at which we categorise a representation can vary depending on the context. A landscape picture may be a seaside landscape picture, or a storm-at-the-seaside picture, or a storm-at-the-sea-side-with-human-figures picture, and so on. The degree of granularity and specificity at which we categorise depends on the purpose of the taxonomic system that we wish to employ. This flexibility should not surprise us, as it is analogous to the different levels of specificity one deals with when in categorising objects in the real world. There is nothing special, then, when it comes to classification systems of representations. And of course, taxonomic flexibility is also a feature of scientific instances of representation. We can be more granular and say that the Z-representation in the case of Watson and Crick's model is a double-helix structure of two chains of nucleotides, which are horizontally linked by bonds between pairs of nitrogenous bases. Or for short, we could say something along these lines: Watson and Crick's model is a double-helix-representation of the DNA.

Someone could now object that it is problematic even to recognise genres or theoretical descriptions (in one word, a type of a representation) without some form of implicit appeal to their alleged referents. How can I categorise, say, a picture as a unicorn-picture, without some reference to unicorns? And what do we do if, as in this example, the relevant referents do not exist? According to Goodman (1976) and Elgin (1983), we learn this by practice and habit, as we learn how to associate words

and predicates with their referents. As Elgin puts it, "this is no more mysterious than learning to recognize landscape[-painting]s without comparing them to the terrain they ostensibly depict" (2010, p. 3). In other words, we constantly taxonomise and organise representations by referring to their properties and to the fields of enquiry that devise and deploy them, and we learn to do this independently of any antecedent classifications of alleged referents of those representations.

The case of *scientific* Z-representations seems even easier to tackle, given the high levels of regimentation and standardisation of scientific practice. There, the Zs are produced within the frameworks of scientific theories and programmes of investigation, and are developed within theoretical domains delineated and shaped by the epistemic community of reference, without any problematic reduction to actual referents.

This is, then, the distinction between representation-of and Z-representation, originated in Goodman and Elgin's work and inherited by the DEKI account. There is still an open question as to how a material object "becomes" a Z-representation in the first place. I will discuss Frigg and Nguyen's original answer to this question below, in section 1.5.

It should be noted upfront that representation-of and representation-as can and do come apart, but also that, for just this reason, they can be used in combination. A caricature of Margaret Thatcher as a boxer, for example, is a representation *of* Margaret Thatcher *as* a boxer. Or, alternatively, it is a boxer-representation of Margaret Thatcher. What is important is that a person looking at the caricature should not think that the author wanted to represent an actual boxer of any sort. There is no boxer in the real world that the picture denotes.[16] One of the main philosophical achievements of this disambiguation is that representation-as, or Z-representation, does not require denotation. This gives us a way to talk about instances of representation even when there is no denoted target system. A painting showing a chimera is not a representation *of* anything, for there is no such a thing as a chimera in the real world. Nevertheless, it is still a type of representation, namely a chimera-representation.[17] As we saw above, we are also usually able to identify types of representations without necessarily requiring a previous classification of the alleged referents of those representations. We learn what unicorns are because we have heard and read stories about unicorns and seen paintings and drawings of them. Indeed, nobody ever saw a unicorn (*ex hypothesi*), but many people would be able

[16] This is the case even though, as we will see in a moment, there is some referential relation between the caricature and some of the properties normally associated with boxers. This form of reference is not denotation, because it does not flow from a symbol to an object. So, we need another type of referential mode. This is what Goodman and Elgin call exemplification and I will illustrate this concept below, as it will be fundamental to what follows.

[17] This puzzle of non-existing objects is a *topos* of the traditional debate in ontology and gained significant traction from the apparent ability of humans to mentally represent non-existent objects. The literature is extensive and cannot be covered here for reasons of space. For an entry point in this literature, see Crane (2013).

to draw, paint, or describe them. This then should not create too many problems if we want our ontology to remain parsimonious with respect to fictional objects.

At this point, one may be worried about the meaning of terms like unicorns and chimeras. If meaning is related to denotation, as it is sometimes assumed, then all fictional terms would have the same meaning, because they all refer to the empty set – that is, to nothing. In her general reconstruction of a Goodmanian theory of reference, Elgin (1983, pp. 43-50) answers to this problem, showing more precisely how we can talk about Zs without making any ontological commitment to cumbersome entities, while at the same time distinguishing the meanings of different terms allegedly referring to fictional objects. We can take fictional terms to refer to a relevant set of descriptions and representations. For example, the term "unicorn" does not refer to actual unicorns, but simply to texts describing unicorns and pictures depicting unicorns – or, more rigorously, to unicorn-descriptions and unicorn-depictions. Also, each unicorn-picture refers to itself and to all of the other members of the class of unicorn-descriptions and unicorn-pictures.

Of course, I acknowledge that deep philosophical discussion concerning non-existent objects may arise here, and there is no space here to do justice to the complexity of such a debate, with all its implications for ontology and for the philosophy of language and mind. The present work, then, simply takes Elgin's development of Goodman's theories as a working theory of reference and of the dual semantic of the term "representation", in order to sheds useful light on crucial features of epistemic representation more generally.

In fact, Goodman and Elgin's strategy seems advantageous in that it allows us to obtain a purely extensional account of reference, whereby we do not (automatically) grant the existence of objects that are ontologically suspicious or at least controversial. Importantly, we can then draw a distinction between what we take to exist and objects that, by assumption, are not part of reality in the same way, such as unicorns and chimeras.

This strategy is of course very useful when we leave the domain of artistic fictions and move to the scientific realm, which also brims with fictional objects such as frictionless planes in physics, immortal rabbits (as in Fibonacci's model of population growth), and the perfectly rational agents of economic models. As Thomson-Jones effectively puts it, in such cases we apparently have "descriptions of a missing system" (Thomson-Jones 2010, p. 284), which are however embedded in what he calls the "face value practice" of science (*ibid.*, p. 285), i.e. the practice of discussing these systems *as if* they were real. With the general representation-as framework I outline above, we can have a first stab at solving this problem. Even if have descriptions of abstract models, these descriptions, in fact, denote actual target system, and epistemically represent them *as* fictional, idealised systems. Thomson-Jones' missing systems, then, are understood as our Zs, i.e., types of representations, or theoretical domains, rather than as existing objects. This allows us to make sense of the "as if" operator that

constitutes the face value practice.[18]

We can thus obtain a general three-place schema: a model system M is a Z-representation of a target system T. This schema applies to all representational models in science: Newton's model represents the solar system as a system of perfect spheres on which only the force of gravity is in act; the Phillips-Newlyn machine represents a national economy as an IS-LM open economy;[19] and the Lotka-Volterra model represents the interactions of populations of prey and predators as a fictional system in which, among peculiar features, prey can die only if killed by a predator, and populations are measured by real rather than integer values.[20]

Here, it is crucial to stress that it is not sufficient a mere conjunction of denotation and that the model is a certain Z-representation. The two things have to work together. As a clear illustration of this issue, consider Elgin's example:

> We could take any [Z]-representation and stipulate that it represents any object. We might, for example, point to a tree-picture and stipulate that it denotes the philosophy department. But our arbitrary stipulation does not bring it about that the tree-representation represents the philosophy department *as a tree*. (2010, p. 4, my emphasis)

For example, the tree-picture may denote the philosophy department in a merely topological sense. Imagine that I have to give you indications of how to find the philosophy department in the university campus, and I want to use the different paintings hanging on the wall in front of us to make myself clear. Then I will say something of this sort: if the wall in front of us is (i.e., denotes by stipulation) the entire campus, and the chair down there is the library, and that other picture (say, a portrait) is the psychology department... then the tree-picture is the philosophy department. In this way, if you know where the library and the psychology department are, I gave you useful information about how to reach your desired destination. In this example, the tree-picture on the wall denotes the department by stipulation, and it is a tree-picture. But it cannot yet be said that it represents the department *as a tree*. That is, the fact that the picture is a tree-picture does not play any role in the representation of the department in terms of its location with respect to the other campus' buildings.

In order to be a tree-representation of the department, something else is needed, namely that some of the properties of the painting *qua* tree-representation are attributed to the department. *Contra* Callender and Cohen, then, stipulation in itself does not seem to suffice here: something more must be said on how a representation

[18] The concept of "as if" has deep philosophical roots in the Kantian tradition, particularly Vaihinger's monumental work *Die Philosophie des Als Ob* in 1911 (for an English translation, see Vaihinger 1924).

[19] "IS" stands for "investment-savings"; and "LM" for "liquidity preference-money supply" (Barr 2000, p. 103). For a discussion of this model see Begg *et al.* (2014, Chapter 20).

[20] Volterra (1928, p. 6). For an analysis of this (family of) model(s), see Weisberg and Reisman (2008).

of a target also represents it *as* a certain Z.[21] This question will be answered in the next section on exemplification.

1.3 Exemplification

The previous section states that a model represents a target by virtue of denoting it and at the same time being a Z-representation of that target. At the same time, as we saw in the case of the tree-picture standing for the philosophy department, the two elements (denotation and Z-representation) do not suffice in and of themselves. They must work in combination. What sorts of thing guarantee that, say, a department is represented as a tree, besides being just a pure matter of stipulation this tree-picture and the department by mere fiat?

As Elgin (2010, p. 4) argues, similarity alone will not help us here, as similarity is ubiquitous and thus not selective enough. Any two objects will always be similar in some respect or other, if similarity is a mere sharing of properties: the picture of the tree and the philosophy department are both objects, they are both entertained by the mind of the reader at this moment, their referents both contain the letter "e", and so on. So, relying merely on the presence of some similarity would leave us in the absurd position that any object is always a Z-representation of any target system.

Let us then return to the example of Watson and Crick's model of DNA. We want to give an account of the intended epistemic dimension of M as a Z-representation of T. As an interpreted object, Watson and Crick's model allows an agent to recognise certain properties that would be possessed by actual DNA: the spiral structure, the bonds between pairs of nitrogenous bases, and so on. Similarly, Newton's model of the Solar System allows us to derive Kepler's three laws, why the Lotka-Volterra model implies that prey-predator systems exhibit the so-called Volterra property (Weisberg and Reisman 2008, p. 113) – namely, that when a biocide occurs in both populations, the relative amount of the prey population increases (Volterra 1926, p. 558).

We can make sense of this capacity of models to highlight certain properties by employing the concept of *exemplification*, which was coined in its technical sense by Goodman (1976, pp. 52-57) and developed by Elgin (1983 pp. 71-95, 1996, pp. 171-183). Samples are paradigm instances of exemplification. For example, when you enter a shop to choose your new curtains, you are shown a series of samples of cloth. These swatches instantiate an indefinite number of properties we are not

[21] Ruyant (2021) has proposed a revision of Cohen and Callender's account. One of the consequences of this revision is exactly the fact that Callender and Cohen's original views about mere stipulation are abandoned, and the relation between epistemic representations and mental representations is articulated in much more detailed than simple appeal to users' fiat. Particularly interesting is Ruyant's emphasis on interpretation, on social epistemic conventions and norms, and on the importance of the practical aims of our epistemic use of a surrogate system. In the end, the revised account ends up being significantly closer to less minimalist accounts of representation, and to DEKI in particular. As such, this account does not threaten my claims about denotation here.

interested in. For example, suppose all of them had been produced in China, with raw materials from Brazil, had been transported to the UK by boat, and are all now displayed in the same room. These properties are not the ones exemplified by the swatches, because in the context of the curtain shop, the swatches do not refer to those properties. The swatches are not to be intended to be exemplars of those properties, but rather exemplars of the colour they instantiate, the fabric, the material, and all the properties which are salient to us as we make our choice regarding type of curtains we want in our house.

More rigorously, exemplification is defined as follows: an object exemplifies a certain property A if and only if the object instantiates A and the object also refers to A. Instantiation, or possession, is thus a necessary condition for exemplification. If an object does not possess a property, it cannot exemplify it. However, instantiation is not in itself sufficient: an object will always instantiate an indefinite number of properties that are not exemplified.

Exemplification, then, is a referential relation: an object refers to a property that it possesses – or, with a more nominalist flavour, to a label or a name that applies to that object. Thus, exemplification is a distinct form of reference from denotation, which, recall, is the referential relation that goes from a symbol, and usually a label or a name, to the object complying with that label. From now on, for simplicity and ease of comprehension, I will dispense with label talk and speak instead of properties or features. Philosophers who are uncomfortable with property talk for reasons of ontological purism can easily translate this into terms of linguistic labels.

More generally, my position will not primarily concern itself with the ontological or metaphysical questions surrounding representation; where I do offer reflections on these matters, it is usually only to demonstrate that my positions on the semantic and epistemological features of representation do not require any radical or controversial metaphysical or ontological commitment. For this reason, for example, I have shown how the account starts from anti-realist assumptions about fictional entities, and that we can talk about properties without committing to their existence. But nothing rules out further ontological inflation. This of course means that pairing my account with different ontological or metaphysical views will give rise to different philosophical consequences.

Two points about exemplification require special emphasis here. First, what properties an individual representation exemplifies are not fixed: depending on the context and on our purpose when using a representation, different salient properties will be brought to the fore and thus exemplified. This is important because certain properties that may seem completely negligible in one context may become salient in another. Imagine that someone takes the swatches from the shop of the previous example and shows them in a university lecture on the globalisation of the economy and production of goods. The fact that those swatches were all produced in China would now become an exemplified property of the sample, while the specific colour

and fabric are no longer referred. The very fact that the swatch exemplifies a property at all is not automatic or intrinsic: it may be used, for example, to clean a stain on a table. In this case, it does not function as an exemplar, as it does not exemplify anything. Ideed, it does not work as a symbol at all.

Second, and relatedly, exemplification is fundamentally selective. In order to highlight and make salient certain features, other properties of the object will have to be ignored or set aside. If an object refers to a certain property by exemplification in a certain context, it will often afford the user a certain privileged epistemic access to that exemplified property.[22] But this, in turn, will require that some other properties be overshadowed. Maps of the London Underground network exemplify certain topological properties of the system, such as connections between different lines, to enable a user to choose the most efficient way to move from point A to point B. However, the fact that a specific line has a specific colour, say, red, is not to be interpreted as though there was an actual red line on the ground, or that the train is red, or anything of the sort. The colours are a convention to identify a line and distinguish it from other lines. The same holds for other properties of the underground map, such as the distances between points on the map.

In the theoretical background of the concept of exemplification, then, there seems then to be no way clearly to define a set of intrinsic or fundamental properties exemplified by an object, independently of the use to which we are putting it. This does not mean that anything goes, however. Given a certain context and a certain purpose, there will be more or less relevant properties we will want to look at, and more or less useful ways to exemplify them. Exemplification also does not seem to be a pure matter of stipulation: in order to refer to a certain property, users must be granted (facilitated) epistemic access to that property (Elgin 2010, p. 9). Again, this is not something that it is intrinsic to the exemplar. The fabric of the swatch has nothing special with respect to any of its other properties. But it is not a pure matter of stipulation that in that situation one has to focus on the fabric of the swatch: it is the epistemic background created by the shop and the purposes of its users that participate in the definition of what the swatch is and is not an exemplar of.[23]

The account of epistemic representation resulting from the above distinctions is the one proposed by Elgin (2010): it combines denotation and exemplification, and describes epistemic representation as a three-place relation. Similarly to the caricature of Mrs. Thatcher, a scientific model represents its target as a Z, and does

[22] Frigg and Nguyen (2020, p. 172) explicitly define exemplification in terms of highlighted properties that are also made epistemically accessible.

[23] One can see here how the focus in this account is shifted away from, for example, a single user and their idiosyncratic mental representations, and into the social context of an epistemic community, with a system of symbols in the background, and a set of legitimate or illegitimate interpretations based on some epistemic purposes set by the community. I take this to be a strong advantage of this view over views of representation that focus on the agent as a single individual (as for example Giere 2010).

so by exemplifying certain properties that then one can impute to the target system. More formally:

Epistemic representation-as$_{(\mathbf{def})}$: A model system M represents a target T as a Z iff:

(i) M is a Z-representation and as such exemplifies a set of properties P_1..., P_n;

(ii) M denotes T;

(iii) P_1..., P_n are imputed to T

As it goes with caricatures and tree-representations, Elgin's account suggests that the same happens with scientific models: Watson and Crick's model privileges the structural aspects of the DNA, and expects us to ignore the dimensions and colours and material used for the carrier. Similarly, the Newtonian model of the Solar System exemplifies the elliptical orbits of the planets, but not the dimensions or shapes of the planets. In the same way, the Phillips-Newlyn machine exemplifies the flow of money from one economic sector to another, but we are supposed to ignore the pump moving the water at the bottom of the machine back at the top (Frigg and Nguyen 2020, p. 173).

We are still a few developments away from the DEKI account. However, some of the elements are already in place: denotation, exemplification, and imputation. Before moving on to the required additions, I will now spend some words on the deep root that this account of representation has with the broader philosophical views held by Goodman and Elgin, who so much contributed to the representation-as framework and influenced the DEKI account so crucially.

1.4 Goodmanian and Elginian roots, and forward

The conception of exemplification I will employ in what follows is linked to a very liberal view of properties and objects that traces back to the work of Goodman and Elgin. In this tradition, there is no true hierarchy of properties – intrinsic vs. extrinsic, essential vs. accidental, internal vs. external (see e.g. Goodman 1978). As a consequence, there seem to be no non-instrumental grounds to define natural or proper kinds as distinct from other kinds.[24] Further, there is no way to distil a

[24] I do not suggest here that it is impossible to offer such grounds for a distinction between essential and non-essential properties, or for similar distinctions. What I want to suggest is simply that the most natural and parsimonious way to identify "important" properties in science is to ask whether they are relevant or useful for a certain epistemic or pragmatic purpose. Thus the order of explanation is reversed: it is not because that some properties are essential, intrinsic, or internal that they are relevant or useful; rather, if they are relevant or useful for certain goals, they become more and more entrenched within our epistemic practices, and we end up considering them as more important, natural, essential, etc.

concept of pure content or pure data that is completely divorced from any conceptual framework used to interpret it.

All of this may at first appear more radical than it actually is. First, while there are no essential properties *simpliciter*, there are still better or worse ways to describe a system in order to achieve a certain particular purpose. While we do not have a list of objectively essential properties, we can still offer reasons to identify some properties as being more relevant than others in order to pursue our goals.

Similarly, just because the world is concept-laden "till the bottom", it does not follow that the way in which we shape and organise the world with our concepts and taxonomies is completely random. It may be arbitrary (i.e. it could have been different), because the resulting organisation is always relative to a theoretical framework of reference. But the choices we are making in conceptualising the world in one way rather than another are usually supported by reasons that can be and in fact are subject to evaluation. And this evaluation needs not to depend in a circular manner on the conceptual framework adopted. Or if it does, the circularity involved is not necessarily harmful. This is because, while all of our concepts are part of a holistic system, different parts of our conceptual framework are more or less separable from other parts, and thus compatible with alternative conceptualisations. If the way in which we conceptually partition the world in certain situations ultimately depends on other conceptual taxonomies, this allows us to reshape different parts of our net of concepts in order to adapt to new information and to address different purposes and values.

Furthermore, my view allows for different worldviews to interact, be used to test one another, modified, mashed up, and so on. Therefore, this view does not open the gates to radical relativism, in the sense of total incommensurability between conflicting positions. What I hold is simply that what counts as relevant in a given context is always relative to what we want to do and what we are interested in when we interact with objects and systems. At the same time, this approach does not entail that we can change our way of understanding the world at will: changing something in the system could affect the entire system, which may be very arduous or ultimately undesired.

At this stage, it will be helpful to clarify what I take a conceptual framework to be. A conceptual framework is a taxonomic system, in which the meaning of each category depends on its relations to the rest of the taxonomic system. In this sense, a conceptual framework is holistic. It does not need to be linguistic or propositional, nor to be endowed with sharp boundaries between the taxonomic categories involved. For example, perception seems to function based on a background conceptual framework in this sense: in order to perceive the colour black, for instance, there must be a conceptual framework that allows a perceiver to distinguish colours, namely the colour black from at least one other non-black colour. Of course, the perceiver does not need a word for black, or for any other colours, in order to distinguish the colour black

from these other colours. This conceptual taxonomy also does not need to entail a complete first-order logic and its operators.[25] Yet, there must still be a system of reference in place that allows for the identification of black in contrast with other non-black colours.

Any property attribution thus presupposes some form of background conceptual framework. This naturally applies to symbols too: an individual symbol's meaning is given by the entire symbol system of which that symbol is part (Goodman 1976). This applies to linguistic symbols, such as the letters of the alphabet, but also to non-linguistic symbols such as the colours in a painting, the lines constituting a fever chart, or the coloured lines of the train lines on our Underground map. Symbol systems, then, are just a special case of the more general class of conceptual frameworks as described above.

Some hardcore Goodmanians (or indeed some of his unshakeable opponents) may expect my endorsements above to force me to endorse all of his other main philosophical positions. However, this is unnecessary. My account so far, for example, does not entail accepting Goodman's strong nominalism about properties and classes. It is simply compatible with it. While I would like to remain as ontologically parsimonious as possible and am inclined, as is Elgin (1983), towards a purely extensional theory of reference, I am not in the business of denying ontological status to properties and classes. Of course, we can admit talk about properties and classes within our discourse as short expressions to refer to labels that stand for series of individuals etc., and with the same spirit I do not have particular reasons to deny talk of more "fancy" properties, such as relations, dispositions, and powers. Thus, until proven otherwise, the view sketched out so far is perfectly compatible with (but importantly, does not imply) more luxuriant or extravagant ontologies or metaphysical views.

Another radical view expressed by Goodman in his later work (1978) is that the plurality of ways we have to carve up reality via different conceptual frameworks implies conflicting worldviews. Given that in Goodman's view there is no such a thing as a world independent of the views we are endorsing, and that we can and do hold different worldviews resulting from different systems of conceptual frameworks, we can obtain "worlds" that are in conflict and eventually incompatible with each other. This may again be too strong: merely because some views originate from different conceptual frameworks, we need not conclude that they are irremediably incompatible. In fact, part of the scientific endeavour is to seek ways to reconcile different conceptual frameworks originating from different goals and contexts of application within a unified conceptual system.[26]

[25] Pictorial, non-linguistic systems such as pictures and maps seem unable to express operators like universal quantification or logical negation, without being further integrated with other more conventional forms of at least semi-linguistic symbols. See Camp (2007) and Rescorla (2009).

[26] Here, I feel the heritage of the argument put forward by Davidson (1973) against incommensurability and the very idea of conceptual schemes – which he calls "the third dogma of empiricism". While philosophers may disagree on whether Davidson succeeds in dispelling the problem (or even

A discussion of this last issue would take us away from our main focus on epistemic representation. The important point here is simply that while my view on representation is inspired by Goodman's work on representation, properties, language, and concepts, it does not include and is not beholden to his other views, including on nominalism and conflicts between worldviews. My position is a more moderate one, where Goodman's antirealism about fictional entities and abstract object is retained, and the possibility of worldviews conflicts is acknowledged as a possible feature of epistemic endeavour, but it is not assumed to be an irresoluble one.[27] What is also retained from Goodman's views is the omnipresence of conceptual frameworks and their constitutive role in the ways we interpret, understand, and interact with the real world, which would seem to be a prerequisite for any minimally sophisticated metaphysical position in the debate.

As regards few words my position relative to Elgin's, her Goodmanian, extensionalist theory of reference in *With Reference to Reference* (1983) provides the basic semantic foundations of my work on epistemic representation in science. Furthermore, as I have tried to show so far, most of the basic elements of the DEKI account of representation, namely denotation exemplification, and imputation, are already present in her account of models put forward in "Telling Instances" (2010).

Moreover, for reasons that will become clearer in the following chapters, my view of epistemic representation naturally fits the general epistemological theory put forward in Elgin's book *Considered Judgement* (1996), in which she argues that knowledge is always holistic. The justification for a belief can be offered only by looking at a general system of beliefs, assumptions, and hypotheses which we already endorse. Progress in our knowledge also implies continuous adjustments to the net of beliefs previously endorsed, aiming at maintaining a reflexive equilibrium of our different beliefs, taxonomic systems, theories, and assumptions. This will become manifest, paradigmatically, when I discuss the justification for our inferences regarding a target system on the basis of a representation, both in this Chapter and when I move to thought experiments, experimental organisms, and scientific pictures. Indeed, a *fil rouge* running through my argument will be that the source of this justification cannot be given by looking at the single representation system in use. The justification for representational inferences is always (at least partially) extrinsic to one representational framework and requires a more holistic understanding of scientific representational practices as integrated into a network of representations, theoretical assumptions, and empirical knowledge.

the idea) of incommensurability, his arguments definitely undermine the problem of conflicting worldviews. I am broadly sympathetic to the idea that, however distant and different from one another our conceptual frameworks may be, there will still be enough common ground for us to understand each other well enough, even if only to recognise the very differences that separate us.

[27] Then, my view remains compatible with a variety of positions in the debate on scientific realism, including metaphysical scientific realism (Psillos 1999), perspectival realism (Massimi 2022), constructive empiricism (van Fraassen 1980), and others.

In her more recent book *True Enough*, Elgin goes further, proposing to abandon a traditional epistemology based on the concept of truth and the consequent notion of knowledge as justified true belief. She argues that it is better to embrace a new epistemology based on a non-factive notion of *understanding*,[28] which is then not founded on truth and is instead contextual, holistic, and action-oriented. On the one hand, she argues that truth is not a sufficient criterion for good scientific reasoning: far too many true propositions are just irrelevant to our scientific purposes (cf. also Cartwright 1983). This position was in a certain sense expected: as objects possess many properties but not all are relevant, we can say many true things about reality, but most of them are uninteresting for any specific purpose.

On the other hand, Elgin points out that truth does not seem to be necessary: many of our scientific theories and models make heavy use of approximations, abstractions, idealisations, and distortions, making them literally and explicitly false about the things they are intended to represent. She also offers further reasons to doubt that the traditional concepts of truth and knowledge can give us a successful epistemology for science (or of art or other epistemic contexts), however these are not my focus here.

I take Elgin's challenge to the traditional factive epistemology as a challenge that must be addressed by any thorough investigation of surrogative reasoning in science. I would agree with Elgin that truth, even when justified, is not the only epistemic value that we consider when it comes to (scientific) knowledge. Many other epistemic and practical virtues enter the picture when it comes to assessing a product of science (theories, models, experimental results, systematisations of data, and technological productions and interventions). I also concede to Elgin that many of our representations provide literally false descriptions of their target systems. This is why we talk about Z-representations, and do not assume that the Z is an actual system or an object in the world. I also suspend judgement on whether Elgin's further arguments against traditional epistemology are sufficient, and whether her positive proposal of a new epistemology based on non-factive understanding is successful. What I want to focus on here is her account of the apparent non-factive character of representation. Following Frigg and Nguyen (2021), I hold that there is a way to respond to the specific problem of the literal falsity of epistemic representation while still retaining the tenets of traditional epistemology. I present this solution in the next section.

[28] The literature on scientific understanding has grown exponentially in the last few years, but relevant entry points include De Regt (2017), Doyle *et al.* (2019), Elgin (2017), Illari (2019), Khalifa (2017), Kostić (2019). Le Bihan (2021), Reutlinger *et al.* (2018), and the papers collected in Grimm *et al.* (2017).

1.5 Interpretative functions: the key and the I

As I stated above, Elgin recognises the problem with making inferences from idealised models to actual targets, because the former include distortions with respect to the latter. Of course, some of these distortions will be taken care of by the selectivity of exemplification: some properties of the model are just to be ignored. However, models often distort properties that constitute the respect in which we study them as surrogate systems of our targets. A harmonic oscillator model of a pendulum is supposed to give us a story about the dynamics of the pendulum, and therefore to explain it in terms of forces. However, the model ignores friction, thus distorting the description of the forces involved. This is by no means an exception, and examples of models that distort salient properties of their targets abound in every scientific discipline.

The philosophical use of terms like approximation, abstraction, idealisation and distortion is extremely varied across different authors.[29] For the purposes of this thesis, it will be useful to regiment the language and conceptually distinguish between these terms. First, I take distortion as a vague enough umbrella concept that encompasses approximation, idealisation and abstraction. Furthermore, to distinguish these three latter terms, I follow the distinctions drawn by Frigg (2022, Chapter 11). I refer the reader to Frigg's book for a detailed illustration, but in a nutshell: approximation expresses quantitative, mathematical closeness between values. Abstraction and idealisation, by contrast, are not necessarily expressible in mathematical terms. Abstraction involves the omission of properties that do not pertain to a respect (or a dimension) that is represented by the model – for example, the colours of objects in a mechanical model that is meant to represent the physical forces acting on a system. Finally, an idealisation consists of any distortion (or even omission) of a property that, unlike in the case of abstraction, falls under a respect that is meant to be represented by the model – such as ignoring friction in a model that represents the mechanical forces acting on a system, of which friction is one example. We have now a way to distinguish a type of model distortions, idealisations, that are particularly problematic for a factive epistemology of scientific representations.

Elgin (2017) suggests that the presence of idealisations in my technical sense is one good reason among many to relax the dependency of our scientific reasoning, and surrogative reasoning in particular, on truth. Models can afford us an understanding of the target even if they are literally inaccurate with respect to their target systems *relative to properties that are relevant to a respect which is meant to be represented.*

Notice that this problem cannot be solved by a simple appeal to competent users. In a recent talk, for example, Alexander Bird attempted to retain the factive

[29]Recent discussions of idealisation and approximation can be found in Batterman (2009), Cartwright (1983), Elliott-Graves and Weisberg (2014), Jebeile and Kennedy (2015) Nguyen (2020), Norton (2012), Portides (2007), Potochnik (2017), and Saatsi (2013).

nature of scientific understanding provided by models based on an appeal to the concept of epistemic representation proposed by Suárez (2004).[30] In his minimalist account, Suárez gives only necessary but not sufficient conditions for a system to be a representation of a target. Roughly, it is enough that a competent user is able to use the model system in order to draw inferences about a target system.[31] Bird argues that such an inferentialist minimalism, paired with the role of a competent user, can solve Elgin's dilemma. For a competent user will be able to understand the model well enough to interpret it correctly and draw only (allegedly) true inferences about the target. Of course, the representation itself could just be incorrect, but that does not concern us as the worry here is representations involving falsehoods (idealisations, abstractions) that nevertheless work well epistemically, namely providing accurate information about or a good understanding of the target. If what we get, Bird argues, is just what a competent user would get, then falsehoods are not really a problem, because the competent user will already be able to sift through and eliminate the non-factive information.

I would argue that while Bird's suggestion is a move in the right direction, this specific strategy focusing on inferences drawn from a competent user does not succeed. This is because the inferences drawn from models and representations in general often have a holistic, non-modular nature. In order to get the results we are interested in, the correct, allegedly true one, and thus in order to obtain a factive understanding, one still has to assume the idealisations in the first place. Indeed, it is exactly because of those idealisations that certain properties can be exemplified, and certain important results about the target obtained. If one assumes an inferentialist approach, however, there seems to be no principled way to "dissect" a representation and select the factive information only, as there is no real way to identify and "delete" the idealisations without then also renouncing their roles in our inferences. From an inferentialist point of view, everything that is part of the inferential reasoning is also part of representation, and thus of the understanding of the target provided by such representational activity. Therefore, the "pick-what-you-like" strategy proposed by Bird seem impossible from the beginning, if one endorses a too-strong inferentialist perspective on representation.

However, there may yet be a way to run with the hares and hunt with the hounds. The solution, proposed by Frigg and Nguyen (2021), is the introduction of a *key*, that is, a mathematical function which maps the properties exemplified by an idealised model system onto the properties we actually want to impute to its designated target system.

For example, in the case of Watson and Crick's model, the general structure of

[30] This talk was given in the form of a comment on the book recently published by Mauricio Suárez (2024) at the HPS Department in Cambridge (22 May 2024).

[31] There is a further condition that the "representational force" of the model points to the target system, but this requirement is not relevant for the argument about the factivity of representations.

the DNA and the bonds between the two strings of nucleotides are imputed as they are to actual DNA. However, models often exemplify properties that are not the ones eventually imputed to the target system. A scale model of a bridge, for example, will also need a scale factor to translate the dimensions of the model system into those of the actual bridge. Sometimes, this factor will be temporal, as in the case of cycles in the Phillips-Newlyn machine (Frigg and Nguyen 2020, p. 174). In mechanical models, the limit values of some parameters will have to be translated into non-limit values (Nguyen and Frigg 2020). Geometrical projections are another example of keys being employed in scientific visual representations, mapping the properties of two-dimensional objects onto those to be imputed to three-dimensional systems. For example, a key is provided to translate distances on a planisphere to actual distances on Earth (Nguyen and Frigg 2022a). When the properties exemplified by the model are exactly the ones we impute to the target system, the key will simply be an identity key, mapping the properties exemplified in the model into themselves.

The concept of a key is important to retain the difference between what the model actually seems to describe – an abstract, idealised system – and what we can take to actually learn from that system about the represented target system. A key, then, is a more or less systematic way of converting knowledge of the model system into hypotheses about the target system. Compare this strategy with Bird application of Suárez' inferentialism. In Frigg and Nguyen's proposal, we can preserve Elgin's intuition that idealisations participate in the production of new knowledge and understanding, because we acknowledge their role in representation. What is factive, though, are the final imputations, i.e. the final outcome of our inferences. Suárez's inferentialism, because it is so minimal, does not make a distinction between inferences about the model and inferences from the model to the target. The model is taken as an inferential instrument about the target from the start. The DEKI strategy here, by contrast, consists of a sort of *divide et impera*: we first appreciate the value of the exemplified properties within the model, then we try to convert this information into usable imputations for the target to the best of our abilities. At the same time, Bird broader aim of retaining the factivity of our final understanding of the target is vindicated. We can now reconcile Elgin's and Bird's views, at least as concerns the nature of the understanding provided by models. While Elgin seems to be right on the ineliminable value of falsehoods for a correct appreciation of a model's functioning and the way in which it provides understanding, the final outputs of the representational process can still be understood in a more traditional way as factive, as Bird sought to argue.

Keys are not always so easy to define: the double-helix model of DNA exemplifies many properties that we may want to impute to the actual DNA, but it is still a strongly idealised representation. Therefore, as philosophers of science, we should recognise the possibility that some properties possessed by the model and made salient within it do not directly translate to the target, and indeed, the model itself will

imply this once properly interpreted.

More generally, when we are dealing with a representation, we will often encounter idealisations, and we will need a way to interpret them. For some properties, no idealisation will actually be involved, and the property can be imputed as is to the target. At other times, the idealisations will be re-translatable in factive information about the target.[32]

As an additional advantage, the key seems also to account for what we would call the *metaphorical* reading of representations in non-scientific cases.[33] A caricature of Winston Churchill as a bulldog, for example, is meant to exemplify certain properties of a canine and attribute them to the politician. But surely the caricaturist, as well as the observers who interpret the work, will not impute the properties of a bulldog to Churchill in a literal sense. The aggressiveness of the dog will become political ruthlessness, the loyalty of the dog to their owner, loyalty towards his own ideals, the animal stubbornness will perhaps be interpreted as perhaps conservatism, and so on. As can be seen here, the keys employed in art tend to be less systematic and more difficult to define than in science, where one aim is to make our inferences easy to share between members of the epistemic community of reference. In this sense, metaphor can be seen as a form of proto-keying-up, with a vague enough mapping function associating words with new fields of applications, thus endowing them with new meanings.

Besides the key, Frigg and Nguyen have also contributed to Elgin's representation-as account some further reflections about how a material object becomes a Z-representation in the first place.

Of course, it is usually the case that a painting can be easily and automatically identified as, say, a tree-representation. More difficult though is to explain how we move from the very mundane plastic object that instantiates the abstract model of Watson and Crick and that we can encounter in science classes in schools to a theoretical representation of the DNA as a double-helix structure of nucleotides chains connected by nitrogenous bonds. More generally, in science we have often material objects that undergo a substantial theoretical revision once they are used as a Z-representation of some external target. The most natural way to think about this, as Frigg and Nguyen (2020, p. 166-171) suggest, is a form of interpretation. First, let's call the material part of the representation the *carrier* of the representation (I will use the letter X for the carrier from now on), which will then be conceptually distinguished from the Z-representation. Remember the case of the caricature of Margaret Thatcher as a boxer. We have a picture on a sheet of newspaper, which is

[32]For a generalisation of the view that, for interpretation-based accounts of representation like DEKI, idealisations are not in fact a necessary evil for feature-selection and salience, but also make an epistemic contribution when dealt with properly, see Nguyen (2020).

[33]On metaphor within the present theoretical framework, cf. also Goodman (1976, pp. 81-84) and more specifically the developments by Elgin (1983, pp. 59-70).

a collection of ink marks that need to be interpreted as a boxer-representation. This requires some interpretation on the observer.

Specifically, one has to endow different material parts of X with some interpretation of those parts, which are then themselves taken as a symbol for something else. For most caricatures and paintings, this interpretation comes automatically. The shape and other visual aspects of the caricature are readily associated with three-dimensional objects, such as boxing gloves and other typical properties of boxer-representations. In the scientific realm, however, interpretation can be less trivial.[34] For example, the two spiral strings in the material instantiation of Watson and Crick's model will have to be interpreted as chains of nucleotides, and the sticks connecting them horizontally as bonds between couples of nitrogenous bases. The recognisable spiralling structure has to be interpreted as the DNA molecular structure. This seems to be the case in most scientific models. The water flowing through pipes into different reservoirs in Phillips-Newlyn machine must be interpreted as the flow of money between different sectors of a national economy. The same with the material example of Newton's model of the Solar System, where balls are interpreted as perfect spheres interacting with each other only in terms of gravitational forces, with no friction involved.

Interpretation, Frigg and Nguyen argue, can be thought of simply as an interpretation function I, which associates the different material elements of the carrier X with properties of the Z. By applying the I-function to the carrier X, we obtain the actual model system M. It is M, and not simply the X, that represents a designated target system T.[35]

The function of the I goes beyond associating material properties with the Z-properties: it can also help us make sense of the selectivity of exemplification. We can just specify that the I does not map all of the properties of the material carrier into properties of the Z-representation. So, the function of our interpretation of the object is not only to assign a theoretical interpretation to a certain material property of the carrier, but also to discriminate between salient and non-salient properties.

The introduction of the I-function gives rise to a significant difference to the original account proposed by Goodman and Elgin. Recall that the definition we gave of exemplification included instantiation. If a property is not actually possessed by the representation, it cannot be exemplified. However, most material models in science do not actually possess the properties that we focus on when we interpret them theoretically – consider Phillips and Newlyn's hydraulic machine interpreted as a Keynesian economy, or Kendrew's (1958)[36] material model of the myoglobin protein. The I-function helps us move towards a theoretical interpretation of the

[34] Even in the arts, symbolism can be arduous to decipher, not only due to historical or cultural gaps, but also depending on how cryptic or esoteric the author wishes to be.

[35] This can be extended to non-concrete models: see Frigg and Nguyen (2020, pp. 185-190).

[36] For an analysis, cf. de Chadarevian (2004) and Frigg and Nguyen (2016).

material object without assuming that that theoretical interpretation actually refers to any existing object in the real world. Plastic balls are interpreted as perfectly spherical objects, wooden sticks as atomic bonds, and water as a hypothetical flow of money that may never be instantiated in the real world. But then, we would seem to lose the actual instantiation of the properties of the Z-representation, and thus, it would seem impossible to talk about exemplification in the first place.

Frigg and Nguyen propose to amend the original account advanced by Elgin and specify that we are not strictly speaking about instantiation and exemplification, but rather of I-instantiation and I-exemplification. It is the model as an object endowed with an interpretation that instantiates and exemplifies the properties in which we are interested, not the material carrier alone. Of course, nothing precludes the I-function from being an identity function, mapping the material properties of the model into exactly the same properties in the Z-representation. This may perhaps occur more often in what we may call more realistic paintings and photographs (where at least the colours of the picture are to be interpreted as the colours the Z-representation). But in science, most of the properties of the carrier will have to undergo substantial re-interpretation in order to become elements of the theoretical domain of the Z.

Now that we have introduced both the key and the I-function, a question arise concerning the relation between these two interpretive functions and their role in DEKI: how do we distinguish the I and the key, given that both are interpretive functions? One might wonder if this is merely a vague, arbitrary, and ultimately useless distinction. To illustrate the problem with a simple example, take the case of the litmus paper used in chemistry to measure the acidity of a solution. We have two extreme views here. On the one hand, we can say that the colours are already mapped via the I-function to different levels of acidity, which are consequently I-exemplified given the context of enquiry. The key in question is then inert, as it is just an identity function between the properties exemplified by the Z-representation and the properties we impute to the chemical solution. At the other extreme, we have no interpretive function: the carrier is the one that in certain context possesses and exemplifies certain colours, and the key translates colours into levels of acidity.[37] So, why do we need both, if one is enough?

There could, however, be more nuanced accounts of what is going on. One may say that the I-function indeed already maps colours to levels of acidity, thus creating an acidity-level-representation of the solution, but we need the key to translate the continuum-like values of the litmus paper into discrete levels (or ranges) of acidity imputed to the solution. This requires a conceptual distinction between the I and the key.

There is clearly some advantage to recognise the conceptual distinction between interpretation and key. First, it allows for nuances in the interpretation of epistemic

[37] Frigg and Nguyen (2020, p. 103) explicitly endorse the latter position for the case of litmus paper.

representations, and if a philosopher decides to describe a model in one way and not in another, they will have to motivate their choice.

Second, the distinction seems the best way to acknowledge a certain autonomy of the Z and the fact that what is true about it is independent from its representational use. This dimension can in principle be distinct from the purely representational purposes of attributing properties and making inferences about the target system. A Z, to paraphrase Hacking's words, has a life of its own, even when it is already interpreted as a representation of a target system.

Third, the same model may apply differently to different target systems and in different contexts, and thus may require different keys. The key is not technically entrenched in the Z-representation *per se*, and thus in the consequent set of exemplified properties. Given that the DEKI account describes the model system as simply as an object endowed with an interpretation, the key is relatively detached from the exemplifying system. So, the key gives us some leeway to articulate how the same model can in fact relate representationally in different ways to different target systems. In the next Chapter, I will use the example of a Galilean thought experiment that functions as a representation in the same that scientific models do, and results of which apply differently depending on the characteristics of different target systems.

Sometimes the distinction between the key and the I-function is simply the product of the historical evolution of a model. The model system in itself may be thought of as a model of a certain target system T and then applied, provided the right key can be found, to new target systems, as in the case of hydrodynamic models were used to represent electromagnetic phenomena. Does the key disappear, in the long run? Sometimes it may. At other times, it persists independently of the interpretation because the Z, as interpreted in a certain way, acquires a certain conceptual autonomy. Think of the model of the ideal gas described as a system of particles that do not interact with each other but only with the surfaces of their container. This model is useful even though it does not give true results about many of the target systems to which we apply it (Elgin 2017). Elgin suggests that the solution to this problem is related to Strevens' (2008) concept of negligibility: what diverges from truth is not a difference maker. I think that the solution can be generalised thanks to the concept of the key. The model in itself is valuable in its abstraction and generality. What we need is a set of different keys to apply the same model to different contexts and target systems. This may be seen as a general solution to Cartwright's famous point that the laws of physics lie (1983). Either we accept that truth is not really the point (Cartwright 1980, Elgin 2017), or we instead recognise that, while it can be very useful to entertain fictional systems in our scientific research, it is also always essential to bring the fiction back to reality via a proper key.

Finally, nothing in the DEKI account requires the various elements (denotation, exemplification, the key, and the I-function) to be static and completely sealed off

from each other. They are interactive: changing the description of the I-function may change the exact definition of the target system denoted by the model, and the exact exemplified properties, which then in turn may require a different key before they can be imputed. I will return to this point and further develop a dynamic, interactive understanding of DEKI in section 1.7.

1.6 Imputation and justification

So far, we have analysed in more detail three of the four elements constituting DEKI: denotation, exemplification, and keying-up. The last ingredient of the account is imputation: we take the set of properties translated by the key and we impute them to the target system T of interest. As in Elgin's original formulation, imputation is to be understood as simple property attribution: we attribute some property to the target system, but this does not imply that our attribution is correct.

We have now all the basic ingredients of the DEKI account.

DEKI epistemic representation$_{\textbf{(def)}}$: Given a model system M, defined as an object X endowed with an interpretation I that makes M a Z-representation; M epistemically represents a target system T *iff* four conditions jointly obtain:

(i) M denotes T,

(ii) M exemplifies properties P_1..., P_n,

(iii) P_1..., P_n are associated with a second set of properties Q_1..., Q_n via a key,

(iv) Q_1..., Q_n are imputed to T.

This is intended to be a substantial development of Elgin's original proposal illustrated in section 1.3, where the original elements of denotation, exemplification and imputation are complemented by (a) the I-function to explain the creation of the Z-representation and (b) the key to account for de-idealising practices and thus retain the factivity of scientific representation.

As mentioned above in relation to the concept of imputation, the DEKI concept of representation provides no guarantee of accuracy or correctness: we could just be wrong. This is important because we want the concept of representation to be able also to include misrepresentations, and to distinguish them from cases of non-representation.

The possibility of misrepresentations being representations sheds light on an important aspect of the inferences we draw from models, and from representations in general. For there are two distinct senses in which an inference drawn from a representation is correct. In a first sense, an inference can be correct in that it follows from the correct interpretation of the representation. I follow a pirate map towards

the "X" by correctly interpreting the "X" as indicating the point where the treasure is buried. My inference is correct insofar as it is based on a correct interpretation of the map. However, the treasure may not be there. While I have read the map as it was supposed to be read, the inference I made simply gave a wrong result: the treasure is in fact not where the map indicates. I follow Frigg and Nguyen (2022, p. 296) and call "derivational correctness" and "factual correctness", respectively, the two type of correctness of an inference drawn from a representation about its designated target system:

> An inference drawn from a representation is derivationally correct [about the target] if the inferential steps that lead to the conclusion are correct with respect to the rules of the representation and only use premises that form part of the representation. The conclusion of an inference is factually correct if the conclusion is true of the representation's target. (*ibid.*)

According to DEKI, the assessment and justification on the first type of correctness is given, jointly, by the I-function, the properties exemplified by the model system, and the key in place.

What about the justification for factual correctness? In previous work (Sartori 2023), I have provided a definition of representational accuracy that basically corresponds to Frigg and Nguyen's (2022) idea of factual correctness of our representation inferences. In a nutshell, a representation is accurate when the inferences we draw from it about a target are factually correct.

Let us give more details on my contribution to the concept of representational accuracy that enriches the original DEKI account. I take it that the accuracy of a representation is always relative to two factors: the designated target of the representation, and the specific set of properties that is eventually imputed to that target. As such, we can never talk of representational accuracy *simpliciter*: we have to specify the target and which property exemplified by the representation we want to impute to the target – once properly translated via a key. By adding the condition of factual correctness, I hereby develop the DEKI account by offering a simple, precise definition of accurate representation:

Accurate representation$_{\textbf{(def)}}$: A model system M is an accurate representation of a designated (i.e., denoted) target T regarding a set of properties $Q_1...$, Q_n *iff* (1) M exemplifies a set of properties $P_1...$, P_n; (2) $P_1...$, P_n are converted via a proper key into $Q_1...$, Q_n;[38] (3) $Q_1...$, Q_n are imputed to T; and (4) T actually possesses $Q_1...$, Q_n.

In other words, a representation is accurate if we impute properties to the target and the target actually possesses those properties. It is important to stress again that my

[38] I remind the reader here that the key may also be a relation of identity, mapping a property P_i onto itself.

definition of accurate representation does not require similarity or isomorphism, nor they are sufficient for accuracy t obtain. They are not required because the presence of interpretation (implicit in the definition of the model system M) and the key (when not an identity function) make similarity unnecessary. They are insufficient, because accuracy requires all four of DEKI's components to obtain.[39]

Clearly, my definition of accuracy does not provide us with a procedure to assess whether a representation is in fact accurate, other than a direct check on the target itself. This is because the assessment of the factual correctness of our inferences depends on the truth values of propositions about the actual states of affairs in the target system. Sometimes, we are able to directly observe and check whether the target system actually is how the model predicts. However, most of the time we cannot make direct observations of the system, hence why we are using a model in the first place. The justification for the factual correctness of our inferences, then, will crucially depend on information that remains extrinsic to the single representation system:[40] theoretical assumptions provided by well-established theories, the results of other, independent models, and the results of experiments and data analysis. Even if we study models and representations as single units of scientific practice, it is almost never the case that these units function in isolation, completely detached from the rest of our theoretical and empirical background knowledge. Knowledge and understanding should thus always be regarded as holistic concepts (Elgin 1996), and the scientific case is not different in this respect.

1.7 Meeting the general *desiderata* for representation

On the basis of what I have said so far, the account should also be able to satisfy six general *desiderata* that any account of epistemic representation is generally expected to meet: our concept of representation should explain for representational directionality, the possibility of misrepresentation, the existence targetless models, the ability to carry out surrogative reasoning, the application of mathematics in our models, and the dynamic nature of representation.

First, denotation and imputation allow the account to describe the *directionality* of representation: a model denotes its target and not vice versa, and we impute properties to the target on the basis of the model but not the other way around.

Second, nothing in the model rules out the possibility of *misrepresenting* the target: imputation is just property attribution. The factual correctness of our inferences is not built into the concept of representation, because it can be conceptually distinguished from the derivational correctness of the inferences made on the basis

[39] By detaching accurate representation from similarity, I agree with Nguyen (2020, §4) and with what he calls "interpretational" accounts of scientific representation, among which he lists DEKI. For a survey of these accounts, see Frigg (2022, Chapter 9).

[40] In this respect, my conclusion is similar to that at which Frigg and Nguyen (2022) also arrive.

of the representation system.

Third, the account also expresses the sense in which a representation can be *targetless* via Goodman's view that representation is an ambiguous concept and can mean both representation-of and Z-representation. With targetless representations, such as the model of phlogiston and Norton's globe, we do not in fact have any denotation, because one of the *relata* – specifically, the *denotatum* – is missing. However, we still have a Z-representation, namely a theoretically interpreted object in the case of a material model, or a fictional system described by a text or illustrated by an image that does not denote anything in the real world.

Fourth, the account gives an answer, if only skeletal, to the question about how a representation allows *surrogative reasoning* about its target: the model system, once interpreted theoretically, exemplifies properties that we then impute to the target system using a certain key. This describes the correctness of an inference based on a representation about its target system, given that specific representational framework. However, we cannot justify the factual correctness of our inferences within that framework alone: factual correctness can be assessed or justified only by looking outside of that single representational framework. However, this implication should not surprise us. In order to deem a misrepresentation still to count as a representation, and to distinguish both from non-representations, the account cannot require our inferences to be successful: our reasoning must be able to bring about wrong outcomes about the target system. Also, this fallibility of surrogative reasoning seems to be expected from a form of reasoning that is non-deductive and ampliative in its essence: moving from a surrogate system, like a model or a picture, to a target system will always be tentative and conjectural.

Fifth, an important feature of scientific models is that they employ mathematics to represent phenomena. So an account of epistemic representation seeking to be applied in scientific contexts should also explain the role mathematics in representational reasoning. The originators of DEKI (Nguyen and Frigg 2017, Frigg and Nguyen 2020, section 9.2) argue that the role of mathematical formalism in modelling can be broken down in two separable questions: how mathematics is applied to a target system, and how a carrier is mathematised. Let us start with the applicability of mathematical statements to target systems. According to DEKI, we end up with a list of propositions that consist of attributions of attribution to a target system. This list of propositions then constitutes a description of the target system. Such a description can take different levels of complexity, but nothing rules out the possibility of attributing mathematical properties to parts of the target system (a population rising exponentially, the elliptical orbit described by a planet, and so on). Of course, this precisely answers the question about the general applicability of mathematics to target systems, while leaving open the question of what makes the attribution of specific mathematical property to a certain target system reasonable or justified. This will of course require a local justification for the specific model's assumptions.

As regards the mathematisation of carriers, again, for material models we endow the carrier with a theoretical interpretation that can include a mathematical description. For non-material models, the fictional system resulting from the description in a text or the depiction in a picture may well instantiate some mathematical properties too, normally specified by the principles of generation describing what we are imagining and how the imagined system (or the inferences we are allowed to perform about that system) evolve(s).

So far, I have listed the *desiderata* that Frigg and Nguyen (2018) themselves have identified in the contemporary debate, and in their book (2020) they also argue in detail that DEKI meets them. To these five criteria, I want to add a new one, namely the *dynamic nature of representation*. In the current debate, many philosophers have insisted that our philosophical analyses of models excessively objectify models: we should not talk about models so much as about *modelling*; less about representations, and more about *representing*. In short, we should move from a perspective that focuses on models and representations as objects, to one that sees them as results of processes, activities, and practices, which are more philosophically interesting and in greater need of investigation. I want to suggest that the DEKI account fits the bill also as concerns this dynamic nature of representation. All of the elements of the account delineated thus far are themselves dynamic processes or activities. The interpretation of the carrier as a Z-representation is a dynamic process carried out by an agent or a group of agents. Exemplification is static as concerns instantiation, but dynamic in its referential dimension, as it strongly depends on the context (and thus also on the purposes, questions, and cognitive abilities of the users). Like the I-function, the key is the product of an interpretive activity of the scientist that de-idealise properties of the model on the basis of their current understanding of it. Finally, imputation is also an action performed by an agent.

As a consequence of these dynamics, the identification of the target itself will be affected. The more we study and understand a model, the more precise we will be about the respects in which it functions as an accurate representation of a target system. Therefore, even denotation will in part be a dynamic process, and we can reasonably talk about *moving targets* – or moving Ts, for short.

An aspect I particularly want to emphasise is the evolution of the I-function. It seems to me quite plausible to assume that there will be a temporal dimension to the interpretation of a representation. Take for example a medical scan, or an astronomical image. A lay person, or even an inexperienced scientist, may not be able to immediately interpret these images correctly. They will certainly be able to detect colours in the astronomical image, or changes of continuity and homogeneity between different shades of grey in a brain scan. Therefore, what the picture exemplifies is in fact just the colours and their distribution. Given the cognitive nature of exemplification, only with time and practice will a beginner start to see colours *as*, say, chemical and electromagnetic properties of astronomical objects, or discontinuities

in colour in a medical scan as indicating the potential presence of a tumour. The distinct Z, then, will emerge progressively through a continuous process of adaptation of the agent to a certain interpretation of the carrier. Not only do we have moving Ts, but also evolving Zs.

The dynamic of representation in DEKI is of course not exhausted by the sum of the individual evolution of each component (interpretation, denotation, exemplification, keying-up and imputation), but is also a product of the mutual interaction between them, as well as of the interaction between the representation system as a whole with the rest of our empirical and theoretical knowledge. This is particularly so when it comes to the justification for the factual correctness of the inferences we draw from representations. If we do not have direct access to a target system, all we can do is to check whether the results of our representation fit well with other results obtained from other experiments, measurements, models, and theories. Also, nothing rules out complex, multi-system forms of representational activity, where inferences are made on the basis of the comparison and integration of more than one representational system (different models, different pictures,[41] or models and pictures together). The point here is to give a philosophical account of what can be considered a unit of representation, but of course DEKI permits the use of more than one unit within the same representational endeavour.

While perhaps implicitly entailed by the original formulation of the DEKI account, this dynamic nature of representation has not yet been made explicit by its originators. I put this reading forward as a further development of the account, and as a pre-emptive defence against potential critiques. Simply the fact that we talk about models and representations should not entail any problematic reification or excessive focus on the products instead of the activities and practices. In the same way, DEKI representation is not a static notion, and the account allows us to see how this dynamics actually occurs, as a result of the evolution of the different constituents of representation, as well as of their interaction with each other and with external factors. In what follows, then, the reader should keep in mind that my understanding of representation is always intrinsically dynamic, as illustrated above.

1.8 DEKI and its rivals

It will be helpful to point out some major features that generally distinguish DEKI from other proposed accounts of representation. This will provide a basis for some specific claims I make in the following chapters on thought experiments, model organisms and pictures, respectively. It will also be useful to respond to some anti-representational sceptics' doubts which I present in Chapter 5, insofar as much of their suspicion of representation follows from a problematic understanding of representation,

[41] I will briefly talk about this in section 4.4.2 as regards the picture(s) of a black hole.

which is different in important respects from the narrower definition offered and developed here. However, I will remain general in my review, simply pointing out the macroscopic differences between DEKI and the other main families of accounts of scientific representation. I will not delve into the detail of each contribution, nor I will elaborate on the various pros and cons of each with respect to the five conditions exposed in section 1.1. The point of this section will be simply to highlight the most macroscopic conceptual differences before turning to the application of DEKI to specific forms of representation beyond models in the following chapters.[42]

To begin with, the account does not ground the relation of representation on the concept of similarity (see, e.g., Giere 2004 and Weisberg 2013),[43] nor on its mathematical relatives – namely any form of mathematical morphism: isomorphism, which is the most common, partial isomorphism, or homomorphism.[44] I concede that, in some instances of representation, some of a carrier's actual properties are interpreted as themselves via the I-function, they also turn out to be exemplified, and they are also imputed to the target system unmodified via an identity key. In such special cases, we can talk about sharing of properties, and thus of similarity, between the model and the target system in the relevant sense. However, even in these cases, the mere sharing of a property is not in itself sufficient: the property has to also be exemplified by the model. Exemplification requires reference to the property, thus requiring an interpretation of that property as salient with respect to other properties. This interpretation can of course depend on objective facts of the matter, as for example the fact that a property, even though instantiated, is not easily detectable in the representation. Elgin (2010, pp. 11-12) offers an example by reinterpreting a case study from Tufte (1997, pp. 17-31). Before the launch of space shuttle Challenger on January 28, 1986, NASA engineers were asked for an assessment of the risks associated with the launch. The summarising tables in the engineers' report did in fact contain evidence about the vulnerability to cold of the so-called O-rings, an important component of the shuttle. However, this crucial information was drowned out by the huge amount of other data present in the summarising tables: the information about the O-rings had not been made salient by the engineers. When the shuttle was launched, it exploded (causing the death of the people inside) because of the effects of low temperatures on the O-rings. This episode illustrates (indeed, exemplifies) the distinction between mere shared instantiation and exemplification: even if the engineers' reports showed a certain property of the target system, they did not highlight it relative to other properties.

[42] The interested reader can find a more comprehensive evaluation of the main proposals in the debate in Frigg and Nguyen (2020), where the authors explicitly assess the main families of views on representation against their own proposal.

[43] Although both authors couple similarity with some reference to a user, they still consider similarity as the relevant core concept in order to understand epistemic representation.

[44] Examples of such accounts of representation can be found in Da Costa and French (1990, 2000), French and Ladyman (1999), and van Fraassen (1980).

The same applies, *mutatis mutandis*, to mathematical-structural relations. It is often the case that the I involves a mathematical description of the carrier, and thus the Z-representation will be able to exemplify certain mathematical properties, which in turn are imputed to the target system via the key. Still, which isomorphism is chosen and how it is applied requires interpretive activity on the part of scientists, and is just one way to connect models and targets, although a particularly common one given the flexibility and generalisability of mathematical formalisations.[45]

As I showed above, DEKI remains silent about the justification for our inferences from representations to target systems. While the account is accordingly compatible with the idea that our inferences may be in part justified by some actual similarity or isomorphism between the representation and the target system, it does not require it. We have seen with Elgin's example that sharing a property, even a structural one, does not necessarily entail a felicitous inference. The very existence of the key in DEKI shows that sharing a property is not necessary either: the property exemplified by the system could require a translation in order to be applied to the target system. Of course, a structuralist could insist that what is preserved in the key translation is some mathematical structure. However, even in a case where the key was a structure-preserving function, the question is *which* structure we want to use to describe the model, and consequently the target. This is however a matter of choice, because the same object can be described in many different structural ways, each of which provides a different perspective on that object, with different inferential paths downstream.[46]

Moving away from the views based on similarity and mathematical structures, the originators of DEKI endorse a form of antirealist fictionalism about models[47] inspired by the work by Walton (1990) on artworks and artistic imagination as a result of games of make-believe. The fictionalism of DEKI specifically concerns the Zs, and in this way distinguishes itself from forms of fictionalism about models (see e.g. Toon 2012 and Levy 2012, 2015). These accounts are also inspired by Walton's theory, but tend to interpret models as *direct* descriptions of their targets that just happen to be fictional. DEKI instead follows Weisberg's (2013) lead and take the relation between models and their targets to be indirect: we describe and reason about a stand-in system, and only as a consequence do we then try to export our results to another context, namely the target system. Even though my focus is on representations and not on models specifically, it should be clear how this difference in understanding fictionalism has repercussions for the representational relations between models and

[45] Against a structuralist understanding of representation, see also Suárez (2003) and Frigg (2002).

[46] See Frigg (2002, pp. 30-31) and the example of the methane molecule therein.

[47] For a discussion see Frigg (2010), Frigg and Nguyen (2016), and Salis, Frigg and Nguyen (2020). This ontological aspect of the account is quite tangential to my main focus on the semantic and epistemic aspects of epistemic representation. However, I will talk more about this when I turn to scientific imagination in Chapter 2 in order to give an account of the activity of imagination involved in scientific thought experiments.

their targets.[48]

Finally, the DEKI account does not reduce to an inferential account of representation (see Hughes 1997, Suárez 2004, and Contessa 2007). Inferentialists focus on the capability of models to afford (justified) inferences about their target systems. Therefore, these accounts seem to take the surrogative reasoning condition as analytically constitutive of representation. As I have shown above, DEKI gives us an account of surrogative reasoning via representation, but the inferential potential of representations is not built into the concept of representation, but is rather a consequence of the interpretative aspects of the account and the referential relations connecting both the model with the target (denotation) and the model with the exemplified properties (exemplification). Thus, within the DEKI framework, inferential power is not an irreducible, primitive notion, but it can be decomposed and explained by interpretation, exemplification, and the key.

The difference between inferentialism and DEKI, and particularly how incompatible the views actually are, is strongly dependent on the details of the inferential account one chooses to adopt. For example, Suárez's (2004, 2024) deflationary account seems intrinsically recalcitrant to any reconciliation with more substantial accounts.[49] Some inferential accounts of representation, like those of Contessa (2007) and Díez (2020), have attempted to enrich the inferential view by introducing some form of interpretation. However, their sense of interpretation is characterised in different terms to that of DEKI.[50] In Chapter 5, I will return to inferentialism, specifically a recently proposed inferential account of representation, proposed by Khalifa, Millson and Risjord (2022). There, I will both emphasise the radical differences between their account and the DEKI account and argue the superiority of the latter.

1.9 Summary of the chapter

This Chapter has defined the scope of the question and reconstructed and developed the basic tenets of the DEKI account of scientific representation. Building on the fundamental work by Goodman and Elgin, Frigg and Nguyen's DEKI recognises the crucial role of interpretation in any component of epistemic representation, from denotation to exemplification (and the selectivity thereof), to the use of de-idealising keys. This basic point has been illustrated by paradigm examples of scientific models. The next step is to show how to apply this account to other forms of scientific representation. In the following chapters, I will apply DEKI's conceptual toolkit to three areas of philosophical debate: thought experiments in physics, organism

[48] For more details on this difference in the application of Walton's fictionalism to scientific models, see Frigg and Nguyen (2016).

[49] For a critical analysis of Suárez's views from the perspective of the DEKI account, see Nguyen and Frigg (2022b).

[50] For a critical analysis, see Frigg and Nguyen (2020, pp. 95-105).

specimens in biology, and mechanically produced pictures in astronomy. I aim to show not only that the DEKI framework helps us reconstruct these representational practices, but also that it can shed light on current semantic and epistemic issues, helping us solve some troublesome puzzles in the specialised debates on thought experiments, model organisms and pictures.

Chapter 2

Putting the 'experiment' back into the 'thought experiment'

2.1 An inflamed debate

In the last decades, philosophers of science have debated the epistemological status of scientific thought experiments (TEs):[51] how they function, what is their role in scientific investigation, and whether we can learn from them about the real world.

The aim of this Chapter is to provide a full-fledged account of the semantic and epistemology of TEs in science. In order to do this, I look at the work that has been done in philosophy of science on material experiments and on scientific models. From the first area, I employ the traditional distinction that philosophers of experimentation draw between internal and external validity of an experiment and show that the same distinction naturally applies to thought experiments as well. From the literature on models (see Chapter 1), I retrieve the concept of representation and argue that it allows us to understand the relation between TEs and the world and shed light on the epistemic use of TEs in science overall.

The plan for this Chapter is as follows. In this section, I reconstruct the major positions in the debate on thought experiments in science and how we learn from them about the world. In section 2.2, I propose to reset the investigation on TEs on the basis of a structural similarity between thought experiments and material experiments (MEs). Specifically, I argue that we need to focus our attention on the distinction between the internal and the external validity of a thought experiment. Then, in

[51] If not specified otherwise, I restrict myself to TEs used in science, and my examples refer to TEs employed in physics. Yet, work has been done on the role of TEs in other scientific disciplines, such as economics (Schabas 2018, Thoma 2016) and biology (Schlaepfer and Weber 2018), as well in not entirely empirical disciplines like mathematics (Starikova and Giacquinto 2018; see also examples of geometrical TEs in Brown 2004) and philosophy. I hold that my argument works with all kinds of scientific TEs. Normative TEs – i.e., TEs the outcome of which is an ought-to type of statement – definitely require further specification. Philosophical TEs, as long as they remain purely descriptive, will also be treatable in my framework. I talk more broadly about this in section 2.2.3.

section 2.3 I offer a more detailed account of both these types of validity in thought experiments. I propose to analyse their internal validity in terms of Walton's games of make-believe, and to interpret external validity in terms of accurate representation. Finally, in section 2.4 I go back to the current debate and show that, thanks to my reconceptualisation, the numerous positions presented in section 2.1 can be explained and then reconciled with each other to a considerable extent.

2.1.1 Kuhn's questions

Scientific TEs[52] began to receive considerable attention from philosophers of science in the wake of Thomas Kuhn's provocative question (1977): what is the epistemological status of TEs in science, given that they are apparently performed only in scientists' mind? Kuhn's question can be broken down in two sub-questions, which allow us to offer a first general taxonomy of philosophers' views on TEs:

1. Do TEs provide new knowledge about the empirical world?

2. If so, how do they do so? And if not, why not?

The first question allows us to divide the positions in the debate up into two main camps, which I call the yes-camp and the no-camp. Positions in the first camp hold that we can achieve new knowledge via TEs and try to explain how this is possible. The positions in the opposite camp contend that TEs cannot provide new knowledge about the world and give reasons to show why so. Within each camp, distinct positions can be further qualified by looking at how they answer the second question. This reconstruction of the many positions in the debate cannot be and does not aim at being exhaustive. Yet, framing the discussion of TEs from this particular perspective allows me to introduce and develop the issues of validity in the context of TEs.

At this point of my discussion, I do not need to commit to any particular theory of knowledge. As we will see, philosophers also introduced different concepts of knowledge when it comes to the one obtained via a TE, so I do not want to commit from the start to a specific notion of knowledge and remain as liberal as possible. A proper concept of knowledge will emerge naturally from my analysis, and I will come back to the nature of the knowledge obtained from TEs, and specifically the question about its justification, in section 2.3.4. For now, I take the term 'knowledge' to refer to whatever notion of knowledge is appropriate for science. The only constraints are that the purported knowledge concerns empirical facts, and that 'new' here is to be interpreted as the fact that the scientist acquires knowledge that they did not possess before performing the TE.

[52] For a general overview, cf. Brown and Fehige (2022); Stuart, Fehige and Brown (2018); and Frappier, Meynell and Brown (2013).

2.1.2 The yes-camp and the no-camp

In the previous section, I said that the philosophical positions in the debate can first of all be divided between those who argue that TEs do in fact provide new knowledge (the yes-camp) and those who deny it (the no camp). Let us look at the main views on the yes-camp first, and then move to the accounts of TEs belonging to the no-camp.

The yes-camp usually stems and acquires momentum from a historical perspective on TEs and their manifest role in modern science, from Galileo to Einstein, passing through Newton, Maxwell, and many other scientists. The yes-camp's leading intuition is that TEs have importantly contributed to substantial scientific progress. Further, as a springboard for the development of their views, most authors in this camp have emphasised, though in different ways, the peculiar role of the imagination in science and how it would allow scientists to go beyond previous empirical knowledge. Among these positions, we can identify three main views, which I call for short Platonism, objectualism, and structuralism. Let us now see how they answer to the second of Kuhn's questions, namely how scientists do in fact obtain new knowledge from thought experiments.

Brown (1992, 2004, 2011) argues that TEs can transcend the empirical knowledge that we possess before performing them. He contends that TEs contain an element of "a priori intuition" (2004, p. 31) that can neither be reduced to logical inference nor to the empirical knowledge already possessed by the scientist.[53] Particularly, TEs would allow us to "see" some of the general laws that govern our world. TEs are thus instances of (potentially) powerful intuitive reasoning, which can make us immediately achieve some a priori content of knowledge about the world without directly interacting with it. Brown further characterises TEs' intuition in terms of a direct act of "perception" or "seeing", performed not through our senses but by our mind itself, and directed to abstract objects, and specifically the very laws that govern nature (Brown 2011, p. 98).

Platonism is not the only option to explain how TEs provide empirical knowledge. Miščević (1992) and Gendler (2004), for example, focus on the objectual, picture-like nature of the imagination involved in TEs. This imagistic dimension of TEs, these authors contend, cannot be entirely reduced to a set of propositions, in the same way as a picture's dense content cannot be reduced to a linguistic description. A picture, the authors suggest, has usually a character of holism and a level of repleteness[54]

[53]The notion of intuition as a source of knowledge alternative to experience well fits the long-standing tradition of rationalism (see Markie and Folescu 2021, §2).

[54]The concept of repleteness here is to be read in the technical sense offered by Goodman (1976). Without delving too much in the very articulated theory of symbol systems that Goodman proposes, a symbol system is more or less replete depending on how many different aspects of a symbol can have an impact on its meaning. Goodman's example is to compare the line of a fever chart with the line of a minimalist painting of a skyline of a mountain chain. Even if very similar in their appearance, the former is less replete than the latter. In the latter, we can look at many aspects of the line – e.g., its width, the colour used, if the line is continuous or not – and all of them has semantic import for our understanding of the painting. In a fever chart, instead, what matters is

that makes it qualitatively different from propositional sources of information. So, by visualising an imaginary system, given its picture-like nature, we would acquire knowledge about how that very system would behave in the real world in a way that is fundamentally different than by deriving it from a theory or a mathematical model.

In a sense, the quasi-perceptual dimension of TEs is here contrasted with the acquisition of knowledge that occur by derivation and proposition-based inferences. The objectual nature of the TEs' imagination is then also what would make TEs useful instruments of scientific investigation and would explain how they provide new epistemic content.

The third view in this camp is what I call structuralism and has been proposed by Nersessian (1992, 2007, 2018). Her account of thought experiments is still an overtly non-propositional view on the TEs' imagination. However, her characterisation of the imagination does not involve mental pictures or objectual visualisation. According to Nersessian, TEs are instead to be understood as instances of a *simulative model-based reasoning*, where a model is defined as a "structural analog" (1992, p. 293) of the scenario targeted by the TE. In Nersessian's terminology, a structural analog is a system isomorphic to the modelled system with respect to spatiotemporal and causal relations.

I want now to suggest that among the accounts of the yes-camp just illustrated, none has been entirely successful in explaining the epistemic role of TEs in science. To begin with, the Platonic account appeals to somewhat mysterious processes of mental vision and a priori truths acquired via intuition. On such an account, it remains unclear how we can perform TEs that lead us to false results (Norton 2004). However, the history of science abounds of TEs the results of which are wrong. In response, Brown could say that, like sensorial perception, Platonic vision can go wrong too. As in the case of optical illusions, we would be sometimes victim of "intellectual" illusions as well.

The problem with this reply is that, while we have a scientific theory that explains how and why the optical and in general perceptual illusions or errors occurs, we have nothing comparable in the case of TEs. This is because intuition, to which Brown mostly appeal as a way to get epistemic access to laws of nature via TEs, works usually as a black box: in contrast with observation and logical inferences, which can be explained as chains of causal events or logical steps, respectively, no analysis of the sort can be offered for intuition, which is normally conceptualised as an automatic and holistic process. Then, the distinction between good or bad TEs resulting from Brown's appeal to Platonic illusions, analogous to the sensorial ones, risks being dangerously *ad hoc*, as there would not be a theory that provides principled reasons for why sometimes we perceive true laws of nature and sometimes we just get it wrong. It would simply be a matter of good or bad intuitions, which however does

just the coordinates of each point of the line with respect to a frame of reference.

not solve the issue but ends up stating it again in new words.

Let us move to objectualism. While the view is intriguing in importing our understanding of perception and picture-like symbol systems, it still does not explain clearly why this objectual nature of the imagination involved in TEs would make an epistemic difference to answer Kuhn's question of how we acquire knowledge about the world. TEs' perception-like nature may certainly make them effective in a pedagogical or heuristic sense, and their picture-like quality may offer different information, perhaps richer and more holistically integrated, than the one conveyed propositionally.[55]

However, this proposal does not seem to provide any new insight on the acquisition of new empirical knowledge via TEs. For it is not real perception, just imaginary one. And even if pictures and images have a different semantics than logically organised propositions, they are still created by our minds. So, either we picture things we already know, but then there is no epistemic gain through TEs; or the objectual imagination allows us to achieve new content of knowledge, but this leads to exactly the same issues that we encounter when we elaborate new information from propositions. In other words, the fact that the type of symbol system (in this case, a pictorial, imagistic one) involved is different does not immediately provide an answer to our question of how we get new knowledge from TEs. We still have to specify on what basis the results obtained via the imagination are valid. All things considered, it would seem that there is not much of a difference between objectual and propositional imagination in this respect, and the burden to prove that an objectualist treatment is more successful for our purposes than a propositional treatment lies entirely on the objectualists' shoulders.

More generally, the appeal to non-propositional forms of the imagination does not seem to answer Kuhn's second question satisfactorily, because from the point of view of knowledge and its validity (and not, say, feeling), any imagined non-propositional content has the same problematic epistemological status as an imagined propositional content.

Nersessian's view seems to make some progress, as she states tha the missing link between what is imagined and the real world is provided by some structural analogy between the thought experiment scenario and some real world phenomena.

However, this approach requires an element of structural isomorphism between TEs and real systems that is problematic, unnecessary, and insufficient. Problematic, because it is difficult to understand why one should perform the TE in the first place, if one already knows the structure of the modelled system, and the structure is all

[55]Linguistic descriptions, mathematical formulae, diagrams, maps, photographs, graphs: all these types of representation have their own semantics, and different ways to convey information. For example, pictures are semantically dense in Goodman's (1976) terminology, and these make them particularly rich representations. See also Camp (2007) for maps, Rescorla (2009) for cognitive maps, Perini (2013) for diagrams, and Perini (2010) for a more general taxonomy of pictures' semiotics.

we need. TEs would then just become pedagogically useful ways to illustrate more abstract, structural aspects, which would make all the actual epistemic work.

The appeal to structural analogy seems also unnecessary, because, at least *prima facie*, there are thought experiments that do not seem analogous to any phenomena in the respects relevant for the epistemic use of the thought experiment. Without further specification, Newton's two spheres orbiting in an empty universe is not analogous to any real system in the world. Same thing happens for Galileo's two bodies falling from a tower according to Aristotelian theories of falling, or Einstein's lift placed away from any gravitation field.[56]

Finally, the account seems also insufficient, because there are many different (and often irrelevant or misleading) ways a system can be structurally analogous to another one. The problem then is to find the *right* structure, among the many that are applicable to the same phenomenon.

In contrast with the positions in the yes-camp just briefly illustrated, philosophers in the no-camp have argued that TEs neither provide new knowledge about the empirical world, nor they actually can. Let us then introduce some views in this second camp.

On the extreme of the spectrum, we can place Dennett's view (1996) that TEs are simply "intuition pumps". Dennett seems to start from the same characterisation Brown gives of TEs, namely them being forms of intuitive reasoning, but he arrives at a radically different conclusion, plausibly given a quite different attitude towards intuition itself. He holds that TEs "are not arguments, they're stories". Instead of having a conclusion, "they pump an intuition" (p. 182). Exactly for this reason, they do not provide new knowledge about the world: they can just provide the initial intuition that needs to be supported by arguments and observations.

A similarly minimalist view is held by Hacking when he says that TEs "can reveal tensions between one vision of the world and another" (1993, p. 307) and that they have no function beyond that. On this view, TEs are simply instruments to put our intuitions about reality into words or images and make the tension between them become apparent. Hacking also emphasises the limit of TEs by a comparison with material experiments. He contends that latter tend to acquire independence from the experimenter, in that they offer new information about the world because they are themselves part of the world, they make it possible to perform new experiments, and they possess an element of surprise, as we do not have complete knowledge of the processes and mechanisms involved. TEs, instead, are inevitably instances of the agent's imagination, and their relation to reality is mediated by the ideas and intuitions of the agent. Because of this dependence on the agent performing them, their epistemic import remains seriously limited.

The answer that Dennett and Hacking give to the second question, i.e., why TEs

[56] I will provide more information on these TEs and the relevant references in section 2.2.2.

cannot provide new knowledge, is not entirely satisfactory. Dennett just stops at the intuitive element and does not focus on the often complex elaboration of information usually involved in thought experimentation. Eliciting intuitions does not seem the only thing we do when performing TEs. For we want to employ TEs to understand whether these intuitions are correct or not, both in philosophy and in science. On his part, Hacking bases his argument on a comparison with MEs, but this is not going very far either: even admitting that TEs are not as independent of the experimenter as MEs, this does not entail that they cannot provide new knowledge about the world. Furthermore, Shinod (2017) has provided compelling arguments against Hacking's argument, showing that TEs can exhibit a life on their own in Hacking's own terms, and they often produce surprising results, like material experiments do.[57] In general, further unpacking is needed.

Norton (1991, 1996, 2004a, 2004b), one of key contributors to the debate, and plausibly the champion of the no-camp, provides a more robust answer to Kuhn's second question. He starts from the empiricist premise that knowledge can be acquired only in two ways: by observation or by logical argument. As TEs do not involve the former, they must rely on the latter. Norton thus argues that TEs are logical arguments, presented in a picturesque fashion. Thus, TEs can also manifest interesting features of the real-world if they exhibit empirically grounded premises. However, they do not actually provide *new* knowledge about reality, as their empirical content was in fact already contained in the premises. TEs then turn out to be a rhetorically effective way to select and logically organise empirical premises in a salient way. Norton further defends his position by showing that one can effectively reconstruct successful TEs in the form of logical arguments with empirical premises. TEs epistemology turns out to be quite simple in Norton's eye: a TE fails when it "is not sound", i.e., when the underlying argument exhibits at least one "false premiss or a fallacious inference" (2004b, p. 51).

Although Norton's account is compelling, it has some problematic implications,[58] the main one concerning the empirical premises required by his approach. For his account to work, we need TEs to be reconstructed as sound arguments about empirical facts. In order to be so, they need empirical, non-trivially true premises. However, it seems that many TEs involve unobserved, and sometimes unobservable things: someone running at the same speed as a beam of light, a lift placed away from any gravitational fields, minuscule demons able to move molecules, or, worse, falling compounds where a lighter object acts as a brake on a heavier one. Of course, Norton can always offer an argument where these imaginative whimsies are swept away and still achieve the same results that we obtained via the TE via an argument that displays only empirically true premises. But then one has the impression that we

[57] On TEs and surprise, see also French and Murphy (2021).

[58] For an exhaustive analysis of Norton's view on TEs and the criticisms raised against it, see Brendel (2018).

are not talking about TEs anymore, as they have been substituted by something else. Then, the fact that TEs, and not those argument reconstructed *à la* Norton, allowed scientists to formulate new ideas and gain knowledge starts looking like cases of epistemic good luck.

The last position from the no-camp I want to introduce here has been defended by El Skaf (2018, 2021). Here, Hacking's characterisation of TEs as expression of a conflict between different visions of the world receives more qualification, coupled with the simultaneous effort to avoid the most problematic issues of Norton's account. For El Skaf, TEs' primary purpose is to detect and sometimes resolve inconsistencies in our scientific theories. This can be achieved both intra-theoretically – when the TE is grounded in the tenets of a single theory – and inter-theoretically – that is, when the principles of two or more theories are combined. As an example of the former, El Skaf appeals to Galileo's TE on falling bodies, used by the Italian physicist as a refutation of the dominant theory of falling bodies defended by Aristotelian physicists of that time.[59] As an example of the latter, inter-theoretical use of TEs, El Skaf proposes Bohr's (1949) objection to Einstein's own TE against the indeterminacy principle, a cornerstone of the newborn theory of quantum mechanics. Bohr manages to show that Einstein's TE fails because it assumes a classical framework, instead of taking into consideration the implications of relativity. Once a relativistic framework is assumed, the TE's results are again compatible with the indeterminacy principle.

For El Skaf, TEs do not in fact provide new epistemic content about the empirical world, as all the premises of the thought experiment's scenario are derived from, or at least compatible with our scientific knowledge. TEs are then theoretical devices available to scientists to check, refine, amend, and sometimes abandon, their own scientific theories. Thus, either TEs' results remain *internal*, so to speak, to the theoretical domain, or they constitute possible models of those theories. In both cases, TEs concern theories, and they are concerned with the empirical world indirectly at best. So, we do not have to meet Norton's strong requirements of empirical truth: all we need is what our scientific theories tell us.

There are two problems with this view. First, it would seem controversial that *all* TEs are about theoretical inconsistencies. There is no inconsistency in Maxwell's demon, just a question about the nature of entropy. In Newton's thought experiment of two spheres rotating in an empty universe, there is no contradiction to be detected, but rather an abductive argument for the existence of absolute space. In Einstein's TE of a scientist trapped in a lift, the point is to show that uniformly accelerated motion is identical to the motion of bodies in free fall, not that their distinction leads to inconsistencies. The second problem is the relation between the TEs and theories, on the one hand, and between TEs and reality on the other. The purpose of TEs seems to tell us something about the empirical world, and in order to do that, they

[59] Galilei (1638, p. 107 of the original, p. 62 of the English translation).

sometimes have to distance themselves from our established theories.

2.1.3 Internal debates and general problems

The positioning of authors with respect to Kuhn's questions has produced internal sub-debates between the protagonists of the discussion. First, Brown and Norton have discussed at length whether the knowledge achieved via TEs transcends empiricism – see, e.g., Brown (2004) and Norton (2004b). While Norton holds that TEs are just logical arguments in disguise, to be filled with available empirical knowledge, Brown argues that there is something more. Logic and empirical premises cannot account for the results we can achieve via thought experimentation, which forces us to postulate some form of a priori knowledge that is achieved via intuition.

Another discussion focused on the nature of the imagination in TEs, asking whether it is propositional or objectual – or, for Nersessian, at least not entirely propositional. This discussion revolves, again, around the epistemology of TEs. The arguments of the non-propositional view on the imagination involved in TEs are motivated by the fact that propositions are not enough to explain the epistemic power of TEs, and their effective use in science.

Finally, as regards El Skaf's account, there is a question about the relation between TEs and scientific theories. He seems to hold that, if we learn something via TEs, it concerns what our theories imply. TEs' epistemic value would then strictly depend on theories, either because they inform relevant aspects of them, or because they show their shortcomings or problematic consequences. Naturally, this is also an attempt to answer the question concerning empiricism: there is actually no issue here, because TEs remain internal to the theoretical background. It seems also the most viable path for the no-camp, as it restricts TEs' knowledge to a purely theoretical one.

Now that the main positions have been introduced, and the sub-questions presented, I want to highlight two general problems. First, from the point of view of content, no account seems completely satisfactory. While the yes-camp remains quite obscure or vague in answering the second question, the no-camp risks to discard the epistemic value of TEs too quickly. I am not in the business here of offering definitive, knock-out arguments against each single contribution, but rather to suggest that all of them present serious issues that have to be addressed.

Second, and more importantly, there seems to be a problem in the structure of the debate overall. As has become clear from our presentation of the state of the art, the problem with the philosophical debate on TEs is not only due to a plurality of epistemological accounts proposed. Rather, the different perspectives on TEs generate philosophical consequences in sharp contrast with each other. This should be clear from the sub-debates that I have just highlighted, where different authors arrive to opposite conclusions. All things considered, the general impression that neophytes to the discussion on TEs in science inevitably have when they approach the debate is that there is little common ground shared by the discussants.

I contend that the reason for such a fundamental disagreement on TEs is that the community still lacks a solid theoretical framework shared by all participants. I suggest that a fruitful way to improve the current situation is to focus on the experimental nature of thought experimentation, and more specifically, on a fundamental but so far largely overlooked characteristic shared by both TEs and MEs: in both kinds of experiments, it is crucial to distinguish between the internal and the external validity of the experiment.

This distinction is compatible in different ways with many views in the debate, and even anticipated in some works on TEs. Nevertheless, it has not been made explicit yet, or remained considerably underdeveloped. In the next section (2.2, after I have illustrated the general distinction for MEs, I show that it naturally applies to TEs. I conclude the section by showing the novelty of this investigation with respect to other analyses of the experimental nature of TEs.

Once we put back the 'experiment' back into the 'thought experiment' by explicitly acknowledging the distinction between these two types of validity, and once we appreciate the deep consequences of this distinction, we acquire a more general conceptual framework, so that we can better understand how the various positions in debate relate to each other. Consequently, the general epistemology of TEs becomes more limpid and more strongly related to the epistemology of experiments and scientific modelling.

2.2 Two kinds of experimental validity

2.2.1 Validity in material experiments

There are two ways in which a material experiment can be valid: internally and externally. Campbell (1957) was the first to formulate such a distinction, and he associates internal validity with whether the experimenter's intervention is causally responsible for the observed outcome(s) occurring in the experimental system. By contrast, external validity concerns whether that very outcome can be generalised to other settings, and if so which ones.

More precisely, *internal* validity is the correctness of the experimental result within the original setting, which is studied by the experimenter under specified boundary conditions. Given an experimental system S, which will usually include a description of a relevant context (initial and boundary conditions) C[60]:

[60]Even though it can often be arduous to clearly delineate the specification of the initial and boundary conditions of a certain experimental setting, a conceptual separation between initial and boundary conditions, on the one hand, and objects and their own dynamics, on the other is nevertheless useful to demarcate different settings, as they may differ from each other because of different conditions, different objects involved, or both.

Internal Validity$_{(\mathbf{def})}$: An internally valid experimental result R is a proposition made on the basis of the observation of and/or an intervention on S, which correctly ascribes a certain property P to S.

Thus, internal validity implies that the observed properties do not depend on (i) the experimenters' mistakes or subjective biases, (ii) faulty measurement devices, (iii) a misinterpretation of the data (e.g., by taking a correlation as a direct sign of causation), or, most importantly, (iv) confounding factors, both expected factors and ones that had not been considered in the first place – or any combination of these reasons.

An experiment is *externally* valid when the internal results are successfully applied, or extrapolated, to other scenarios beyond the original experimental system. Extrapolation here should be understood in a very general sense: induction from one specimen to other specimens; predictions about the outcome of an intervention when performed in a different context; hypothesising similar causal patterns in a new system; a statistical inference from a sample to a larger population.[61]

Considering an experimental system S as defined above, and introducing a target system T external validity can be defined as follows:

External Validity$_{(\mathbf{def})}$: An experimental result R is externally valid with respect to a designated target system T *iff* R is a proposition derived from the observation of, and/or an intervention on, an experimental system S, and R is true about T.

Thus, the external validity of R is always relative to a specific target system T, which can differ from the observed and/or manipulated system S on the basis of the objects involved, the surrounding context, or both.

It is important to distinguish between these two types of validity because, clearly, internal validity does not entail external validity. A result can be correctly achieved in an experimental setting, but this is not in itself sufficient to predict that the same result will obtain in another experimental scenario.

While the separability of the two types of validity is a general feature of experimental extrapolation, it has been emblematically acknowledged for so-called Randomised Controlled Trials (RCTs), a very common methodology employed particularly in medical research. A RCT normally consists of two randomly created groups of individuals, the test group and the control group. A certain intervention is performed on the test group, while no intervention (or a placebo sort of intervention) is performed on the control group. If the intervention makes a difference in the test group relatively to what happens to the control group, we have a RCT result. The specifics will depend on the qualitative and quantitative difference produced by the intervention in the

[61] Christensen and Waraczynski (1988, §4) propose a similar articulation of different types of external validity.

test group with respect to the control group.

RCTs are undoubtedly accepted as a gold standard for internal validity, because their randomised sample selection usually succeeds in cancelling confounding causal factors out. The randomisation is normally taken also as a feature that makes the RCTs' results exportable to other contexts. The intuition is that, exactly because the selection of the members of the groups are random, a result of a RCT should not depend on the specifics of the group chosen. For this very reason, however, it is also difficult to determine why a RCTs result was not extrapolable to other specific contexts. This is because, given the randomisation, we usually ignore what factors could be confounding or not (cf. e.g., Cartwright 2010). This results in serious issues for establishing the reasons of failures of external validity.[62]

Consider the well-known case of the drug Benoxaprofen: although this drug had proven effective in several RCTs, it later turned out to be harmful when applied to the actual target group of patients (Worrall 2007, pp. 994-995). So, we have an internally valid RCT result that yet was not externally valid for the intended target population that would have received the drug. The explanation of this specific mismatch was that while the tested groups included individuals of many different ages (indeed, a true random sample of the entire population), the actual target population was mostly composed by old people. Thus, the percentage of side effects, quite negligible in the random trials, was significant in the group of elderly people outside the laboratory walls.

This is a clear example of a case where an internally valid result is not externally valid in the designated target. There is of course nothing logically inconsistent in imagining the reverse case: an experiment mistakenly considered as internally valid that turned out to be externally valid. This is the reason for which, in my definition of external validity above (2.2.1), internal validity as a necessary condition for the external one. However, it would certainly be unwise to build an experimental methodology from this sort of cases, where external validity is acquired by mere luck. So, even if conceptually possible, I will not further consider the cases of experimental results that are externally valid but not internally so.

Another example of a mismatch between internal and external validity has been discussed by Cartwright (2012) and concerns the application of a plan for improving child nutrition. A policy focused on providing nutritional education to mothers had been introduced in a Northern region of India and was showing very promising results. Then, an attempt was made to export the same strategy to a similar problem of child nutrition in Bangladesh. This time, the intervention was a failure. The reason was that in Bangladesh mothers had less control over food shopping than the mothers in Northern India, where these decisions were under the control of their

[62] See also Stegenga (2015, pp. 68-69), with which though I disagree concerning potential solutions to extrapolation problems and publication biases (see Hoefer and Krauss 2021).

mothers-in-law.[63]

Since Campbell's reflections, the distinction between internal and external validity has generally framed the debate about the methodology of experiments.[64] The upshot of this excursion into MEs is that the there are two distinct types of validity of an experiment, and we have to keep this distinction in mind when it comes to understanding what knowledge an experiment can provide. Let me call this methodological thesis the Internal-External Validity Distinction (IEVD). I now intend to apply IEVD to the current philosophical debate on scientific TEs.

2.2.2 Validity in thought experiments

I now give an example to show that it is useful to apply IEVD to TEs as well. Then, I will argue that when the distinction is in place, the best way to understand a TE's internal validity is in terms of games of make-believe, while the best way to interpret a TE's external validity is in terms of accurate representation.

In his *Dialogues Concerning Two New Sciences*, Galileo employs a TE to demonstrate the so-called law of equal heights.[65] Imagine a V-shaped cavity with a bottom that approximates a curve, to allow a ball to roll smoothly between the planes. Galileo demonstrates that if the ball rolls down one plane of the cavity, it will reach the same height on the other plane. This result will occur independently of the relative inclination of the two planes. For example, if one imagines bending one of the planes downward, leaving the other side unchanged, the ball will still reach the same height on the former side.

In modern physics, this follows directly from the conservation of energy principle. Galileo, instead, arrives at the law of equal heights via a long series of demonstrations, which concern the general features of uniformly accelerated motion. It is crucial to note that in both cases, the derivation of the law still strictly depends on some important idealisations. First, like most of the results obtained in Galileo's work on kinematics, this law depends on assuming the total absence of friction: the ball is perfectly spherical, the inclined planes are hard and smooth, and the air provides no resistance (Galileo 1638, p. 166 in the original, p. 170 in the English translation). Second, specifically to this TE, once the ball reaches the vertex of the cavity and changes inclination of motion, it moves *as if* the conjunction of the two planes formed a curve. This second distortion remains implicit in Galileo's reasoning, but it is necessary because he restrains his analysis to accelerated motion along straight lines.

[63] For other examples, see Cartwright and Hardie (2012).

[64] In experimental economics, cf. Guala (2005, §7) and Cartwright (2007, §15), who has also anticipated the same problems in physics (1983); in psychology, see Berkovitz and Donnerstein (1982) and bibliography; in biochemistry, see Strand *et al.* (1996).

[65] Cf. Galilei (1638), *Third Day*, Proposition XIII, p. 208. Particularly, see the comment to the figure at page 210. In the 1954 English translation, see pp. 216-218. This TE is also discussed in Sorensen (1998, pp. 8-9) and Salis and Frigg (2020, pp. 20-21).

Immediately before presenting the TE, Galileo introduced what we now call the *law of inertia*, which asserts that if no force acts on a body, then that body will either remain at rest or keep moving with constant velocity. This was a significant change of paradigm in physics, as people previously thought, in a generally Aristotelian theory of motion, that a moving object would grind to a halt once the cause of its motion ceased. The TE shows how the law of inertia follows from the law of equal heights. For, if one continues to bend, say, the left side of the plane indefinitely, we obtain a case where it is actually horizontal. The law of equal heights states that a ball dropped on the unchanged right side will not stop until it reaches, on the left side, the same height from which it fell. However, given that the left plane is now completely flat, the ball never reaches the same height, and thus it must keep moving indefinitely. So, an object can and does move with uniform velocity with no force acting on it. The law of equal heights, then, is an intermediate inferential step to demonstrate the law of inertia.

This perfectly fits my definition of internal validity in section 2.2.1: the law is an internally valid result as it correctly ascribes properties to the experimental system described by the thought experiment. However, the same result is not universally applicable, because in our world objects do not move perpetually, and a ball rolling down one side of a cavity will not reach exactly the same height – only approximately so. Therefore, there is clearly a difference between a result that is valid within the thought experiment's scenario, and result which is also valid in a different context. Most importantly, the internal validity of the result, as it is evident in the case of Galileo's TE, does not *per se* entail external validity in any target T.

Galileo's case is no exception, and I contend that many other famous TEs can be fruitfully analysed through the lens of IEVD. For example, Maxwell considers a minuscule demon that is able to separate fast molecules from slow molecules in a box of gas.[66] The demon concentrates all of the fastest molecules in one half of the box, thus creating a considerable difference of temperature between the two sides of the box. This scenario exemplifies a reduction of entropy in a closed physical system, and therefore counts as a counterexample to the second law of thermodynamics. This further imply that the second law is at most statistical: it is possible, just highly unlikely, that entropy decreases in a closed system. However, even in the case that the result were valid in the TE, the applicability of this result to the external thermodynamic phenomena would not immediately follow – there are no such things as Maxwellian demons in our world.

Similarly, Newton asks us to imagine two spheres tied to each other with a rope in an otherwise empty universe.[67] The spheres are assumed to rotate around the

[66] This TE was mentioned in a letter to Tait in 1867 (cf. Knott 1911, pp. 213-215) and published in Maxwell (1871). Cf. Norton (2018) for an analysis of this TE and its recent reformulations.

[67] Cf. Isaac Newton, *Philosoph. Nat. princ. math.* (1687), *Definitiones*, 17, 11-12. Eng. transl. in Cohen (1999, pp. 414-415).

common centre of mass, so each sphere is at rest relative to the other. If they are in motion, Newton argues, then there should be a force acting on the rope. Were the spheres at rest, this force would instead be absent.

Newton suggests that this TE offers an argument against a relationist view of space. The relationist holds that only material objects exist, and space is just the set of spatial relations between them. Now, Newton argues, relationists are unable to offer any explanation for the presence of the force acting on the rope between the two spheres orbiting in the empty universe. This is because, from their point of view, motion is always relative to something else, and here the spheres are at rest with respect to each other. So, they do not move with respect to anything, and hence it can't be motion that explains the force between them. By contrast, the absolutist about space can provide an explanation for the presence of the force, namely that the two spheres are moving relative to absolute space itself; conversely, the force is not present when the two spheres are at rest, because they are not moving with respect to absolute space.

Newton's TE has an abductive nature: absolute space would provide the best explanation for the presence of the force in the experimental scenario. However, this inference only concerns what is true within the TE. It is not evident *per se* that such a result is also true about the nature of space in our actual universe, which is importantly different from the scenario described by Newton – for example, it is not empty.

Einstein's TE involves a scientist closed in a uniformly accelerated lift, which is not subject to any gravitational force.[68] In such a scenario, the scientist will see the objects moving as if they were subject to the effect of gravity. The TE's gist is that motion in accelerated frames of reference is observationally identical to gravitational motion. From this, Einstein derives a structural identity between uniformly accelerated motion and motion in a gravitational field. Then, he is able to posit the *principle of equivalence*, which establishes an identity between inertial and gravitational masses, which classical mechanics deemed as theoretically distinct. This result applies to the imaginary scenario with the scientist trapped in the lift. The point is then to show that what is true in the TE is also a thesis about real-world mass.

Generally, in scientific TEs, the passage from the imagined scenario to external targets is often mediated by theoretical and empirical assumptions. However, the inference from an imagined scenario to the behaviour of a real system is not as straightforward as it may seem at first glance. Nancy Cartwright highlighted this problem regarding what she calls Galilean experiments – MEs or TEs that isolate "a single factor as best as possible to observe its natural effect when it operates 'on its own' with no other causes at work" (2010, p. 23). She focuses specifically on

[68] Cf. Einstein (2002, pp. 68-69) and Einstein and Infeld (1938, pp. 230-235). For an analysis, cf. Norton (1985).

cases where an experiment involves unrealistic assumptions because in such cases it is necessary to "climb up the ladder of abstraction" in order to get "from falsehood to truth" (*ibid.*, p. 20).

My attempt here is to provide a generalisation of Cartwright's point, namely a systematic account of the distinction between internal and external validity in thought experiments. Looking at this issue in relation to the case of MEs, one realises that the issues relative to extrapolation do not depend on unrealistic assumptions only, as Cartwright seems to suggest. Sometimes, the assumptions made in the TE are realistic for some application, and unrealistic in other contexts. So, it is not simply a matter of falsehood and truth of the assumptions made in the experiment. Rather, I suggest it is better to interpret the issue in terms of internal and external validity, and regarding the latter, I will argue in section 2.3 that the passage from the experimental assumptions to the extrapolation to an external target is better understood as a matter of representation, understood in the terms explicated in Chapter 1.

2.2.3 The experimental nature of TEs

Mine is not the first attempt to put the 'experiment' back into the 'thought experiment', in the general sense of relating TEs and MEs in order to reveal features of the former. Indeed, the literature on TEs contains numerous suggestions of this sort. For this reason, I will briefly review the most relevant contributions on the experimental nature of TEs and show how my treatment differs from or goes beyond previous proposals. Readers unconcerned with matters of novelty can fast forward to section 2.3.

Some of the contributions highlighting the relation between TEs and MEs insist on their mutual irreducibility or complementarity. For example, Sorensen (1998) draws important epistemological parallels between TEs and MEs, but his primary aim is to show how they serve different tasks, so concluding that the latter cannot entirely replace the former. In particular, he argues that TEs are examples of ideal experiments that are impossible to perform in the real world. In a similar vein, Buzzoni (2008, 2018) puts forward a transcendental interpretation of TEs, in which they would constitute the condition of possibility of MEs, both by framing the modal space of events and by providing a conceptual background to design and produce actual material experiments. Häggqvist also focuses on the modal features of TEs and claim that TEs are defined as hypothetical tests for theories. So, TEs do not directly provide knowledge about the empirical world (2009, pp. 59-60) – in this sense, he is close to El Skaf's account.

None of these three authors make explicit reference to validity. Furthermore, they all seem to share the core idea that TEs' results should be constrained when we turn our attention to the real world. If TEs do teach us something about the empirical world, it is just in the sense of delineating the possibility space for actual phenomena

to occur.[69] I think that these accounts have problems in defining what "possible" exactly means, and to tailor it so that it can encompass all the rich varieties of "possibilities" drawn by TEs. In fact, many TEs employed in science describe scenarios that are just impossible. As Stuart (2020, pp. 972-976) points out, the imaginative activity in TEs, even scientific ones, is sometimes productive exactly because it is "anarchic", i.e. radically independent of previous theories and assumptions. This freedom can lead to imagine impossible scenarios, where physical laws are explicitly violated – think of Maxwell's demon, or even Einstein's scientist running at the same speed of a beam of light (cf. Norton 2013). Moreover, these authors still express a unidimensional characterisation of TEs, without considering the two-pronged nature of validity emphasised in this Chapter.

Stuart (2016) proposes a "material" account of TEs, where the justification for their results depends not on the formal or logical relations between the propositions expressed by the TE, but on the material ones. He turns to Franklin's (1986) criteria for good material experiments in order to delineate analogous criteria for TEs. He takes into consideration the isolation of the experimental settings, the elimination of experimental bias, the identification of potential sources of error, the calibration of instruments, and the specification of a theory of measurement (Stuart 2016, p. 460). While I agree with Stuart's application of these procedural guidelines to TEs, they seem to have little to do with IEVD. Indeed, some, if not all of these criteria have a different meaning depending on whether one is concerned with internal validity or with external validity. While isolating causal features can be relatively unconstrained in the internal dimension, the process of controlling causal factors will be more difficult when one applies the results to a real-world scenario. Different types of biases can affect the construction of the scenario and the inferences we draw about the external targets of our investigation. The theory of measurement Stuart proposes for TEs is a theory of inference making (*ibid.*, p. 461), but it disregards the fact that the inferences warranted within the TE may be very different from the ones concerning the world of phenomena, which is what we have to be more careful about. Hence, I take it that Stuart's methodological recommendations, while very helpful in general for any sort of experiment, remain orthogonal to my analysis of TEs.

IEVD should not be confused with other dichotomies that have already been drawn in the literature. For example, the one between *interpretation* and *material realisation* suggested by Radder (1996, pp. 12-13) for MEs and applied by De Mey (2003) to TEs. According to Radder, while interpretation concerns the outcome of an experiment *vis à vis* a precise theoretical background, material realisation is the idealised concept of an experiment *qua* a mere set of actions. Even if we grant appeal to a notion of an experimental action deprived of any theoretical interpretation, this distinction has nothing to do with IEVD. This is because both internal and external

[69] See also Häggqvist (2013) for how philosophical TEs do not by themselves provide justification for this sort of modal knowledge.

validity crucially depend on conceptual and theoretical assumptions.

Inspired by Mach's work (1896), Arcangeli (2018) distinguishes a dimension of production from a dimension of presentation in TEs. The dimension of production is the actual mental process of selecting and isolating features of the TE's scenario, manipulating of the imagined systems, and observing results. The dimension of presentation corresponds to the interpretation of the results in the light of a theory (*ibid.* p. 17). Again, this distinction targets something very different than what IEVD does. First, I take both the internal and external validity of a TE to depend on the scientific, theoretical background. Similarly to what I have said about Radder's distinction, it is not theories that get the lion share of the work in separating internal and external validity. Second, Arcangeli focuses on the dimension of production in order to show that the imagination involved in TEs is best characterised by appeal to mental models. This supports her view that TEs are useful because they allow us to "perceive" and "believe" from perspectives that are not directly present to our senses, and this constitutes the "experimental character" of TEs (*ibid.*, p. 15).[70] As I will show in section 2.3.1, I offer a fictional treatment of the internal dimension of TEs, thus explicitly denying that belief is the cognitive attitude that the experimenter either does or indeed should entertain when they perform a TE. In this respect, Arcangeli's view is very different from my own account.

Let us now look at studies that have more or less explicitly appealed to IEVD. Wilson (2016), for example, has proposed such a distinction concerning moral TEs. However, his account does not give a precise account of either type of validity. More importantly, his analysis restricts itself to normative cases, namely to moral TEs, which are relevantly different from factual cases as regards a potential definition, and method of assessment, of external validity. Even if there is not space to delve into this issue, the gist is that it is not clear whether we can think of the external validity of a normative TE in terms of representation at all – and if so, representation of what.[71]

In her doctoral dissertation, Murphy (2020) develops a rich comparison between TEs, MEs, and computer simulations. Although she is clearly aware of the fact that the IEVD can be drawn in all three activities, she does not give a precise account of both types of validity when it concerns TEs. She mentions external validity issues only when she criticises the alleged superiority of MEs over computer simulations and, implicitly, TEs (cf. *ibid.*, pp. 33-37). The development I offer in the following sections, as well as the positive consequences of my treatment of TEs' validity in section 2.4, thus go beyond Murphy's remarks while remaining compatible with them.

Finally, El Skaf and Imbert (2013) have offered a number of theses that are close to

[70] This bodily dimension, in that it allows us to feel, perceive, and thus believe, seems to be the main reason for which Arcangeli relates TEs to MEs also in other works, like her (2010, p. 584).

[71] Interesting thoughts on the same question, though applied to cases of normative models in economics and moral theory, emerge in Beck and Jahn (2021) and Roussos (2022).

my own, though they differ in both their perspective and goal. They argue that TEs, MEs, and computer simulations share a "functional description", which is articulated as follows. They all are (i) question-oriented activities, and they involve (ii) a scenario, (iii) an unfolding of that scenario, (iv) the achievement of some results in the scenario, and finally (v) the obtaining of a scientific conclusion – i.e., an answer to the original question. While they neither talk about validity, nor refer to IEVD, El Skaf and Imbert clearly separate the results of the scenario from the answer to the scientific question. Thus, they have hit the same nerve on which I intend to focus, but without framing it in terms of validity.

The account proposed in this Chapter develops their ideas by being more general in certain respects, and by diverging from theirs in others. First, El Skaf and Imbert explicitly focus on the unfolding of the scenario, which corresponds to my internal dimension. My fictional treatment of internal validity is different and more general from theirs, insofar as I relate the internal dimension to the literature on fictions and the imagination in the context of science. Also, this fictional characterisation of TEs' scenarios distinguishes them in a relevant way from MEs and computer simulations, while it makes them more similar to scientific models. Furthermore, I offer a deep analysis of external validity, which is at best a secondary concern of theirs. I connect external validity with representation, which allows me to develop the view in a new direction. Their focus on the internal dimension leads them to say that the primary task of TEs is "explicatory" (*ibid.*, pp. 3463-3464) with respect to background theories, which means that their primary function is to develop, analyse and assess theoretical assumptions. Instead, I want to insist on representation, and thus on the *external* relation between TEs and the world. Despite these differences, the fact that my account generally converges with El Skaf and Imbert's characterisation of TEs is a sign of how useful the distinction of validity can be at many different levels of the discussion.

Before I show how IEVD positively contributes to solving the controversy presented in section 2.1, I need to spell out a precise account for both the internal and external validity of TEs, which I develop in the next section. This is a required step in order to fully appreciate the potential benefits of IEVD when applied to TEs, as well as to the issues currently troubling the relative philosophical debate.

2.3 Developing the account

2.3.1 Internal validity and games of make-believe

In this section, I propose an account for the internal validity of TEs in terms of Walton's (1990) games of make-believe. The benefits of connecting TEs to Walton's

treatment of artworks have been noted before.[72] In particular, Meynell (2014) was the first to explicitly suggest interpreting TEs in this way. However, as it often happens with fictionalist approaches to scientific contexts, Meynell does not offer a precise account of how exactly TEs provide knowledge about the empirical world (cf. *ibid.*, p. 4165), thereby failing to address the relevant problem of TEs, as identified by Kuhn. In contrast, Meynell seems satisfied with Walton's account as providing a definitive answer to the question regarding the epistemological status of TEs – or, at least, providing the fundamental ground for any philosophical analysis of them.

I intend to argue that Walton's semantics does indeed provide a neat account of the internal validity of a TE. However, in contrast with Meynell, I contend that this approach gives us just half the story: further work is needed to analyse the external validity of TEs and address the related issue of how TEs may offer new knowledge about the empirical world.

Salis and Frigg (2020) also employ Walton's games for an analysis of the scientific imagination involved in TEs and scientific models. The view put forward by these authors is more similar to mine than to Meynell's, insofar as they restrict themselves to the internal dimension of the imagination involved in TEs, without developing their account any further about the possible epistemic import of TEs for the empirical world.

In this section, I first introduce the broad outlines of a Waltonian account of TEs on the basis of the previous works I have just mentioned, and then apply it as a general framework for the internal validity of TEs.

When we look at a piece of art, Walton argues, we are often engaging in a game where the material elements of the artwork have to be interpreted according to specific rules. For example, most bi-dimensional coloured canvases must be interpreted as presenting three-dimensional objects; the statue of a blindfolded woman with a scale in her hand should be understood as an allegory of justice, and so on. This also applies to many works from the non-material arts, like literature or music. Beethoven's *Pastoral* should make us imagine a rural landscape and atmospheric events, and Abbott's novel *Flatland* tells the adventures of a two-dimensional object that encounters objects from a three-dimensional reality. Here too, the sound of a composition, or the written text and the pictures in a book prompt our imagination to create a fictional scenario and develop the narrative.

In Walton's terminology, the material vehicles of the game, like the pictures, the sounds of music, and the written texts, are called *props*, while the rules that guide the construction of the game's scenario and its further unfolding are called *principles*

[72] Godfrey-Smith (2006), Frigg (2010), Toon (2010), Levy (2012), Frigg and Nguyen (2016) also proposed a fictionalist account of models and suggested that we have similar epistemic attitude, namely one of imagination. Besides Godfrey-Smith, all authors mentioned explicitly appeal to Walton's theory of imagination as a game of make-believe. Readers interested in comprehensive surveys on the accounts proposed in the literature about imagination and fiction in science can look at the volume edited by Levy and Godfrey-Smith (2020).

of generation.

In this game activity then one generates a fictional scenario, or fictional world. A fictional world can be best determined as the set of propositions that are true in it (Walton 1990, pp. 35 ff.). Let us call 'w-fictional' a proposition that is true in a game of make-believe w. A proposition then is w-fictional, or simply fictionally true in w, if and only if it is directly expressed by the prop (the writing, or the image) or derived from it through the principles of generation assumed in w. Salis and Frigg (2020, p. 35), who have applied the concept of game of make-believe to scientific cases, call the former set of truths *primary truths*, and the latter *implied* or *derivative truths*. Derivative truths may not be explicitly stated in the description of the fictional scenario. They are inferred by further reasoning, on the basis of prop together with the principles of generation that are in play in that specific game.

It is important to insist on the intrinsically normative nature of Waltonian games (*ibid.*, p. 36). The game prescribes us to imagine certain contents and rules others out. More generally, there are licit and illicit acts of imagination. These are determined more or less explicitly by the principles of the imagination, which can be dependent on the constraints an epistemic community holds regarding the specific context of investigation. Thus, the results of our interpretation of, and reasoning about, the fictional scenario can be evaluated on the basis of the legitimacy of the individual imagining. This process becomes even more rigorous in the scientific imagination, where the principles of generation are inferential schemes, mathematical theorems, evidence-based assumptions, and tenets of our most general theories. In other words, the fact that we are imagining does not mean that anything goes.

It is also crucial to highlight that what is true in a fictional world is independent of the concept of truth *simpliciter*, whatever theory of truth one may want to endorse. What is true or false in the game is solely determined by the prop and the assumed principles of generation. The fact that there has actually never been a fabulous treasure on the little island of Monte Cristo, for example, is just irrelevant for the game we play when we read Dumas' book *The Count of Monte Cristo*.

The same, I want to suggest, holds for cases of the scientific imagination, when for example we are asked to imagine frictionless planes and perfectly elastic bodies, fully rational agents and completely free markets, or very general structures of a neuron or a gene that are not instantiated by any actual organism. In fact, the scientific imagination often presents a mixture of truth and falsehood, with observation-based elements merging with theoretical assumptions, idealisations, and abstractions. Fiction should therefore not be confused neither with truth nor falsity.[73]

This semantic independence of Walton's games with respect to truth has also an important consequence at the epistemic level, namely that one is neither committed to *believe* in any particular content of the game, nor vice versa to believe that such

[73] Cf. also Frigg and Nguyen 2020, Chapter 6.

content is false. So, what is required to be imagined has no bearing on the credential attitude we ought to take towards it, and the prescriptive character of Waltonian imagination does not entail any kind of epistemic commitment.

Walton's original theory of pretence is much more articulated and richer than the brief exposition given here. I want to stress that I am not endorsing every part of Walton's account. The ingredients I need for a fictional theory of models are only the three main concepts that I have introduced so fat: fictional truth, props, and principles of generation. Many other aspects of Walton's theory are simply not required in the present investigation. As a general maxim, then, any other position that Walton has held in his work should not be associated with the account put forward here.

Moreover, Walton's account originally allowed imagination to be objectual (e.g. imagining a tree, like producing a mental image of it) or propositional (imagining *that* there is tree), like forming some form of proposition via a mental language[74]. In this work, I don't take a stance on the debate about whether mental representations associated with scientific models are propositional or not, as it has no direct effect on the nature of epistemic representations like models and their use in science. I then suspend judgement on whether mental representation has to be understood propositionally, non-propositionally, or both depending on context.

This approach to artworks, as Salis and Frigg (2020) have already shown, neatly captures the key features of the scientific imagination, i.e., the kind of imagination involved in scientific models and TEs. In order to show how this works, let us return to Galileo's TE from the previous section. One has a text, written by Galileo himself, that works as a prop for our imaginative activity. Further, this activity is governed by explicit but also implicit assumptions and rules of inference, given by the scientific context in which we are operating. The primary truths provided by the prop concern the existence of one ball on the edge of a V-shaped cavity and the idealised features of these objects, as well as the absence of friction. The explicit principles of generation at play are: the definition of uniformly accelerated motion; the absence of friction; and further idealisations about the motion of the ball when it approaches the vertex. However, other principles are in place, for example classic mathematical derivations and logical inferences. As it is evident at this point, we are prescribed to imagine a scenario where false statements and true ones are irremediably intertwined. From the combination of prop and principles we obtain the derivative truth that, once the ball starts rolling down the slope, it will reach the same height on the other side of the cavity. Furthermore, it is true in the fictional scenario that, were one of the arms of the cavity bent till being horizontal, the ball would proceed to roll forever, thus exemplifying the modern law of inertia.

This view can just as easily be applied to most TEs used in science: Maxwell's

[74] See Fodor (1975).

demon in the box, the imaginary scientist that Einstein conceived trapped in an elevator, Newton's two lonely spheres rotating in an otherwise empty universe, and so on. In all these cases, a prop and a set of principles of generation can be identified. It is possible to further infer the derivative truths to understand the properties exemplified by the TE's scenario. Accordingly, we can say that a result is internally valid in the thought experimental setting w if and only if it is w-fictional, that is, it is part of the explicit description of the fictional world, or it is derived via a principle of generation.

Sometimes, internal validity is not easy to establish. For example, Mach (1919, pp. 228-238) contests that there is no way, in Newton's TE, to ascertain that the rope is actually undergoing that force. For there is no observable difference between the case where the spheres are at rest and the case where the spheres are rotating. In other words, Mach is accusing Newton of begging the question: he is already assuming the negation of the relationist thesis, namely that there is any physical difference between the two cases. In my framework, this concerns the internal validity of the TE, as it concerns what is true in the thought experiment's scenario. Here, I do not wish to take a stance on who is right or wrong in the controversy. What I want to show is only that this should be conceived of as a debate on the internal validity of Newton's TE, and that one can analyse the internal tenability of a TE by investigating what is actually true in the imaginary system.

In a Waltonian framework, there can be countless derivative truths, which all have the potential to be relevant results for subsequent scientific investigations. In this sense, the account is different from the one proposed by El Skaf and Imbert (2013): they insist on a prominent difference between the unfolding of the (TE's) scenario and the internal results. In contrast, my account has no principled way to identify "the" results. All derivative truths are on the same level and can in principle play the role of internal results. What counts as a "result" of the TE will depend on the context of external application, once we try to extrapolate the properties of the surrogative system to real targets.[75]

It is important to also remind about the intrinsically normative nature of Waltonian games (Walton 2020, p. 36). The game prescribes us to imagine certain contents and rules others out. For example, we are not allowed to ignore that there is no friction between objects in Galileo's TE. More generally, there are licit and illicit acts of imagination, particularly in the context of scientific imagination, where the generative rules will be of the sort of mathematical theorems, logical inferences, and constraints dictated by empirical observation.

All things considered, there are two main advantages of treating the internal

[75] I think that El Skaf and Imbert will have to refer to external factors in order to distinguish the results from the general unfolding of the scenario. For example, they may appeal to the scientific question the TE is meant to answer. I want to resist this, because the internal dimension should retain enough independence of the specifics of the external application.

validity of TEs in this way. First, this approach allows enough freedom to employ false assumptions without requiring that our attitude to them is belief. Consequently, the account also keeps explicitly distinct the internal level of analysis from the external one. As we are not concerned with truth *simpliciter* from the start, we refrain ourselves from making any inference about external phenomena. Second, once the game is on, it imposes strict rules, which simultaneously captures the normative dimension of the scientific imagination, its social dimension, and its potential for rigour. Thus, the rules of the game endow it with a prescriptive nature, balanced against the freedom of the imagination.

The make-believe account, when applied to the scientific imagination, has been the target of many criticisms. Thomasson (2020) and Friend (2020) have cast doubts on the ability of fictions to denote and to be elements of comparisons with real world systems. Todd (2020) has also raised some doubts on Salis and Frigg's proposal: if all we achieve from TEs and scientific models already depends on the principles we started with, how can we learn something new? The make-believe account also has to compete against alternative accounts – see e.g., Godfrey-Smith (2020), who treats the scientific imagination in terms of counterfactual conditionals, and French (2020), who characterises TEs credence status in terms of "quasi-truth".

In defence of the make-believe account, it must first be noted that much of the general criticism raised against the fiction view of scientific models is not relevant to the present work, because my aim is to restrict Walton's account to what concerns internal validity only. Therefore, I grant that this view is in itself limited and demands to be integrated with an account of the epistemic relation to the external world. Such an account will be offered in sections 2.3.2 and 2.3.3. Besides, even if the objections raised against the fiction view were relevant to this restricted application, much work has already been done to answer them. For example, Frigg and Nguyen (2021) debunk several commonplaces about the fiction view on models. Salis, Frigg and Nguyen (2020) also offer important reflections on how scientific fictions can denote, without renouncing basic anti-realist tenets of their account of fiction. Moreover, Salis (2016) offers a way to make sense of fiction-world comparisons in a fictionalist framework. Furthermore, in their paper, Salis and Frigg (2020) not only introduce the make-believe account, but also give compelling reasons for why it is preferable to other treatments (like the counterfactual one), and why it is important to clearly distinguish the imaginative attitude from the belief attitude.[76] Concerning Todd's remarks, one can first answer that the internally valid results still have to be applied to real phenomena, and in this step, as I will show presently, we could acquire new knowledge. But, even focusing now simply on what happens within the thought

[76]This last point is in continuity with the remarks offered by Stuart (2020): we must be free to explore different points of views and apparently counterintuitive ideas in order to progress in our scientific understanding, and this makes much more sense if our imaginative activities do not require us to epistemically commit to the fictions we entertain in our minds.

experimental scenario, I think that Todd's worries may be misplaced: scientists are not logically omniscient, so even if they knew all the initial assumptions that govern the imaginary system, this would not rule out the possibility of them being genuinely surprised by the results they achieve by studying the scenario's implications.[77] At the same time, scientists may not be able to explicitly list from the start all the necessary and sufficient assumptions required for our scenario to work. It is also the task of the philosopher of science to analyse scientific fictions and reconstruct the relevant assumptions at play.

Finally, my aim here is not to defend Walton's view *per se*, but to show that its main tenets provide an optimal way to address the problem of TEs' internal validity. Thus, my thesis here should be taken as purely conditional: if one applies Walton's make-believe semantics, then one achieves a working account of internal validity for scientific TEs. There is unfortunately no space here to delve into the debate on fictionalism any further. However, at least abductively speaking, the fact that Walton's account provides a neat answer to the definition of internal validity in scientific thought experiments should already count as a good argument for considering the view seriously.

Because of the intrinsic epistemic restrictions that I have been applying to the fictionalist approach, I have yet to clarify the relation between the scientific imagination and the knowledge about the empirical world. In my framework, this concerns the problem of external validity of TEs and is the topic of the next two sections.

2.3.2 TEs as representations

TEs employed in science are, of course, not only games of make-believe.[78] *Qua* instruments of scientific investigation, they can also be evaluated as tools to investigate the real world. Galileo's TE with the V-shaped plane does not seem to be just a speculative exercise to show that the law of inertia is true in *that* fictional scenario: it is meant to be an argument for the truth of that law in the actual world. Then, to paraphrase Brown (2011), the game of make-believe of section 2.3.1 should become an optimal "laboratory of ideas", where theoretical hypotheses, empirical observation and purely fictional elements interact fruitfully, the ultimate goal being the discovery of interesting features of reality.

The question, then, concerns how to move from the TE to the external target system. In other words, we are now supposed to reflect on the external validity of a TE, namely the validity of its results outside the fictional world and in real scenarios. Ultimately, this is the central question for the epistemology of TEs: whether our

[77] It is not always just a matter of lack of logical omniscience: cf. French and Murphy (2021) and the element of genuine surprise of TEs.

[78] Nor often are artworks: paintings and novels are often telling about the real world as well. For example, *Flatland* is meant to be a mordant parody of the hypocrisy and closed-mindedness of the Victorian society, and Orwell's *Animal Farm* should be read as an allegory of Stalin's regime.

imaginative activity can lead us to knowledge about the world.

I propose to preliminarily distinguish two aspects of the question of external validity, which have often been merged in the debate on the epistemology of TEs. The first question concerns the *definition* of external validity. The second question regards the *criteria* to assess external validity. The second question in fact corresponds to a question about the *justification* of our extrapolations: how we justify our reasoning from the TE to a real target system. The same distinction between definition and justification holds for internal validity: one can define internal validity in terms of fictional truths, and then assess whether a claim is internally valid by checking whether it follows from the prop combined with the principles of generation. Once this distinction is made, I argue below that an answer to the definitional question about external validity does not determinate a universal answer to the methodological question. However, the way in which we define external validity will have important consequences for our way to assess it.

Let us start with the definitional question. Instead of imposing the definition of external validity from the outset, I suggest looking at Galileo's TE once again and see how it inferentially relates to an external target.

First, this TE describes a fictional scenario that, once interpreted as a surrogative system, is meant to highlight some specific properties possessed by real physical systems. Clearly, the TE is *about* motion of objects in space. This aboutness is, I take, a form of reference: the TE refers to other things. Moreover, the fictional scenario is about these other target systems only insofar as it is endowed with a theoretical interpretation that makes it refer to other target systems. Because it is an interpreted object that is also about something else in the world, we can deem the fictional world a symbol. Now, the traditional name we give to the referential relation between a symbol and an object, or a class of objects, is *denotation*. Therefore, the TE first of all is supposed to denote some real target systems in the world.

Second, the TE instantiates many properties, but it definitely seems to be able to emphasise or highlight some of them. For example, the law of equal heights and the principle of inertia are made salient in the fictional scenario. So, the fictional system also refers to some properties that are instantiated by the system itself. This form of reference, that we have encountered in Chapter 1, has been theorised by Goodman and Elgin and is called *exemplification*. It is useful here to repeat the theoretical definition of exemplification: an object X exemplifies the property A *iff* (i) X possesses A and (ii) X refers to A. In this case, we can call X an *exemplar* of A. The notion of exemplification, in turn, is the reason for which some properties of a system are made salient and epistemically accessible (Elgin 1996, Chapter 6). Consequently, some properties will be put in greater evidence than others, which could be eventually distorted or neglected. While Galileo's TE exemplifies the law of equal heights and the law of inertia, it distorts the vertex of the cavity and ignores friction.

Third, these properties are intended to be *imputed* to systems in the world. For example, one may want to say that the motion of the ball in the V-shaped cavity is meant to exemplify the same properties of the oscillatory motion of a pendulum, or the motion of rolling objects on a curved surface – like a skateboarder on a two-sided ramp. The law of inertia is meant to be a general feature of motion in our world, expressed in the form of a counterfactual statement.

The problem is that the properties exemplified by the TE's scenario are not really the ones we end up imputing to real target systems. Skateboarders on ramps and real pendulums do not exemplify neither the law of equal heights nor the principle of inertia. Therefore, we need a way to translate the ideal properties exemplified by the TE into the ones we actually want to impute to a specific target system. As we have seen in Chapter 1, this is what Frigg and Nguyen (2020, pp. 174-176) call *keying-up*. Again, the key is thought of as a function, mapping the properties exemplified by the representation onto the properties that we actually want to impute to the target. You can think of keys as the ones we find with maps, specifying how to read them correctly in order to orient ourselves in real space.

In this sense, Galileo's TE exemplifies the law of equal height, which is then translated via the key in the form of an approximation of the original law. The approximation will mathematically depend on the amount of friction present in the real target system.

A further, interesting case of key is provided by the application of the principle of inertia. In this respect, the TE establishes the following property: if an object were not subject to any force, it would persist in its state of motion (i.e., it would either be at rest or move in a straight line with constant speed). This counterfactual[79] statement is true about all moving systems in our world. However, this counterfactual key is required only if we take the extrapolation class to be such a wide array of scenarios. If the extrapolation class becomes more specific, the key required could be different. For example, objects in interstellar space come pretty close to being described by this law too. In this sense, the key is re-adapted so that the law can be applied to real physical systems via approximation, and not counterfactually. This is particularly insightful as it shows that the properties exemplified by a TE map onto different targets in different ways.

Clearly, the four terms italicised correspond to the four crucial elements defining the relation between the TE and its target: denotation, exemplification, keying-up, and imputation. As we have seen in Chapter 1, these are also the fundamental ingredients of the so-called DEKI account of scientific representation, developed by Frigg and Nguyen (2020, pp. 159-215). Given the analysis of Galileo's TE just offered, it is then natural to interpret this TE as an epistemic *representation* of mechanic

[79] A counterfactual interpretation seems the most natural way to go. At the same time, this seems close to Nguyen's (2020) appeal to "susceptibility" when he describes the application of results that we achieve from extremely idealised models.

motion in the real world. This is exactly the suggestion I want to put forward, namely to interpret TEs as representations in the sense expressed by DEKI, and to define external validity as accurate representation of a target system.

DEKI's concept of representation is useful because, as we have seen, it allows misrepresentation. This means that we can consider TEs as representations even when they misrepresent their designated target system, which provide results we know are false when applied to real scenarios.[80] An account of representation based on a relation of similarity – or its formalised version, isomorphism – would instead not permit this solution. This is because, whatever similarity is, either two things are similar, or they are not, so as soon as the TE and the target are recognised as dissimilar, the former does not count anymore as a representation of the latter.

Furthermore, DEKI allows a thought experiment's scenario, taken *in their entirety* or as a whole, to be very different from any real target. Of course, the target scenario can in principle be identical to the imagined one. Then, the TE's results are trivially true about the target as well. However, these cases of identity are rare, because the very point of a TE is to investigate aspects of reality that we do not know yet.[81] In fact, we can seldom be aware, in advance and with certainty, of all the relevant features of the target, particularly the ones we want to reveal via a TE.

Accordingly, DEKI underlines the importance of both exemplification and the key. Exemplification is intrinsically selective, because TEs can highlight some properties at the expense of others that are overshadowed or distorted. Furthermore, the key is also crucial because the properties exemplified by a surrogative system usually need a translating function that maps them onto the intended target.

We have already seen this with the different ways the law of inertia is mapped onto different classes of target systems. So, there is an evident need for different keys, depending on the specifics of the designated target. This does not only mean that the imputed properties may vary depending on the target: they are also patently different from the ones literally exemplified by the TE.[82]

Besides what we do not know yet, DEKI also sheds light on the fact that TEs can involve elements that *we know* are just false when applied to their target. If what is needed in the first place is a true or realistic description of the target in order to assess its external validity, then Galileo failed from the outset, regardless of the kind of extrapolation employed. For example, objects are not perfectly smooth in the real world. This is again accommodated by the functioning of exemplification: in order

[80] The history of science is full of examples of this sort. Lucretius' (*De rerum natura* I, 968-983) argument against the idea of a finite universe is one of them.

[81] As we have seen in section 2.2.3, Sorensen (1998) indeed describes TEs as experiments performed in the mind of scientists because impossible to be carried out as material experiments.

[82] If one was troubled by the idea that fictional scenarios instantiate properties, there are good news: the model system instantiates properties only as an object interpreted by some function I. So, the properties are I-instantiated and therefore I-exemplified. See Frigg and Nguyen (2020, pp. 172-173).

to emphasise some properties, others will inevitably be omitted from consideration. This aspect is also in accordance with the freedom implied by a fictional treatment of TEs internal validity.

My analysis seems to naturally account for the functioning of Galileo's TE. It also fits the mould of all the other TEs that I mentioned so far. For example, Newton's TE aims at establishing a property of real space, namely its independence of the objects inhabiting it, by exemplifying that property. Here, the key keeps the property of being absolute unchanged. In the case of Maxwell's TE, the actual contradiction of the second law of thermodynamics is best be converted into a statistical interpretation of the same law. Finally, concerning Einstein's lift, nobody has in fact ever observed a uniformly accelerated box away from any gravitational field. The observational identity between uniformly accelerated motion and the motion in a gravitational field has to be translated into a true identity between the two types of motions.[83]

This gap between what is true in the TE and what is true in the world follows from the very characteristics of TEs as fictional scenarios employed to represent external targets. While there are prescriptions about the content to be imagined, nothing yet forces us to believe the contents of our imaginings as true. At the same time, TEs can then still be used as tools to understand something about the empirical world. I contend that this is possible because TEs exemplify certain properties, thus selecting them and making them salient, and these exemplified properties can then be applied to a target system in the real world via a proper key.

Note that all the aspects that I have illustrated so far regarding TEs also apply to scientific models. Models are often employed to discover features of targets about which we are still unaware or uncertain; they usually include plainly false assumptions or highly idealised controls; and these assumptions are not just a necessary evil, they also positively contribute to the success of the model's external validity. Moreover, despite not commonly being framed in terms of internal and external validity, the literature on scientific models offers a great deal of analysis on the distinction between truth within the model and truth about the target.[84] Given these relevant similarities, I suggest addressing the issue of TEs' external validity in close analogy with how we address related questions about representation in the context of scientific modelling.[85]

At this point, the reader may suspect that I am just identifying TEs with models.

[83]This key will require some sort of justification, along these lines: states of affairs which are not observationally distinct should not be distinguished by scientific theories. Interestingly, as Norton (1991, p. 136) highlights, this meta-theoretical principle plays some role also in at least another TE by Einstein, namely the one of a magnet and a conductor, illustrated at the very beginning of his seminal paper "Zur Elektrodynamik bewegter Körper" (1905), where Einstein proposed for the first time his revolutionary new Special Theory of Relativity.

[84]This distinction is a basic tenet of the so-called representation-as accounts (Goodman 1976, Elgin 1983, 1996) and of the DEKI account (Frigg and Nguyen 2020, §8). Similar intuitions are expressed also in Hughes (1997, p. S332) and Tan (2021, pp. 16-18).

[85]Salis and Frigg (2020) have shown that the internal activity of the imagination conducted in TEs and in scientific models is fundamentally the same. Here, I am completing the picture by showing the similarity between TEs and models when it comes to external validity.

However, I contend that such a total equation is not warranted on the basis of what I have said so far. I take TEs and models to share a structural feature in their methodology, namely a conceptual distinction between internal and external validity, but this holds for MEs as well, when employed for extrapolation. However, this does not imply that MEs, TEs and models identical.

The identification of TEs and models does not follow even if we add, as I did, that the imagination involved in both models and TEs can be treated as games of make-believe. For Walton's account functions well with both scientific surrogative systems and with works of art, but this does not entail that art fictions function exactly like scientific ones, even within the internal dimension. Indeed, we reasonably expect that the principles of generation involved in the two contexts will diverge to a considerable extent. Finally, the DEKI account of representation applies to many different types of representations: maps, diagrams, scans, simulations, and material and theoretical models. Nevertheless, the generality of DEKI should not be understood as implying an indiscriminate identity between all these different types of surrogative systems.

Thus, besides acknowledging the striking similarities between TEs and models in terms of representation, once representation is understood in DEKI's terms, I put the question about the relation between models and TEs aside. I simply intended to look at how one deals with external validity in models in terms of representation and take inspiration from that as concerns TEs.

Now, I want to focus on the idea that models facilitate successful surrogative reasoning about their targets by means of *accurately* representing them.[86] I suggest that the same holds for TEs. This is the topic of the next section, where I propose to define external validity of TEs in terms of accurate representation.

2.3.3 External validity as accurate representation

I have already introduced the concept of representational accuracy in section 1.6. For the sake of clarity, let me re-propose the definition here:

Accurate representation$_{\textbf{(def)}}$: a model system M is an accurate representation of a designated (i.e., denoted) target T regarding a set of properties Q *iff* (1) M exemplifies a set of properties P; (2) P is converted via a proper key into Q;[87] (3) Q is imputed to T; and (4) T actually possesses Q.

On the basis of this, we can give a clear definition of external validity for TEs:

[86] As we have seen in Chapter 1, some authors have undermined the role of representation to explain the epistemic success of models. The readers can find my replies to their concerns in Chapter 5.

[87] I remind the reader here that the key may also be a relation of identity, mapping P onto itself.

TE external validity$_{(def)}$: A TE is externally valid with respect to a designated target T relative to a set of properties Q *iff* the TE is an accurate representation of T relative to Q.

Thus, I suggest that the extent to which our TE-based extrapolations are valid depends on whether the TE exemplifies properties that, once translated via the appropriate key, are correctly ascribed to the designated target. We need a key in order to address the fact that, depending on the target, the same property can map in numerous ways from the same TE to distinct target systems. As I highlighted above, the law of inertia applies in different ways depending on the type of target – sometimes counterfactually, sometimes as an approximation. Of course, the key should not be completely *ad hoc*, but rather it should associate the properties of the TE and the target in a systematic way. This is the case in Galileo's TE and the other examples mentioned so far.

It is important to stress here that my definition of accurate representation, which offers the ground for my definition of external validity of TEs, does not involve any assumption about the *similarity* between the representation system and its target. Indeed, accuracy here only concerns the results of a TE once they have already been properly translated by a key. The key included in my definition, inherited from the DEKI account, allows for very different, and sometimes purely conventional ways to connect the properties of the representation with the properties of the target.

The fact that accuracy is independent of similarity makes my account compatible with, if not in fact a theoretical ground for, what Stuart (2020) calls the "productive anarchy" of TEs.[88] Stuart argues that TEs are sometimes useful exactly because they challenge our theories and intuitions in a revolutionary, radical way. Consequently, they often involve scenarios that are extremely different from the usual ones, if not even impossible according to our best scientific theories. Despite their anarchic nature, these TEs are still important, Stuart argues, because they make us critically reflect on our theories and unchallenged intuitions. My definition of accurate representation gives us another good reason for not being too troubled about TEs' "anarchy": a representation can be accurate, despite its lack of realism or even its conflict with our current theories and intuitions. Therefore, once we introduce a key that translates the more recalcitrant, "anarchic" properties of TEs, we do not need to renounce accuracy in order to account for their potentially revolutionary nature.

Let us now discuss possible types of targets of TEs. The target of a TE can of course be one single object. For example, I take Newton's TE with the two rotating spheres to target one single object, namely physical space. Yet, this is rarely the case: TEs, like models, normally tend to represent classes of systems or types of mechanisms – what Weisberg (2013, §7) calls "non-specific targets". As

[88] Similar ideas are expressed in Murphy (2022, §4).

stated before, Galileo's TE represents pendulum-like motions, and the law of inertia is counterfactually targeting any motion in the world. Similarly, Einstein's lift is a representation of a type of motion, specifically the one instantiated by objects subject to gravitational fields and by objects moving with uniform acceleration. Maxwell's TE targets closed thermodynamic systems and exemplifies the statistical nature of phenomena occurring at the level of the fundamental components of matter. This should not worry us, as TEs are no exception in this respect: scientific models usually target broad, and sometimes even vaguely defined, classes of phenomena. The exact limits of an extrapolation class for both TEs and models is something that we clarify only a posteriori and can change through time.

It is important to notice that nothing in my analysis rules out the possibility of a targetless TE, whose goal is a purely theoretical investigation. For instance, one may wonder if the famous TE that Galileo offers in his *Dialogues* (cf. *infra*, fn. 4) against Aristotle's theory of falling bodies is actually a representation of anything. The TE is meant to be a *reductio ad absurdum* of the Aristotelian theory of falling bodies, which states that heavier bodies fall faster than lighter ones. But what if, Galileo reasons, we connect one heavy object L and a lighter object l, made of the same material, with a rope, and we let them fall from Pisa's tower? If the fictional system satisfies the laws of the Aristotelian doctrine of motion, then we obtain two mutually contradictory answers. On the one hand, the lighter object l should act as a brake, so that the resulting velocity of the compound is somewhere in between l's and L's velocity. On the other hand, l+L is itself one single compound, which as a whole is strictly heavier than L alone. So, the compound's velocity must be strictly greater than L's one. Therefore, either Aristotle's theory is contradictory, or it is vague enough to produce contradictions.[89]

This TE may be said to be targetless: it is only a fictional scenario employed to reflect on the implications of Aristotle's theory.[90] Alternatively, one may argue that Galileo's falling bodies has a target after all, however general and vague it may be. In fact, we would be imputing to physical systems involving falling bodies the property that the speed of falling objects does not depend on their weight. Brown (2004, pp. 30-31) goes even further and contends that there is a general claim we can make about reality on the basis of this TE. Namely, that it is impossible that the velocity of a falling body depends on an extensive property of the object – that is, properties that can be added and subtracted as if they were real numbers.

I am not taking position on the specific question about whether this TE has a target in the real world or not, as it concerns the independent problem of tackling representations with vague targets.[91] What matters is that, even in cases of targetless

[89] For a thorough analysis of this TE, see Gendler (1998) and El Skaf (2018).

[90] Thus, recalling the distinction I made in section 1.1, this TE would work more like a model of a theory than a model of phenomena, for it provides a scenario that simply tries to make Aristotle's tenets true.

[91] As Frigg and Nguyen (2020, pp. 13-14) notice, the very question whether a model is targetless

TEs, the framework that I have sketched still applies. In fact, one can still talk about representation here: as we have seen in Chapter 1, we can adopt Goodman's (1976) terminology and talk about targetless TEs as Z-representations. In this sense, Galileo's TE on falling bodies is an Aristotelian-falling-bodies representation, with possibly no targets in the world – just like pictures of unicorns.

2.3.4 The justification for external validity

Once external validity of a TE is defined, we need to discuss the issue of how a scientists can assess whether a TE is externally valid. This is not a question about what external validity is, but rather *how* we find out whether a TE is accurately representing something or not. In other words: what is the epistemology of external validity and how do we justify our inferences from the surrogative system to the target one?

Besides testing our target system directly, there is no ready-made, universal recipe to determine whether a TE is externally valid. In this, TEs are like other surrogative systems such as models or MEs aiming at extrapolation. Think of an experiment performed, say, on mice in order to study the neurological mechanisms of memory. Of course, in the case we have direct epistemic access to the target system we can also assess the external validity of the experiment. The experimental results on the mice are externally valid with respect to, e.g., humans *iff* what we find out in mice turns out to be true in humans as well. But how do we know that the inference is also justified? That is, besides being delivering a correct result, it is also supported by sufficient reason, or it is just the result of mere epistemic luck? I contend that the answer to this question largely lies outside of the experiment itself, and the methods to provide it can be numerous: performing further experiments on different organisms; establishing parallelisms between mice brains and human memories; evaluation, via archaeological and genetic data, of the tenability of phylogenetic assumptions about rodents and primates. Furthermore, all these forms of investigation may find foundation in overarching theories – in this case, a valid candidate would be the contemporary theory of evolution – the justification for which again relies upon observations, experiments, and other mathematical, physical, and statistical theories, without even starting with our theories on DNA, ontogenesis and so on.

Of course, even if we performed many experiments, severe tests of our hypotheses, and robust analyses of our measurements, this will not necessarily entail that we will have achieved an irrevocable justification for the extrapolation. In principle, the process of justification can go on indefinitely. There will simply be a moment when the scientific community will reach a provisional consensus on the tenability of the required assumptions.

The same holds for models: there is not a one-size-fits-all method to decide

or not is tricky, as it strongly depends on how the target is defined.

whether a model's results are externally valid about a target system, just by looking at the model itself. Fine-grained analysis must be carried out on the theoretical assumptions involved in the model, together with empirical investigation. This is the reason why some models' external validity is so difficult to assess. In some cases, it is because past empirical data are not entirely sufficient for justification – e.g., consider the case of the case climate models, where the effects that humanity started to have on the climate around 200 years ago make the data about farther past less relevant. In other cases, it is the overarching theories that are not a matter of consensus yet – e.g., in economics and psychology – so they are not sufficient to justify the model-based inferences. Generally, the grounds to justify both experimental extrapolations and model-based inferences partially lie outside them. The surrogative system of course participates in the justification for the extrapolation, but it is in itself insufficient to ground it completely.

The partially extrinsic nature of justification holds for scientific TEs as well. Take once again Galileo's cavity TE. We are warranted in applying the result of the TE to real motion because some assumptions are empirically grounded, the idealisations are known to function well in that context of application, and other empirical evidence (e.g., on pendulums) indicate the same conclusions. All this is just extrinsic to the TE itself. Again, this does not undermine our definition of external validity in terms of accurate representation. It just shows that sometimes it is difficult to assess whether a representation is accurate or not. In this respect, the structure of justification is the same for TEs, MEs, and models, when they are used to make surrogative reasoning about a target system.

This extrinsic character of justification that I am emphasising here is just a special instance of the holistic nature of knowledge as described by Elgin (1996) that I anticipated in section 1.4: each of our beliefs is part of a net of other beliefs and endorsements that tend to support each other. To use Dennett's (1995, p. 74) beautiful metaphor in a new context, there are no *epistemic* "skyhooks":[92] all beliefs we have hang on and are supported by other beliefs, in a holistic network that changes and updates itself continuously in interaction with new concepts and experiences. Therefore, our representations are no skyhooks either: they are not self-supporting symbols, providing their self-justification. They are part of symbolic and conceptual systems that support them, in turn supported by other systems of beliefs, assumptions and hypotheses.

At this point, one may complain that TEs seem much more "distant" from empirical reality than scientific models or MEs. Thus, we may not be allowed to answer the problem of justification simply by associating TEs with the other two types of surrogative reasoning. Irrespective of how distance is interpreted here, this will be

[92] In his book, Dennett is talking about Darwin's theory of evolution and its revolutionary result of explaining the design of life without any appeal to an intelligent mind, and instead reducing it to the result of an extremely simple, algorithmic process.

a matter of degree. If we understand distance in terms of similarity, we can have both really "realistic" TEs and very idealised ones, as well as more or less artificial experiments; analogously, we can have both very unrealistic models, and models that describe accurately many aspects of the target. So, there does not seem to be an essential difference between TEs, models and experiments in terms of their distance from reality. Whatever surrogative system one employs, an extrapolation to a target will always require justification. In addition, as I have already emphasised, the definition of accuracy that I am suggesting is independent of the similarity with the target. Therefore, we should not worry about the lack of realism of most TEs, because they can be accurate representations of their targets even when they consist of very unrealistic fictional systems.

One final caveat concerns the relation between the internal and the external dimension. It is important to clarify that IEVD does not imply that the imaginary scenario and its representational function have nothing to do with each other. When it comes to actually constructing TEs, scientists will obviously make considerations of empirical and theoretical nature. Therefore, they will aim from the start at accurately representing something in the world. This does not undermine IEVD, because the two types of validity, although being intertwined in practice, remain conceptually distinct.

In conclusion, I do not think that a definitive set of universal criteria to assess the accuracy of a representation can be established. Like in the case of MEs and scientific models, the success of exporting information from the surrogative system to the target one can be assessed only a posteriori. As in all cases of reasoning based on surrogative systems, results are always tentative and conjectural, just as any other scientific result. In this sense again, we can appreciate the true experimental nature of TEs.

2.4 Stabilising the debate

2.4.1 Amending the yes-no debate

We can now answer Kuhn's questions: TEs can produce knowledge about the empirical world by providing true propositions about real target systems. These true propositions are justified by both the TE itself and theoretical assumptions and their empirical support. This latter ground of justification remains importantly extrinsic to the TE itself. However, I have argued that this is not problematic in principle, as the same occurs in the case of model-based inferences and experimental extrapolation. From this perspective, the problem of justifying the external validity of TEs is just a special case of the broad problem of justifying scientific inferences, and this can be solved only by (i) looking at background scientific knowledge more holistically, and (ii) taking into consideration the specific context in which the inference is performed.

In addition to giving an answer to Kuhn's questions, IEVD and the consequent diarchic account of TEs' validity also allow us to re-frame the debate on the epistemological status of TEs. Let us start with the yes-camp, and specifically with Brown. He has tried to answer Kuhn's questions by focusing on the internal dimension of TEs. His account tends to lump together all the aspects that I have distinguished in this article: internal validity, external validity, and their justification collapse and remain within the TE's scenario itself. Through imaginative activity, Brown argues, we can infer not only what is true in the TE's scenario, but also about the real world, namely true laws of nature. The warrant of our inference is given in both cases by the strength and immediacy of our intuition. If one accepts my distinction between internal validity and external validity, this is too quick. For we cannot automatically infer external validity from the internal one. One has to recognise that there is a fundamental difference between what occurs in the scenario, and how the target system behaves in the world.

One may want to redefine Brown's account in the light of my distinction. We first find out what is true in the TE, under the assumption that some laws govern the scenario; and *if* the same laws also govern our world, then the TE is externally valid. Although Brown does not seem to have this in mind, as he argues that the laws themselves are discovered via the TE (2004, p. 34), I find this reinterpretation of Brown's account more appealing. At the same time, I find any reference to laws of nature, a priori intuition, and Platonic perception unnecessary. My account thus seems preferable, insofar as it does not require laws to be the same in the surrogative system and in the target system, even though this is certainly the case in many scientific TEs. At the same time, Brown's main intuition is retained: we can learn something about the real world via thought experimentation, and this occurs by interpreting TEs as representations in DEKI's sense.

With respect to the yes-camp in general, my account provides a more qualified answer to how we obtain knowledge about the world. The flexibility of the DEKI account is useful to capture the different ways in which a TE can relate to the external target. Compare, for example, the representational account of external validity with the objectualist accounts. As I noticed above, it is not entirely clear why the picture-like nature of the TEs' imagination should add something crucial for our TEs to succeed. Instead, my representational view accounts for it and allows for both propositional and non-propositional treatments of imagination. This is because the account holds that the TE, *qua* representation, exemplifies certain properties, and it includes the use of keys to properly translate TEs' properties into features of real targets.

The way in which the selection of exemplified features is achieved will of course depend on the type of representation: pictures and images have syntactic and semantic features that may be important to understand how some properties are exemplified in certain TEs. Here, the work of Goodman (1976) and Perini (2010) helps highlight the

peculiarity of visual and in general non-propositional representations. For example, picture-like images tend to be syntactically and semantically dense, thus making them rich in detail while remaining concise. Also, spatial relations play an important role as representing other forms of relations. All these features may be relevant to understand how some TEs exemplify properties that are then imputed to a target.

My point is that, first, when we are concerned with validity, a distinction between visual and propositional representation does not help us much at the general level, but rather only at the local one of exemplification. Second, at least when one takes into consideration examples of TEs in physics, they do not seem to require an appeal to specifically non-propositional features to be valid, either internally or externally.

The advantages of the account are evident also with respect to Nersessian's model-based account. In a sense, Nersessian's idea that a TE is a structural analog of a real system is similar to the idea of external validity that I put forward here. However, her concept of structural analog does not provide a satisfactory account of scientific TEs' validity. For the isomorphism (even if spatio-temporal or causal) is not sufficient for a TE to be externally valid. In fact, many things we imagine are often isomorphic in some way to external systems, but this does not make them externally valid. This is because isomorphism is a very abstract notion that can be easily instantiated. We need a richer conceptual background to tell us which isomorphism is the relevant one. In this sense, Nersessian's structural analogy is not a sufficient condition for external validity.

Furthermore, being a structural analogy does not seem to be particularly useful to understand the internal validity of TEs, and thus investigate their imaginary dimension. For one may want to perform TEs that are not externally valid but only internally so, like Galileo's TE on falling bodies. Here, it is at least unclear to what the fictional system is isomorphic, given that nothing in nature seems to instantiate the contradiction exemplified by Galileo's fictional system.

Finally, Nersessian seems to insist that the fictional scenario itself must be relevantly analogous to real world settings, and this is problematic for many TEs that involve events and facts that are implausible or even impossible in the real world (from elevators lost in the middle of nowhere, to scientists who run at the speed of light, to demons or mechanisms that act against the second law of thermodynamics). My account solves this issue nicely, focusing not much on the structural analogies between the TE and the target, but simply on a selected number of exemplified features, which are also de-idealised via a key before being imputed to the target system.

More generally, it seems that a structural analogy, in Nersessian's sense, pertains to a possible method of justifying the external validity of TEs with respect to a designated target system. In fact, it can be *part* of the explanation of why a TE is an accurate representation of a target. This method of justification for external validity, though, for the reasons just given, should neither be confused with external validity

itself, nor with the TE *tout court*. My account can incorporate Nersessian's view as a part of the strategy to justify TE-based inferences. Then, isomorphisms will usually have to be further qualified by an interpretation of the fictional system that will explain *in what sense* it is relevantly isomorphic to real ones. In other words, we will need a key, and thus some further work is required to motivate the choice of the designated isomorphism over other ones.

As regards the no-camp, my account firstly explains why TEs can be intuition pumps in the first place. This is because they offer a free space for our imagination to combine true elements with fictional ones, opening up in front of us possibilities and hidden aspects of the investigation that may have remained implicit or unknown. It also adds crucial information about why this is relevant for science, by providing an analysis of the relation between TEs and external world in terms of representation.

About this last point, El Skaf thinks that we usually do not get "outside" TEs' internal domain. Thus, the only epistemic function TEs can serve is a theoretical one. More specifically, they reveal inconsistencies in our theories (and potential solutions to them). My reply to this view is twofold. First, even remaining within the internal domain, we do not need to restrict ourselves to inconsistencies. Galileo's cavity does not bring up any contradiction, just the implications of some general assumptions combined with empirical data and some relevant idealisations. Similarly, Newton's spheres, Einstein's lift, and Maxwell's demon are not meant to identify contradictions in our scientific theories, but rather to make their consequences manifest. The imagination allows us to stretch the limits of our theoretical knowledge, not only to challenge it. Furthermore, even though TEs instantiate many theoretical assumptions, they are not entirely reducible to them, given the presence of fictional particulars and idealisations. Second, while the results of TEs may sometimes be only internal (e.g., Galileo's falling bodies), this is not necessarily the case for all scientific TEs. Actually, most of them are meant to give us information about the real world. If we think TEs in terms of DEKI, we can account for this in terms of surrogative reasoning. Therefore, El Skaf's view can also be incorporated in my general framework, as it describes the special case of TEs that reveal and solve contradictions in our theories.

Finally, let us turn to Norton. His method consists in getting rid of the imaginary elements and reconstructing the TE as a logical argument, where the premises express empirical knowledge or well-grounded theoretical claims. Therefore, trying to explain TEs' external validity, he imposes constraints on internal validity. Furthermore, by identifying TEs with underlying logical arguments based on empirical premises, he also seems to conflate external validity with the method employed to justify it. As a consequence, he has to reject the view that there is something new that we discover via TEs because all the empirical content is already present in the premises already.

I contend that this strategy is problematic. To begin with, his account is unable to explain the importance of the imagination and fictions, which play a crucial role in TEs. In fact, fictional elements are vital to achieve the internal results, and also

to make salient those properties that we want to impute to the target. Without assuming imaginary objects and their behaviour, the TE normally does not work. In Galileo's cavity, as almost always in physics, we deal with idealised objects, the behaviour of which is rarely if ever approximated in the empirical world. Salis and Frigg (2020, p. 37) have shown that the same holds in Galileo's TE on falling bodies: without assuming events that actually never obtain in the world, Galileo would be unable to prove the inconsistency of Aristotle's theory. The same holds for Einstein's scientist running parallel to a beam of light (Stuart 2020).

One may wonder whether this is actually in contrast with Norton, who sometimes allows some role for the imagination and the particular elements described in the narrative. Fiction would then be useful because it facilitates reasoning. I have two main reasons for doubting that Norton's view in its current formulation is able to take this route. First, while Norton understands fictions and imaginary particulars as merely allowed, I intend to stress that TEs are usually epistemically valid, in the internal sense, exactly *because* they involve idealisations, abstractions, approximations, false assumptions, and so on. Moreover, even when it comes to external application, idealisations are not just a necessary evil: like models, TEs use these distortions fruitfully in order to make some properties salient at the expense of others. Without idealised assumptions on the absence of friction, Galileo would have been incapable of formulating the law of inertia. Therefore, we actually want fictional particulars and idealisations: they are not synonyms of inaccuracy,[93] they are an essential element of scientific enquiry.

Second, the main argument that Norton gives for the superiority of his account over the others is that it gives a clear method to choose between TEs with mutually contradictory results. He calls these "thought experiment – anti thought experiment pairs" (2004b, p. 45). If two TEs have mutually contradictory results, we need a systematic way to understand which one is the correct one. Now, Norton's account envisages two possibilities: either (at least) one of the two underlying arguments is logically invalid, or (at least) one has false premises. However, the first option is quite rare. So, his solution to the problem, which is put forward as a crucial reason to prefer his account, basically relies on the assumption that a good TE has true premises. However, if Norton allows for fictions and imaginary elements to play a role in the scenario, then he loses this option, and his account is not better off than those that he criticises.

My view offers a neat solution to this problem: Norton just needs to split his account in two and consider each TE as (usually) composed by *two* arguments: one is internal to the scenario and ruled by the chosen principles of generations, the other concerns the extrapolation of results from the scenario to an external target. In this way, we can allow scientists to use fictions in the internal dimension, while

[93] This very point is made by Nguyen (2020) concerning toy models.

our empiricist tenets can be retained in the argument we give for external validity. The challenge then becomes to seeing whether there is a good key to translate the exemplified properties into ones to be imputed successfully.

Finally, it is worth noticing that my proposal is in line with other authors' concerns about Norton's account. For example, Stuart (2016), who argues that Norton seems forced to renounce at least one of following: (i) TEs provide knowledge, (ii) an empiricist theory of TEs justification, or (iii) Norton's (2021) material theory of induction:

> There is therefore a serious internal tension between Norton's account of thought experiments according to which thought experiments are filled with irrelevant but picturesque details on the one hand, and his account of induction according to which the particular details are crucially important for justification, on the other [...] it would be instructive to consider whether another empiricist or naturalist account of thought experiments can be created that explains their ultimate source of justification (Stuart 2016, pp. 458-59).

Stuart then goes on in specifying the criteria for a good TE that I summarised in 2.2.3. I contend that my proposal provides the empiricist account of TEs that Stuart hopes for. Following my distinction of internal and external validity, I acknowledge the importance of imaginary particulars as concerns the internal results, while at the same time I place the justification for the external applications on theoretical assumptions and empirical knowledge, which remain largely extrinsic to the TE itself *qua* representation. In this sense, also the material theory of induction seems to be retained, given the justificatory role of background, material knowledge, both empirical and theoretical, that has already been taken for granted by scientists. Finally, my proposal permits us to see how TEs provide new knowledge: they can produce new justified true beliefs about their targets by accurately representing them.

What I have said does not undermine in any way the importance of Norton's work. By offering a rational reconstruction of TEs as an argument, Norton is normally able to assess their external (and sometimes internal) validity. What I have argued is simply that this method of assessing TEs' external validity should neither be confused with TEs themselves, nor we should collapse internal and external validity. In the end, Norton's argument view is perfectly adaptable to my IEVD: it is sufficient to acknowledge that there are usually two arguments involved in TEs. Moreover, I think that my account adds something, as it gives a more qualified characterisation of what *type* of "arguments" are involved in the two cases, respectively: one is a game of pretence, the other a representation-wise inference. In this sense, I am providing the material aspects to fill Norton's formalist-empiricist account.

2.4.2 Remedying the sub-debates

Now, I can also show how IEVD and my diarchic account help mitigating the contrast developed by the protagonists of the debate. First of all, with my distinction in place,

we can both retain Norton's empiricism and accommodate Brown's concerns. Given the freedom, autonomy, and even the anarchy allowed in a game of make-believe, a TE can include elements that go well beyond background theories, play with possible and impossible situations, and mix them creatively with both empirical observations and intuitions. The point is that the internally valid results are not unconditionally valid for any external target: whether they are valid or not depends on their accuracy, which in turn depends on the specification of the exemplified properties, the designated target, and the key involved. Moreover, the justification for the key is importantly extrinsic to the single TE and rely on theoretical and empirical knowledge, and this is in perfect harmony with Norton's empiricist requirements.

Furthermore, the discussion of propositional vs. objectual nature of the imagination finds a more precise place into the debate. Once we recognise that TEs work as representations in DEKI's sense, both the propositional and objectual imagination can be accounted. The difference between picture-like and linguistic representations, in terms of their syntactic and semantic peculiarities (cf. aforementioned works by Goodman and Perini), will become important to understand a specific "step" of the representation process, namely exemplification. However, despite the importance of this aspect, it is clear now that a focus on the non-propositional nature of some representations does not solve the questions of whether and how TEs generally produce knowledge about the empirical world. My account succeeds in that by highlighting more general features of TE-based reasoning. In addition, at least in the case of the TEs in physics that I take as case studies, there seems to be no reason to appeal to non-linguistic representational features. A reference to these peculiarly non-propositional properties could nevertheless be useful for other examples of TEs, or different kinds of reasoning.

Finally, in line with what I said about TEs and empiricism, we have a valid alternative to El Skaf's strategy to restrict TEs' results to the internal, theoretical domain. For IEVD gives us a way to distinguish between two different interpretations of TEs' results and allow the externally valid results to be different from the internally valid results. In addition, the appeal to the concept of representation in DEKI's terms allows us to re-connect TEs with the empirical world in a straightforward way. A further, interesting consequence of my account is that TEs can thus sometimes be relevantly independent of our scientific theories. Like scientific models, they will then be able to play an autonomous, auxiliary role in bridging the gap between theories and phenomena. In this way, we are able to answer El Skaf's concerns optimistically.

2.5 Summary of the chapter

In this Chapter, I proposed to re-frame the debate on TEs on the basis of the distinction between internal and external validity, borrowed from the parallel distinction employed in the philosophical literature about material experiments. I illustrate the

distinction by analysing Galileo's thought experiment of a ball rolling in a V-shaped cavity. Then, I provide two detailed accounts of internal and external validity of TEs, respectively. I suggest that we should think of the former in terms of Walton's games of make-believe, and that the valid exportation of the internal results of TEs to external-world contexts is best interpreted as a process of accurate representation. On the basis of this diarchic account, I have provided an answer to Kuhn's initial questions: TEs are games of make-believe that provide knowledge about the real world by representing their targets accurately. Fictions are best interpreted in terms of Walton's games of make-believe, and the concept of accurate representation is clarified by the use of the DEKI account of representation. Finally, my account offers the opportunity to re-interpret previous positions and disagreement by providing a common conceptual framework. With my diarchic account of TEs' validity in place, I could both do justice to the different voices introduced in section 2.1 and establish a fruitful dialogue between them. Especially, the account explains the reasons supporting both camps, while at the same time giving an escape route to the impasses produced by the radically different views animating the debate.

Chapter 3

Model organisms as scientific representations

3.1 Model organisms, models, and representation

Recent debate in philosophy of biology shows considerable disagreement over the role of so-called model organisms (MOs), a group of organisms intensively studied in biological and medical research. Characteristic examples of MOs are the fruit fly (*Drosophila melanogaster*), the nematode worm *Caenorhabditis elegans*, several strands of mice (for example, *Mus musculus*), the plant *Arabidopsis thaliana*, and the bacterium *Escherichia coli*. MOs constitute one of the main instruments for discoveries in biological research and still occupy centre stage in the investigation of several medical conditions in humans (cf. Ankeny and Leonelli 2020 and references therein). It is common to use MOs to draw inferences about properties of other, often very different, organisms – unless specified otherwise, in this Chapter I will call the latter organisms *target organisms*, and the inferences from MOs to target organisms *MO-based inferences*.

One important philosophical question that results from this inferential activity in the biomedical sciences, then, is how to justify MO-based inferences. Rather than addressing the question of justification directly, though, the discussion of MOs in the philosophical literature tended to focus on whether MOs are scientific models. This way of approaching the issue becomes understandable once the two questions are related by the observation that both models and MOs are surrogate systems used to draw inferences about other objects. In the case of models, we draw inferences about their target systems, and in the case of MOs, about other organisms. It is often emphasised that models are idealised versions of their targets, meaning that models end up being different than their targets. The same holds for MOs, which are used to draw inferences about very different organisms. The philosophical discussion of models has provided us with an understanding of the issue of justification concerning model inferences: inferences from models to targets are justified if the model is an

accurate representation of the target. Hence, if we take MOs to be models, this justificatory strategy carries over to MOs: a MO-based inference is justified insofar as the MO is an accurate representation of its designated target. This puts the spotlight on the questions whether MOs are representations and, if so, how exactly inferences from MOs to other organisms can be justified in representational terms.

I identify three relevant philosophical positions in the current debate concerning the relation between MOs and models. Ankeny and Leonelli (2011, 2020), take MOs to be a type of model. They support this claim by assuming that MOs function as representations, where representation is understood along the lines of the DEKI account (Frigg and Nguyen 2020, pp. 159-213). Then, there are two non-representational accounts of MOs. First, Levy and Currie (2015, 2019) explicitly deny that MOs are representations like models are, and classify them instead as specimens of a larger class of organisms. Second, Weber (2004) takes representation to be a non-primary feature of MOs and focuses instead on their use as tools to develop new experimental techniques. This leads him (2014) to conclude that there is a substantial difference between MOs and models.

The central claim of this Chapter is that a representational view of MOs is correct, and that inferences from MOs to other organisms are justified by appeal to MOs representational capacities. In section 3.2, I introduce and further develop Ankeny and Leonelli's (2011, 2020) analysis of MOs. I argue that this account is on the right track but lacks an articulation of many important points, specifically concerning justification. I fill these lacunae by deploying the resources offered by DEKI to the case of MOs, with a particular emphasis on the concepts of exemplification and keying-up, which I have introduced in Chapter 1. Then, I turn to the two alternative views on MOs. In section 3.3, I discuss Levy and Currie's (2015, 2019) account and argue that they are mistaken in drawing a sharp distinction between MOs and models. Their view, I will show, is in fact the result of a general understanding of representation and models that remains substantially wanting. Finally, in section 3.4, I argue that Weber's (2004) account is compatible with the representation view defended in this Chapter once we drop his excessively demanding requirements on model-target relations. Also, I will suggest that his positive account of MOs as tools to develop new know-how is much more connected to the representation view of MOs than it appears at first sight.

3.2 The representation view of MOs, upgraded

3.2.1 MOs and DEKI

As the name suggests, the representation view takes MOs to function primarily as representations of other organisms. Ankeny and Leonelli (AL) start from the unproblematic assumption that MOs are a subclass of the vaster group of "experimental

organisms", namely any organism that is studied in biology laboratories. According to AL, what distinguishes MOs from the rest of this broader class "is the representational power attributed to them" (2020, p. 9). As a matter of degree, Ankeny and Leonelli would be happy to concede that this distinction between MOs and other experimental organisms is not qualitative, but only concerns the potential of using these organisms in order to produce new testable hypotheses about some other target organism (or class of organisms).

The representational power of a system, AL hold, has two conceptually distinct dimensions, namely the "representational scope" of an experimental organism, and its "representational target". The representational scope describes "how extensively the results of research conducted on a group of specimens [...] can be projected onto a wider group of organisms" (2020, p. 6). For example, the representational scope of a property observed in a laboratory population of, say, mice, may correspond to the class of all mice, or it could include other species (like humans), the entire family of mammals, or even all animals. However, what really makes MOs a special class for AL is their very broad and systematic representational target, by which AL mean the number and detail of mechanisms of other organisms that MOs can represent. Instead of representing only a specific mechanism – say, respiration, or flowering – MOs are taken to represent a very wide range of mechanisms, usually considered essential for a great number of species (*ibid.*, p. 8).

AL then try to find the roots of MOs' representational power in what they call the *repertoire*. AL's repertoire is defined as the entire corpus of background knowledge that informs and supports MO-based research, comprising various aspects of the research practice in biological and medical sciences. As paradigm components of the repertoire, AL mention general theoretical "principles (e.g., evolutionary conservation)", "the fit with other models such as simulations, diagrams, and mathematical models of development"; "pragmatic factors", like tractability and accessibility; and also methodological norms, institutions, and the collaboration between different laboratories (*ibid.*, p. 27).

The question whether MOs have more representational power than other experimental organisms can however be answered only once we specify what exactly "representation" means. AL fill this gap by adopting the DEKI account of scientific representation, developed by Frigg and Nguyen (2020, pp. 159-213).

As we saw in Chapter 1, the DEKI account takes representation to be a matter of interpretation rather than similarity. An interpreted object M is an epistemic representation of a target system T if four conditions apply:

(i) M *denotes* T,

(ii) M *exemplifies* properties P_1..., P_n,

(iii) P_1..., P_n are associated with a second set of properties Q_1..., Q_n via a *key*,

(iv) Q_1..., Q_n are *imputed* to T.

Before moving on, though, it is important to emphasise again that the key is a pivotal element of DEKI because it allows us to impute some properties to the target without M instantiating them. A model of a bridge exemplifies spatial measures that are translated by the key into another set of spatial measures by a scale factor. Newton's model of the solar system exemplifies orbits that are not exactly the ones then imputed to the real planets. Scientists have thus the possibility of approximating their predictions by "correcting" the model's properties. Approximations, limit functions, scale factors, and projections are all examples of keys in DEKI's terminology.

While AL explicitly endorse DEKI, they do not give much detail of how exactly MOs fit in the account. Aim of the next two subsections is to fill this gap. This will also allow us to address the question of justification, which remains mostly unexplored in AL's analysis of MOs.

3.2.2 Justification: the key and the repertoire

In their application of the DEKI account to MOs, AL pay little attention to exemplification and the keys. Concerning exemplification, AL only briefly mention it in the context of their illustration of DEKI (2020, p. 26), but they remain non-committal about which properties are exemplified in certain contexts and how. The same happens with the keys, as there is no clear example in AL's book of such a mapping function in the context of MO research.

In part this may be a consequence of the fact that AL define the key in a way that diverges considerably from the way in which it was initially conceptualised. For AL identify the key with their repertoire, which, as we have seen, contains the entire theoretical, pragmatical and institutional background that inform our scientific practices on MOs. Of course, the repertoire as a whole cannot count as an interpretive key of any sort, as it would be not specific enough to translate information about a MO into information about a target system.[94] In a charitable reading of their work, I thus take AL to say that the key is drawn from the repertoire and can in principle include any of the theoretical, empirical and pragmatic element that the repertoire contains.

Still, this idea of a key seems pretty different from the original notion proposed by Frigg and Nguyen, who identify the key with a mapping function that associates the properties of the model and the properties that are eventually imputed to the target. Certainly, keys can become extremely complex, involving many different steps. Sometimes they require multiple ways of property-mapping, depending on the specific property or the exact designated target system. As in the example that I have analysed in Chapter 2, the Galilean thought experiment of a ball rolling in a V-shaped cavity

[94] I thank an anonymous reviewer for raising this point in private conversation.

exemplifies the principle of inertia. The principle is only counterfactually true for most mechanical systems, but it becomes approximately so with interstellar objects, where the effect of friction becomes negligible, and not simply abstracted away (cf. Sartori 2023, pp. 9-10, 20). In any case, in Frigg and Nguyen's account, keys always remain mapping functions, and as such they are not equivalent to AL's repertoire, nor to single elements of it, nor finally to combinations of different aspects of the repertoire.

To keep the two distinct, I refer to AL's keys as repertoire keys, and to Frigg and Nguyen's as local keys. I suggest that the two keys are useful for two different purposes, both concerning the justification for our inferences from MOs. Yet, I also argue that the repertoire key remains inert as regards justification if a clear specification of the local key is not provided.

To see how keys function in justifications, it is helpful to distinguish between two different justificatory tasks. The two tasks concern two different notions of correctness applicable to inferences drawn from a representation, which I introduce in Chapter 1 *derivational* correctness and *factual* correctness, introduced by Frigg and Nguyen (2022, p. 296). The reader will remember that derivational correctness is entirely depending on the representational system at stake: an inference is derivationally correct if it follows from the interpretation given to a certain carrier and it applies the key correctly. Still, derivational correctness does not imply that the target system actually possesses the properties that we eventually impute to it. An inference from a representation is then factually correct if the target system also possesses the property we want to impute to it. Importantly, we can now see that factual correctness defined as above entails what I called representational accuracy in section 1.6). So, if our inferences from a model about the target having a certain property P are factually correct, then that model is also an accurate representation of the target relative to P.

Given that there are two types of inferential correctness, there will also be two corresponding types of justification for our inferences from a representation. The first type of justification concerns the question whether the inferences we draw in a representation are correct according to the representational framework set by the representation itself. For example, when I read a map, I have to interpret it as the legend says I should. Frigg and Nguyen hold that DEKI's key is the locus of justification for derivational correctness, as it specifies the rules to interpret the representation and to perform inferences about the target. These rules are associated with the representation itself, like a legend is associated with a map. In order to understand the map, I have to read it in the light of the legend associated with it. To this analysis, I add that also the interpretive function I of the object as a Z-representation and exemplification play an important role in this context. The interpretation as a Z gives a way to read the material properties of the model carrier as a symbol, and exemplification provides the necessary input for the key, which is

just a set of formal instructions. The map, *qua* territory-representation, exemplifies certain properties – say, distances and topological relations – and exclude others – say, altitude. The legend specifies how the exemplified properties are to be translated into information about the actual territory. Of course, nothing rules out the possibility that the map be wrong: the "X" indicating the treasure on the map could just point to an empty cavern. This is tantamount to say that misrepresentation still counts as representation, just factually incorrect.

The justification for factual correctness, on the other hand, is an altogether different matter. For factual correctness concerns the truth-values of the claims that we obtain from the representation through the imputation of the Q-properties to T. In my previous example, we are not asking anymore whether the map says that the treasure is supposed to be where the "X" is, but rather, whether the treasure is in fact there. Frigg and Nguyen insist that these truth-values "are not something that the representation adjudicates, let alone justifies" (*ibid.*, 297). They argue that to justify the factual correctness of a representational inference, we need to look outside of the representation. This can be done in many, not mutually exclusive ways. If we can, we perform an observation or an experiment directly on the target and see whether the claims bear out. In our example, we go and check if the treasure is buried where the map indicates it is. If we cannot check directly, other ways to justify our inferences from a representation are available. We may, for example ask whether the results obtained from the representation are compatible with the rest of our empirical and theoretical knowledge. Also, in a specific epistemic context, we may know that the principles of generation used to interpret the carrier and the relative keys have proved to be particularly successful in that type of context. Using Goodman's (1983) terminology about predicates, we can then say that our keys are taken to be projectible: they are well entrenched with the development of those specific scientific practices.

Now, let us recall that AL's main goal was not an account of justification for MO-based inferences. They wanted to spell out all the factors playing a relevant role in contemporary MO-based research, with an emphasis on institutional, organisational, and pragmatic dimension of these factors. Yet, their concept of repertoire is clearly relevant for the problem of justification: in the absence of direct experimental tests on the target, AL's repertoire can furnish arguments for the factual correctness of our inferences. This is because the repertoire encompasses all potential elements and factors that may be relevant to the justificatory analysis of a representation's factual correctness: overarching theoretical principles, currently employed models in the field, and empirical data. AL's key, then, is not a key, but an indication of how to justify a certain local key in Frigg and Nguyen's original sense.

The importance of this type of justification is indisputable. However, and this is my main point here, in order to assess the factual correctness of a representational inference, one has to specify what counts as derivationally correct. Namely, we need

to clarify what the MO exemplifies, and what key is involved.

Assume we have a MO: we observe an interesting property in it and want to export our results to other organisms of interest, like humans. Before drawing on the repertoire to justify the factual correctness of an inference, we must specify what we are actually inferring. To this end, we need to spell out what local key we employ and what properties the key representation is imputing to the target. We can ask whether the property attribution is correct only once we are clear about what the property in fact is.

This is tantamount to saying that we have to specify the semantics of our representation (what does the representation mean?) before delving into the epistemology of it (is a representation's claim about the target true and justified?). This, in turn, is equivalent to saying that we have first to ascertain the derivational correctness of a representation, and only afterwards try to justify the claims thus generated on the basis of the repertoire.[95]

The impossibility of moving to the level of factual correctness without dealing with derivational correctness may become even clearer if we compare the case of MOs with the case of models employed in mechanics. Mathematical models in mechanics exhibit, among a vast array of keys, a considerable use of limit keys, a use that has been extensively spelled out (Nguyen and Frigg 2020, pp. 195-203). Limit keys have a precise way to associate limit values of quantities in the model to more realistic values in the target system. Thus, it is clear how to interpret our models in mechanics because, while models exemplify values at the limit, we have a key to take these values back to other values. Once the key is specified, and thus the derivational correctness of the results is established, one can proceed and offer a justification for their factual correctness. Here, we do not always need to perform direct experimental tests on our targets to know that our models are accurate – for example, when we have to send rockets to space. Contextual knowledge (AL's repertoire, but in the context of mechanics) is usually sufficient to justify our inferences from mathematical models used in mechanics because we have specified that limit keys come into play and how we are expected to deal with them. In contrast, we could not do this in the case of MOs because we lack a precise illustration of the derivational correctness of our inferences – which properties are exemplified and what key is employed. Without such a clarification, it is impossible to employ the resources of the repertoire in the first place.

In sum, before we can provide a justification for the factual correctness of our inferences from a representation by appealing to the repertoire, we need a clarification of what properties are exemplified, and what key is employed. In the next section I give examples of how this can be done.

[95]The temporal adverbs here only express a logical priority for a philosophical analysis of the justification for these inferences. In practice, the repertoire and the local keys continuously interact with each other (see below).

3.2.3 Exemplification and local keys in MOs

To understand exemplification and the key more precisely in the context of MOs, let us consider the case of *Drosophila melanogaster*. This MO was studied in order to understand the mechanisms of the so-called chromosomal crossover – the exchange of genetic material during sexual reproduction between two homologous chromosomes' non-sister chromatids. These studies, as well as their implications for genomic selection, have been crucial to understand the same mechanisms in more complex organisms; among them "moths, pigeons, cats, silkworms, rabbits, and several species of plants" (Levy and Currie 2015, p. 333).[96] The *Drosophila* then instantiates and exemplifies the mechanism of chromosomal crossover. This is because, besides exhibiting the crossover itself, it allowed scientists to epistemically highlight the mechanism, due to giant chromosomes isolated from larval salivary glands of the *Drosophila*. In this sense, the concept of exemplification makes sense of the fact that scientists gained better epistemic access to certain previously unknown aspects of the crossover.

As we have seen in section 1.3, exemplification is not an intrinsic feature of a system: it depends on the context and the interpretation with which we endow the system. The fruit flies that get inside your kitchen, for example, do not exemplify anything *per se*. The fact that *Drosophila* exemplifies chromosomal crossover crucially depends on the interpretation of a *Drosophila* population as a genome-representation, as well as the modifications and controls applied to the laboratory populations on the part of the scientists. According to DEKI, the interpretation of an organism as an exemplar of a property must be coupled with also denotation and property imputation via the key. Denotation and imputation are granted by the empirical fact that the scientific community used and still use *Drosophila*'s genetic mechanisms to generate hypotheses about other species.

Let us then turn to the key. It is important to recall that keys in DEKI simply provide a way to read our results when we try to export them to the target system. As part of the representation system, therefore, keys can be wrong.

Also, some representations, like some modern maps, have their keys stipulated from the start. This is usually not the case with MOs. They are no exception though in the domain of scientific representation: it is often difficult to understand how to translate the properties of a model onto properties of a target. This is just a version of the general problem of external validity of our models and experimental results, as we have encountered it in Chapter 2.

A further complication for an analysis of keys in MO research is that current scientific works on MOs often are not clear on how exactly we learn from a MO about other organisms. Scientists often leave this implication implicit. For instance, in a recent study of the neurobehavioural impairment caused by anaesthetic drugs

[96] For more extensive analysis of the role of *Drosophila* in genetics, cf. Weber (2004, sections 3.2 and 6.1) and Oriel and Lasko (2018).

in *C. elegans* (Nambyiah and Brown 2021), the results are explicitly taken to be relevant for humans. Nevertheless, the article gives little detail on how exactly the behavioural features of the worms translate into behavioural features of humans.

The fact that keys do not come out explicitly from biomedical research practice on MOs, however, does not entail that keys are not in fact required and implicitly at work in that context. For instance, the dosage of a certain substance (like a drug, or a toxin) necessary for a specific effect (say, healing from a disease, or having substantial impairment) will of course have to be adapted from the MO to the case of humans, given the clear physiological differences (e.g., dimensions). One way to do this is to multiply quantities on the basis of, say, body weight and other basic differences between the MO and the target organism.

In other cases, we have model values that tend to a limit and then have to be re-adapted by the sort of limit keys that Nguyen and Frigg (2020) have studied in the context of mechanical models. For example, Seim (2019) discusses the case of an immortalised cell line of macrophages (called RAW 264.7), in which cells keep undergoing division for their entire life, making their population grow to infinity. This does not happen with normal cells and must be considered when the results are extrapolated to non-modified organisms.

3.2.4 Special keys in MO research

As we have seen, we have reasonable grounds to think that some easily conceptualised keys are already identifiable in many cases of MO-based inferences, like multiplying factors and limit keys. However, most interesting discoveries in MOs require even more complex keys. This complexity is a challenge for both the scientist and the philosopher and calls for a deeper conceptual analysis.

An interesting example of a key which seems highly peculiar to biology research in general, and model organisms in particular, can be found in the review of Moretti *et al.* (2020), which reconstructs the results obtained from the study of the three-dimensional organisation of genome in *Drosophila*. How the genome bends and arranges itself spatially is crucial for processes like the "regulation of gene expression during development, cell differentiation, and cell identity maintenance" in many metazoans (p. 92). Now, there are many relevant differences between *Drosophila* and humans in the way their genome organises in three spatial dimensions. I focus on one difference specifically, which concerns the set of architectural proteins responsible for the shape of the chromosomal bending.

These proteins in *Drosophila* are different from the proteins in other organisms (in particular, humans). For example, "dCTCF, the main driver of TAD [topologically-associated domain] formation in vertebrates [...] is only found [in *Drosophila*] at 28% of TAD borders, with no evidence for a specific motif orientation, contrary to what is observed in vertebrates" (p. 95). Moreover, "[i]n flies, other architectural proteins such as BEAF-32, CP190 and Chromator are probably more important than dCTCF

in TAD boundary formation" (*ibid.*).

So, we can see that the set of architectural proteins responsible of the chromosomal topology in *Drosophila* are associated with different, yet functionally analogous proteins in the humans. This association is done via a key, which maps the *Drosophila*'s architectural proteins to corresponding proteins in humans. The two sets of proteins are different, as is also the overall mechanisms, and of course the resulting chromosomal topology. However, the general mechanisms have the same function – shaping the genome 3-dimensionally. Also, the two sets of proteins serve the same sub-function within their respective mechanism. The reference to mechanisms will also allow for different levels of detail, depending on different goals set by the relevant scientific enquiry.

I call *functional identity key* (FIK) this novel type of key. More precisely, a FIK is a function mapping a (set of) elements E_1 of a mechanism M_1 in the model system onto a (set of) element(s) E_2 of another mechanism M_2 in the target system, where M_1 and M_2 have the same overarching function, and the elements E_1 and E_2 associated by the FIK are identical with respect to their sub-function within their respective mechanisms.

This characterisation is intentionally flexible in order to be adaptable to different conceptualisations of "mechanism" and "function" that may play a role in biology. Nevertheless, the reference to the concepts of function and mechanism is by no means arbitrary or vague, as it can count on a vast and deep philosophical investigation (cf. Huneman 2013, Nicholson 2012, Wouters 2003). More specifically, the idea to refer to mechanistic organisation and structure, combined with some information about the associated biological functions of such mechanisms, when it comes to comparing different species of organisms is well established in the philosophical and scientific literature, with obvious differences across the various positions (see DiFrisco *et al.* 2020 and references therein). It is also important to specify that here I do not make any specific ontological assumptions, limiting myself to understanding mechanisms as theoretical descriptions (not necessarily linguistic ones) of possible causal structures. Given that DEKI already implies an interpretation of the carrier, it takes a model to always involve a model description. It is this description of the interpreted carrier that includes the exemplified mechanisms. These properties are then mapped via the key onto properties that constitute the final description of the target (the set of propositions describing the imputed properties). In this sense, the material genes, cells, and tissues of the MO are connected to the corresponding elements of reality of the target, but this connection is mediated via mechanistic descriptions.

Drug dosages, limit keys and functional identity keys provide a first illustration of what kind of local keys we can find in MOs. This list is of course not exhaustive. Different keys are at work when MOs are used to find correlations between genes and diseases, a common application of MOs. For example, some variants of *C. elegans* exhibit a gene homologous to the human gene BRCA1, which is now known to be

associated with human breast cancer (Ankeny and Leonelli 2020, p. 8). A key then is required to translate the gene-cancer correlation in *C. elegans* into the gene-cancer correlation in humans. This will require a complex key that associates, on the one hand, homologous genes with each other via phylogenetic relations, and different types of cancerous cellular development on the basis of comparable cellular development mechanisms, on the other.

When I talk about phylogenetic relations, we can distinguish two levels. On the one hand, there is a key that translates a gene and its correlation to a phenotype to another, homologous gene. Homologous genes are phylogenetically related in the sense that they are the result of the evolution of one more ancient gene. In this sense, we have a local key that we could call a phylogenetic key, that is, a function that associates a gene g_1 with the class of genes homologous to g_1. This kind of keys are definitely used in biology and specifically for MOs.

This is one level of philosophical analysis. A second level, instead, concerns the justification for a MO-based inference. Suppose we take a mechanism present in one organism to be present as it is in other organisms phylogenetically related to the first one. We can see here that the key at play is a simple identity key. The justification offered for the factual correctness of our hypothesis, however, strongly relies on phylogenetic assumptions. I will come back to phylogenetic assumptions later in the Chapter when I discuss Levy and Currie's account of MOs. It is important to emphasise, however, the difference between the two levels in which phylogeny can play a role in our MO-based inferences: in one, we recognise as part of the interpretation of the MO system the genetic difference from the target system, and we employ a key to correctly associate the gene in the MO to another gene or class of genes in the target system. In the second case, the key is not our concern, but we are instead reasoning to support our imputation by looking at information that come from "outside" the single MO under consideration.

3.2.5 Implications

All these examples reveal a general pattern relating MOs to their targets. MOs denote other biological systems, which grounds their use as surrogates to draw inferences about their targets. MOs drive these inferences by exemplifying specific properties, and thus allowing epistemic access to them. The imputation of related properties to the target system is done via a key that correlates properties of the MO with the properties of a designated target system. So, we see that we have all four conditions of DEKI satisfied, and we have a general framework to investigate each specific case study. Hence, MOs represent other organisms in the sense of the DEKI account, which also provides an answer to the question of derivational correctness: our MO-based inferences are justified insofar as we impute properties to a target that are exemplified by MOs, by applying the proper key function.

The justification for the factual correctness of our results, if any, remains largely

extrinsic to the specific representational system, and can be achieved only by looking at the repertoire, which constitutes AL's repertoire key. Once we have a local key at work, the repertoire will provide the contextual knowledge to justify the claims that it generates. In the case of a FIK, the repertoire provides the theoretical knowledge and empirical evidence to justify the functional equivalence of mechanisms and their elements. For example, in the case of the *Drosophila*'s 3D genome organisation, the repertoire will give us the grounds to associate different sets of architectural proteins, even though those sets of proteins diverge at a considerable extent.

In practice, the relation between the repertoire key and the local key becomes a process of continuous feedback between the specific characteristics of the specific representational system and a more general conceptual framework. Scientists develop keys on the basis of their background knowledge but at the same time also try to find reasons to ground their local inferences from their models in further confirmation from further, independent means of investigations (experiments, simulations, models). The two levels, the local key and the repertoire, are of course intertwined and may end up redefining each other, until they eventually reach a reflective equilibrium. This form of holism, as we have already seen with thought experiments in Chapter 2, is just an intrinsic feature of scientific knowledge and of knowledge more generally, as theorised by Elgin (1996).

While this form of holism is inescapable, and the two levels of justification that I described here are in endless interaction in everyday scientific practice, it is still important to keep them conceptually distinct in order to achieve a better understanding of our inferences from representations and their justification.

Also, it does not seem to be the case, as AL hold, that MOs generally exhibit greater representational power than other experimental organisms. For representational power becomes a function of the set of exemplified properties and the local key. For example, taken as a 3D-genome-organisation model, *Drosophila* has a very narrow representational target (in AL's sense). More generally, even if a taxon is used to represent numerous mechanisms in several species, DEKI pushes us to distinguish different studies performed on that taxon, depending on which specific properties are exemplified and which keys are employed.

Finally, one may wonder what is exactly that counts as MOs' representational carrier: the species, a particular strain (or a number of strains), a laboratory population, or an individual organism? From what I said so far, a natural answer is that what counts as the carrier depends on the specific context, the purpose of investigation, and the assumptions of the relevant epistemic community. In the case of MOs, the carrier is usually identified with a laboratory population, because it is that population that has undergone the procedures of selection that allows it to exemplify certain relevant properties. Yet, as AL (2020, p. 31-33) also argued at length, different laboratories adopt common standards to select and modify the same MOs in order to share their results – ending up creating what AL call "worm community", "*Arabidopsis*

community", and so on. Insofar as this ideal of shared standards is approximated by different research groups, the carrier becomes the entire set of the MO's laboratory populations complying to those standards.

Finally, one may ask why I keep calling "representation" the MO as a whole, or the MO laboratory population, if what we need is just the set of exemplified properties and the key to translate them. For what concerns the MO as a whole, the reason is that the properties that a MO exemplifies are usually inseparable, at least in a practical sense, from the rest of the MO's properties. As other types of models, MOs are often non-modular: we cannot "extract" the exemplified properties without keeping into consideration the relation of these properties with the others possessed (but not necessarily exemplified) by the MO under study.

This holistic nature of MOs is also considered one of the greatest epistemic advantages of complex, *in vivo* representations with respect to abstract models and computer simulations: they are expected to provide a much more complete picture of the complexity of how a certain mechanism relates to other mechanisms.

Similarly, when it comes to the key to use, one will have to consider the context in which a specific property is embedded. The same reasoning extends naturally to MO populations: they may exemplify relevant properties only in a statistical way, thus not reducible to observations of individual organisms. All this is captured by the DEKI account: the representation is not the final outcome of our investigation, but the entire model system that, as a whole, exemplifies only certain properties among the ones it instantiates.

3.3 Levy and Currie's account and its difficulties

Let us now turn to the main accounts of MOs that seem to conflict with the representation view that I have just presented. The first account I consider is developed by Levy and Currie in their article "Model Organisms Are Not (Theoretical) Models" (2015).

From the beginning, Levy and Currie (LC) specify that their "discussion doesn't touch on ontological or semantic questions, such as what models are or how they represent [... Our] aim is to account for the justificatory structure underlying the inferential move from models to targets" (*ibid.*, p. 329). So, their focus is on justification, but in contrast to the view presented so far, they separate this issue from semantics and, more precisely, from the fact that MOs represent and how they do so. They go on to argue that "inferences from work on [MOs] are empirical extrapolations, whereby biologists treat the organism as a representative specimen of a broader class" (*ibid.*, p. 332). They note that, while inferences made from MOs "are broadly model-like, [...] they diverge in their epistemic roles from theoretical models. The type of stand-in at issue is different" (*ibid.*, p. 336).

Let us look at the terminology used more closely. LC take the expression "theo-

retical model" to denote any scientific model, from material scale models of bridges to mathematical models like the Lotka-Volterra model (*ibid.*, pp. 329-31).[97] In their (2015), LC do not give a detailed illustration of their concepts of "specimen" and "representative", but they do so in a later paper (Currie and Levy 2019): they define a specimen as a "typical instance", where being typical "can be understood in terms of similarity – in the limit, sameness – of focal properties" to the other members of the relevant class of extrapolation (*ibid.*, p. 1072). In the same article (*ibid.*, p 1078), they also explicitly contrast being a specimen with being a model representation:

> a specimen is a *representative instance* of the target. But this sense of "representation" is critically different from that applicable to models. Experimental systems [...] represent similarly to how statistical samples do – by being not unusual subsets of the larger class [...] This is not representation in an intentional sense, and the difference is reflected in the epistemology: in a successful experiment the object is a specimen, and confirmation is possible because it has been procured via an unbiased procedure [...] In contrast, a model represents the world by being *about* it.

Because they talk about intentionality, LC here distinguish specimens and models on a semantic level too, and this would in turn affect the epistemological nature of MO-based inferences, in contrast with what they say in their 2015 paper. Therefore, LC's overall distinction between MOs and models leaves some room to interpretation.

I suggest that there are at least two plausible readings of their claims. On the one hand, strictly following their 2019 paper, one can read LC as arguing that there is an essential difference, both semantic and epistemological, between specimens and models. Alternatively, one can read LC as making the weaker claim that, while specimens are representations in a loose semantic sense, MOs and models still exhibit important epistemological differences concerning the justification for the inferences that we draw from them about their targets.

In section 3.3.1, I show that the strong interpretation of LC's distinction does not stand up scrutiny, irrespective of the philosophical account of representation one adopts. The weak interpretation needs some further unpacking of the claims that the two authors make in their 2015 paper, and in section 3.3.2 I look at the arguments that LC offer to support their epistemological distinction. My conclusion is that these arguments fail to show that there is a principled distinction between justification for model inferences and justification for MO inferences, the difference resulting to be at most a matter of degree.

Therefore, for both interpretations of LC's point, I show that their arguments remain wanting and their views do not provide an accurate understanding of the use of MOs particularly, and models more generally.

[97] For details on the model, see Weisberg and Reisman (2008).

3.3.1 Specimen vs. representation

Let us now scrutinise the stronger reading of LC's thesis. On this reading, a MO is a specimen of a target class $\mathfrak{T}$ *iff* it instantiates focal properties similar to the properties instantiated by the members of $\mathfrak{T}$, where a focal property is a property regarded as relevant in our extrapolation. Finally, specimens are not intentional – they are not about their targets – while models are.

The first difficulty for this account is that, as I argued in section 3.2.3, not all MOs work with a simple identity key, so, not all MOs are literally specimens of $\mathfrak{T}$, or the relevant biological kind. For we need a key to translate the exemplified properties into the imputed properties, and $\mathfrak{T}$ is defined by the latter. So, *contra* LC's claim, MOs often do not function as specimens in their sense.

The second difficulty is that, even when MOs do function as specimens, it is just wrong to say that they are not intentional: if MOs are studied in order to formulate hypotheses concerning the other organisms in $\mathfrak{T}$, then the results of our investigations on a MO are in effect *about* those other organisms. And this is the case for all experimental specimens that are used to formulate hypotheses about other systems. But this is exactly the meaning of intentional in this context.

Certainly, an experiment can be performed on a system because we are interested only in that specific system. In this sense, the experimental system, or the experimental population, is not intentional in the relevant sense.[98] However, it is not clear if such non-intentional experimental systems still function as a specimen, at least according to Levy and Currie's definition: for it has no extrapolation class, so it cannot be a typical instance of anything.

Now, it is certainly true that not all representations are specimens in LC's sense. A painting of a horse is not a member of the class of horses. However, it seems clear that the converse holds: all specimens, including MOs that require non-identity keys, are representations. And this is all I need to undermine the strong interpretation of LC's claim that MOs' semantics is qualitatively different from the semantics of models.

The fact that specimens are representational symbols is indeed a cornerstone of DEKI, as well as the representation-as view (Goodman 1976, Elgin 1983, 1996): both hold that one of the main aspects of representation is exemplification, of which specimens are paradigm instances. The turquoise patch in a draper's window represents the turquoise clothes in the shop by being a specimen of turquoise clothing, that is, by instantiating that colour and referring to it given the contextual interpretation of the relevant agents.[99] In fact, in these accounts, specimens are regarded as paradigm examples of representation.

[98] The same occurs in the case of targetless models – e.g., cf. Weisberg's (2013, § 7) models of four-sex organism populations, Norton's dome (2008) and Hartmann's (1995) toy models in chromodynamics.

[99] Cf. Goodman (1976, pp. 52-56) and Elgin (1983, pp. 71-95, 1996, pp. 171-86).

Nevertheless, LC (2019, p. 1073) attempt to drive a wedge between their "specimens" and exemplification:

> [I]n contrast to Elgin, we place weight on *how* a specimen was obtained [...] a specimen is an object drawn (in an unbiased way) from the world. Assumptions about how it was obtained [...] matter for justifying conclusions drawn from the specimen regarding the object of study.

Here, two aspects would distinguish specimens from Elgin's exemplars (that is, exemplifying systems): specimens are (1) drawn from the wild via (2) an unbiased selection process. However, on a closer look, the purported difference dissolves. I postpone a discussion of the epistemological value of (1) to section 3.3.2. For the purpose of analysing the strong interpretation of LC's distinction between specimens and models, however, we can recall that nothing stops Elgin's exemplars from being drawn from the wild too. Later, I will show that the distinction though is far less insightful than it may appear at first sight.

For what concerns (2), LC characterise specimens as obtained via an unbiased selection process. In general, as LC put it, "statistics provides methods for making such selections", and we should "understand an unbiased selection process as one that reduces the risk of selecting an unusual object, and which preserves typicality [...] relative to the aims of the experiment" (*ibid.*).

However, statistics does not help us much here, because a statistical analysis alone won't tell us whether, say, *Drosophila* is a specimen of a class $\mathfrak{T}$ that includes humans. For whether something is or isn't a specimen depends on what features we focus on and what reference class we consider. These are decisions we have to make prior to any statistical analysis. And it may well be that the same organism is a specimen with respect to feature F_1 and reference class C_1, but not with respect to feature F_2 and reference class C_2. Thus, once we have established that something is a specimen, then it is automatically also an exemplar (in Elgin's sense) that eventually proved to be successful for a specific purpose.

In addition to this, one can also see that the concept of specimen as tailored by LC seem inadequate to capture the actual meaning of experimental specimen so that it makes sense of the rationality of scientists' practices. Let me illustrate this via a thought experiment. Let us consider, once again, a group of fruit flies flying around in your kitchen. Let us assume that they are a truly random sample of all *Drosophilae* on the planet, and the selection was done by other people. Does this group of fruit flies count as a specimen? From the point of view of LC, it meets the criterion for being a specimen. However, it seems that until someone does not make use of the properties of the sample in order to conduct inferences about another group, and decides which properties are the salient ones, the fruit flies in question are still not counting for actual specimens, only potential ones. Here, again, we see that the statistical aspects and the objective properties of the fruit flies are not

enough to make sense of the concept of experimental specimen as we normally intend it. What misses is the referential relation, and thus the intentionality: without the interpretation of the sample as a sample, without the use of the properties of the sample to make hypotheses about an extrapolation class, it does not seem right to talk about experimental specimens yet.

Since I have so far focused on the representation-as accounts and DEKI, one might worry that I am assuming a concept of representation that already presupposes my conclusion that specimens are representations. But, luckily for those who do not like these accounts of representation, the same conclusion can be reached from all main accounts of representation. And this is important in order to at least create a general common ground with people belonging to different camps in the debate on representation. Let us briefly look at the other main families of views on representation and what they would say about specimens.

The accounts of representation based on similarity (see Weisberg 2013 or Giere 2004) generally hold that, for a model M to be a representation of a target T, it must be the case that M is similar to T (namely they share some properties) and that an agent uses this similarity for certain (epistemic) purposes. Now, by definition, LC's specimens instantiate the focal properties of the target class, and scientists definitely use these similarities to draw inferences about that target class. Therefore, in the framework of the similarity views, LC's specimens are representations.

The same holds for structuralist accounts (see Bueno *et al.* 2012, Da Costa and French 1990, French and Ladyman 1999), which define representation in terms of (partial) isomorphism (or homomorphism) between the mathematical structure of the model and the mathematical structure of the target. Given the broad way in which an object can instantiate such a structure, here too a specimen of a class represents that class insofar as it instantiates structural properties that are also possessed by the members of the class. Interestingly, if one is a realist about mathematical structures, we have the converse implication that all representations, by instantiating a certain mathematical structure, will also be specimens of the class of objects instantiating that structure.

Finally, the inferential accounts of representation – see Hughes (1997), Suárez (2004) and Contessa (2007) – normally require representations simply to allow inferences about their targets. In these accounts too, there is no principled reason to distinguish specimens from other types of representations, as the former also function as epistemic surrogative systems.

In sum, not all MOs are specimens in LC's sense, all specimens are intentional, and all philosophical accounts of scientific representation unanimously recognise specimens as paradigm instances of representation. Therefore, the strong interpretation of LC's distinction between MOs and models, at least as semantic and not only epistemological, is untenable.

3.3.2 Are MOs different from models?

Let's now turn to the weaker interpretation of LC's characterisation of MOs. LC's point would be that, even if we grant that both MOs and models are representations in a loose sense, there is still an important difference in their epistemology, namely in the way we justify the inferences we draw from them. They offer arguments in support of this claim in their 2015.

First, they claim that while models are intrinsically "idealized constructions", MOs, "in contrast, are drawn from a wild population" (*ibid.*, p. 334), a point that we have already encountered in the previous subsection.

Second, they say that a model's "properties are either wholly stipulated or specified so as to represent some target", which is the reason for "the modeller's intimate knowledge of, and high degree of control over, the model's] makeup" (*ibid.*, p. 331). This is contrasted with the lesser amount of knowledge biologists possess about MOs.

Third, LC hold that "theoretical models are assessed for structural resemblance to real world targets" (*ibid.*, p. 337), and the relation between models and their targets is a "direct comparison" (*ibid.*, p. 339), "grounded in an explicit procedure of feature-matching" (*ibid.*, p. 336). Instead, in "model organism work, the inference from model to target is mediated via indirect evidence [...] One kind of indirect evidence is what we have called circumstantial evidence, the other is shared phylogeny" (*ibid.*). While LC admit that phylogeny is not the only way in which MOs relate to their targets,[100] they repeatedly highlight that this form of inference "sets apart [MO] work from other kinds of theoretical methods" (*ibid.*) and that standard use of MOs "is best understood as an application of phylogenetic inference" (*ibid.*, p. 339).

We can then summarise LC's position in three main claims:

(a) MOs are drawn from the wild while models are idealised constructions.

(b) The full specification of models' properties allows for a more intimate examination of their properties than with MOs.

(c) Models are directly analogous to their targets, while MO-based inferences are typically mediated by phylogenetic assumptions.

On the basis of these three claims, LC insist that MOs are relevantly different from models. These claims do not affect my overall application of DEKI to MOs. However, as LC's differences would bear on the justification for MO-based inferences, I also need to show that these differences should not worry us. So, I argue that these differences are at best a matter of degree rather than principle.

[100] As Bolker (1995) and Gilbert (2009) argue, some emblematic MOs are clearly taxonomic outliers, as their genetic sequences often relevantly diverge from the ones possessed by their targets (Ankeny and Leonelli 2011, p. 318). Phylogenetic relations are sometimes just irrelevant: we have seen in section 3.2.3 that the FIK associates sets of proteins on the basis of functional identity. Fagan (2016, p. 133) also criticises LC's excessive emphasis on phylogeny.

Let us start with (a), which I will call the "materiality of MOs" to facilitate the discussion. First, materiality is neither a sufficient nor necessary condition for the justification for a MO-based inference. That *Drosophila* is drawn from the wild is not sufficient to justify, say, inferences from how the *Drosophila*'s genome folding works to how it works in humans. Nor it is necessary: we have seen that the material elements of the mechanism involved are different in *Drosophila* and humans, and we need a key to associate them.

Therefore, the burden of cashing out how exactly the materiality of MOs justifies the inferences we draw from them falls back on LC's shoulders. The crucial issue, I suggest, is not really whether MOs are drawn from the wild, but rather how their material features are interpreted for our inferences, an act of interpretation that DEKI captures nicely. As we have seen, sometimes the material properties instantiated by the MO are not exemplified; and those that are exemplified are sometimes translated into different ones via the key. Therefore, the justificatory role of MOs' material features is a matter of degree, depending on the context. But then, materiality does not amount to a clear-cut epistemological distinction between MOs and the rest of material models.

Let us move on to point (b), namely that models' constructed nature facilitate a more intimate, detailed knowledge and examination. First, as LC acknowledge (2015, p. 333), MOs too are idealised and controlled, as they usually undergo a sophisticated process of selection and genetic engineering. In addition, the experimental settings of MO laboratories are highly idealised. Illustrations of these elements of artificiality abound in the literature.[101] All this being said, it is also not generally true that (i) models have their properties wholly specified, nor that (ii) even if that was the case, this necessarily facilitates a more intimate examination.

Concerning (i), models' assumptions are seldom known from the start and have no fixed interpretation. The elements of the model have to be interpreted as standing-in for elements of reality, and which interpretation is the "right" one depends on the purpose of the representation (see also below, section 3.4.3). In other words, we need to use an appropriate key in each specific context. Therefore, the difference between models and MOs is again a matter of degree, and along this dimension, not all models are better off with respect to all MOs.

Concerning (ii), models are dynamic instruments, from which we constantly obtain new information. This is because scientists are not logically omniscient. It took physicists 200 years to realise that Newtonian models can exhibit stochastic behaviour (cf. Parker 1998). The more complex models are, the less they allow for the sort of intimate, detailed examination LC take for granted. Except for toy models like the original Lotka-Volterra model, models can be considerably opaque in their inferential patterns. So, it is not always easy to recognise the relations

[101] On *Drosophila*, see Kohler (1991, 1993); on *C. elegans*, cf. Ankeny (2000); on *Arabidopsis thaliana*, see Leonelli (2007).

between the different inferential steps, or what are the rules that govern the evolution of the model system. The concept of "opacity" has indeed become central in the literature on formalised models and computer simulations:[102] for example, Beisbart (2021) illustrates different levels of opacity via a case study from the science of climate, namely the Hadley Centre Coupled Model 3 (HadCM3). In this and many other cases, how the model "works" internally is not entirely known from the beginning: it must be studied in itself. Again, how much a model's properties can be transparent varies in degree. In this respect, many models exhibit a level of complexity that makes them not very different to MOs.

Let us now turn to thesis (c), which I take to be LC's strongest argument for there being a difference between MOs and models. LC argue that the inferences linking MOs to their targets, being mediated by indirect evidence and phylogenetic assumptions, are crucially different from the "analogical", resemblance-based, unmediated inferences that characterise models. I argue that, even when phylogenetic assumptions are in place, this does not constitute a clear rupture with other models. For virtually all model inferences are mediated by some assumptions and indirect evidence, exactly like MO-based inferences are mediated by phylogenetic assumptions and indirect evidence.

To begin with, it is worth recalling that analogy, in its technical meaning (see Bartha 2010), is just one among the many ways in which a model can relate to its target: approximation, projections, limit functions, and conventional rules are other distinct ways in which the properties of a model relate to the properties of the target.[103] Each of these model-target relations, moreover, is rarely based on "direct" or evident similarities, as LC seem to assume. On the contrary, basically any feature-matching activity involved in modelling is usually "mediated" by some assumptions, empirical or theoretical. For example, the system described by the Lotka-Volterra equations is not intrinsically similar to any real population's dynamics. LC take this simple mathematical model as a paradigm example of a theoretical model, so it is worth it to have a look at it more closely. The model describes a fictional system composed by two populations, one of prey and one of predators, whose dynamic is expressed by the means of the two following differential equations:

$$\frac{dV}{dt} = rV - (aV)P$$

$$\frac{dP}{dt} = b(aV)P - mP$$

Being V and P the number of prey and predators, respectively, the first equation says that the rate of change of V is equal to a certain natural increase of V by the growth

[102] See Beisbart (2021) and references within. In contrast with some authors (Humphreys 2009, Winsberg 2001), I do not acknowledge any philosophical novelty of computer simulations with respect to scientific models (cf. Frigg and Reiss 2009).

[103] For an overview on different types of model-target relations, see Frigg (2022, pp. 468-74).

rate r, minus the so-called functional response $(aV)P$, which accounts for the fact that predators eat the prey proportionally to the sizes of the two populations. The second equation is specular: m is the rate of natural death among the predators, and $b(aV)P$ is the so-called numerical response – i.e., the extent to which the predator population grows by eating the prey.

The model system describes the two populations in a very idealised and distorted way. As Volterra (1928, p. 6) himself noted, the population sizes are given in real values rather than with integer numbers; the populations constantly reproduce and are taken to be homogeneous, neglecting completely any difference in age, size, or other potentially relevant individual features. Furthermore, no characterisation of the environment is given.[104]

Then, given all these distortions, what is it that provides the theoretical justification for considering this fictional system as analogous, and relevantly so, to real populations? LC pass over this aspect of models in silence, but this is crucial for the tenability of their argument. We need basic assumptions concerning ecologic regularities and empirical observations on real populations in order to warrant an analogy between the model and the target populations.

Generally, the distortions in the model must be accounted for in some way. Either they are shown to be acceptable in a specific context – for instance, the absence of air resistance in kinematic models is deemed legitimate because it is negligible for some epistemic enquiry – or there is a key that translates the distortions in a meaningful property. For example, in the case of a Mercator map of the Earth, we have precise equations to convert the distorted distances between points on the bidimensional map into actual distances on the planet's curved surface (cf. Nguyen and Frigg 2022a).

As we have seen in section 3.2.4, phylogeny can enter the picture of our MO-based inferences in two ways, namely as a key or as a justificatory tool of our inferences. My account thus can easily incorporate LC's point on phylogenetic assumptions, but it further clarifies that phylogenetic considerations can play in two distinct justificatory level, one concerning the justification what I call derivational correctness of our inferences from a representation and one concerning the justification for its factual correctness.

Of course, reference to phylogenetic relations in the case of MOs is not arbitrary: they derive from the theory of evolution, one of the cornerstones of modern biology. So, it is reasonable to find phylogenetic assumptions in play. But this does not distinguish MOs from the rest of models: all the relevant "similarities" between a representation and its target are in fact mediated by some assumptions, empirical data, rules, and conventions (see Nguyen 2020 for a generalised argument).

In conclusion, all the differences that LC highlight between models and MOs are, at best, a matter of degree, and they do not undermine my general claim that MOs

[104]For more details on this model, see Weisberg and Reisman (2008), and Nguyen (2020), pp. 1018-1019.

are representations like models are.

3.4 Weber's concerns about representation

Let's now look at Weber's concerns about MOs as representations. While Weber expressed his disagreement with a representational view of MOs during the online presentation of Ankeny and Leonelli's volume (2020),[105] he did not argue at length against it in print. However, in a footnote (2014, p. 758) he writes:

> [M]y argument against [MOs] being theoretical models is that any theoretical model must be associated with a mapping function that specifies what part or aspect of the model is supposed to represent or stand for what (e.g., that the function symbol "F" in a mechanical model stands for the mechanical force). Such a function is not uniquely defined for a [MO] because such organisms may serve a variety of different purposes, only some of which are representational.

We can then individuate two main theses: (1) MOs have many purposes, only some of which are representational, and (2) because of it, MOs cannot be equipped with a uniquely defined function that univocally assigns to each element of the model an element in the target system, a requisite for a system to be a model.

I first deal with (1) and show that it is not problematic for the account presented in this Chapter. Then I move to (2) and show that Weber's requirement of a univocal function is too demanding because his view of the model-target relation is too simplistic: the interpretation of the model varies on the basis of the target and the specific purposes of our study.

3.4.1 The multiple functions of MOs

Weber talks about a "variety of different purposes" for MOs besides representation. What are then the functions of a MO that are non-representational, at least in DEKI's broad sense? Weber (2004, Chapter 6) has put forward his own view of the main function of MOs in biological research, which he calls "preparative experimentation". On this view, MOs are best understood as material and conceptual arenas: the space where new experimental procedures arise and scientists learn how to manipulate biological systems experimentally, achieving a know-how that may become useful for future applications or interventions. Weber's preferred example is the work carried out on *Drosophila* in order to understand how the process of genetic crossover occurs and its implications for genome selection. As he shows, the study of *Drosophila* has brought about the acquisition of new techniques to be employed in new contexts of investigation. He particularly insists on the development of the so-called "chromosomal walking" (*ibid.*, pp. 160-62), a technique to clone DNA sequences about which only their chromosomal location is known.

[105] The integral video can be found at this link.

I agree with Weber that preparative experimentation is not a representational use of MOs, at least in DEKI's sense. One may interpret it as a prescriptive form of representation: a system exemplifies the possibility to perform some techniques and indicates or prescribes how to apply them in other contexts. Yet, in its present form, DEKI cannot deal with prescriptive cases.[106] However, this does not seem to be a problem. In the literature on models, nobody argues that scientific models are only functioning as representations (cf. Frigg and Nguyen 2020, p. xii). Representation is indeed just one of the several functions that models can serve.

So, in order for Weber's argument to be effective against the representation view presented here, either (a) Weber needs to identify a function of MOs that is incompatible with representation in DEKI's sense; or (b) he must argue that MOs' representational function is secondary or negligible to understand MOs' use in biology, even when we interpret representation in DEKI's sense. Neither of these options look promising. As regards (a), it should be clear that there is no logical incompatibility between Weber's preparative experimentation and representation. Indeed, it is common to find examples of models used for preparative experimentation that are then also used as representations.[107] This is just because models are useful tools not only to identify properties to map onto our targets, but also to test or develop theories, to integrate them with data or other theories, and to develop new formal or empirical methods of analysis. And, in the case of many models, MOs included, these functions are carried out with the same material objects. Namely, the MO populations studied in the laboratory serve to represent other organisms, to develop explanations and predictions of certain phenomena, to enhance our understanding (both know-that and know-how). Therefore, the plurality of MOs' functions in biology is not necessarily a problem, nor an exception with respect to the rest of scientific models.

Concerning (b), is representation a negligible use of MOs? This is an empirical question that has to be answered by investigating how MO-based inferences can be reconstructed and justified. In section 3.2, I argued that this requires an account of representation, and that the best account to offer such a reconstruction is DEKI. Until we have an argument to the contrary, showing that representation is not needed after all, or that DEKI is the wrong account to fit the bill, the conclusion stands: understanding MO-based inferences is best done with an account of representation, and hence representation is not negligible.

[106] On normative models and their relation to descriptive ones, see Beck and Jahn (2021) and Roussos (2022).

[107] For example, see Hartmann (1995, p. 9) for a use of models to develop new formal techniques in quantum chromodynamics.

3.4.2 The synergy between representation and preparative experimentation

One final point to take into consideration is whether and possibly how Weber's view affect my proposal about the justification for MO-based inferences. Now, Weber's *pars construens*, the preparative experimentation view, focuses on the exportation of techniques and know-how, which does not necessarily require representation. If I exercise my use of hammer and nails on tables, and then I use the same know-how on chairs, this does not require me to think of tables as a representation of chairs. It is instead the mere fact that I can do similar actions and interventions on two different systems.

However, this is not necessarily problematic for my proposal. The fact that we do not need representation to export experimental techniques is perfectly compatible with my suggestion that the justification for our MO-based inferences still derives from the exemplification of properties that are successfully translated via a key. In other words, my account does not require representation to ground all uses we can make of a MO. It is useful to interpret MOs as representations when inferences are carried out from the study of MOs to other organisms. Especially, the fact that MOs are employed for preparative experimentation does not threaten my claim that MO-based descriptive inferences are justified insofar as MOs exemplify properties that are imputed to the denoted target via the application of the appropriate key.

In this sense, we also see that DEKI is not a completely trivialised account: something like preparative experimentation does *not* count as representation because it does not fit the mould of the account.

However, while preparative experimentation in Weber's terms and representation as conceptualised within the DEKI framework are conceptually distinct, I want to suggest that there is an interesting interaction between these two functions of MOs. Indeed, I contend that, when the know-how transfer requires a great deal of theoretical knowledge about the systems involved, representation is crucial to Weber's preparative experimentation as well. In order to export techniques and experimental methodologies in other contexts, one has to assume that the preparative experimental scenario and the application scenario are similar in the relevant way. This is not meant to be a concession to the similarity view: the point is that, in order for a MO to function as an arena for preparative experimentation and development of new know-how to apply to other organisms, we have to at least hypothesise that the MO in question exemplifies certain properties that make the development and further application of that know-how possible.

This point seems in line with Leonelli's reply to Weber during the presentation of the book *Model Organisms*:

> "[Ankeny and I] don't really see our account as contrasting with [Weber's]. Rather, what we see is that our account is adding to [Weber's] by stressing

> the fact that even in situations where you are really using MOs purely for intervention... there is still an important need to pay attention to which *representational assumptions* are been slipped under the carpet by considering a MO as a *plausible* tool [for such an intervention]". (1:12:02, my emphasises)[108]

Ankeny and Leonelli too, then, contend that we are often allowed to use MOs as means of preparative experimentation *because* of some (perhaps hidden, perhaps wrong) representational assumptions. So, MOs would still function as representations, even in the case of preparative experimentation. I want then to suggest that MOs used as preparative experimentation tools would be better understood as something similar to a *normative*, or practical model, like the ones that we sometimes find in economics (Beck and Jahn 2021). As such, a MO would not work as a surrogative system for reasoning about the designated target, but rather as a way to identify practical guidelines for the experimenter or the policy maker.

Here, I am not suggesting that these normative models are treatable with the DEKI account: some new elements should be added to the account in order to make it a good framework for prescriptive symbols. This is only a suggestion for a future investigation. For the present purposes, I do not even need to go as far as Ankeny and Leonelli and argue that descriptive representation is a *necessary* condition for preparative experimentation. All I need to say is that representation and preparative experimentation are perfectly compatible, and that an analysis of each of them is beneficial to understand the other.

The dependence of preparative experimentation on representational assumptions, of course, will be a matter of degree. When we talk about tables and chairs, one does not need much theoretical knowledge and representational inferences about these systems in order to export intervention techniques. The more our techniques require theoretical knowledge, though, the more justificatory role some theoretical assumptions will play in our know-how transfer.

Of course, it is important to recognise that our representational assumptions will be affected by our progress on the know-how too: the more we improve our experimental techniques, the more refined understanding of our MOs will be. And hopefully, this will also imply a better picture of a MO's representational potential with respect to other life forms. This dynamic interaction between representation and preparative experimentation should not be a surprise, nor taken as something characterising MO research peculiarly: all sciences seem to exhibit this synergy between theoretical assumptions and know-that with technological advancement and know-how – again, this is an overall picture of science in agreement with the generally holistic nature of knowledge and the process of finding reflective equilibrium in it, like suggested by Elgin (1996).

To summarise, my new version of the representation view of MOs and its im-

[108] Cf. fn. 105.

plications on the justification for MO-based inferences is not only compatible with Weber's positive account of the use of MOs as tools for preparative experimentation, but it is also useful to shed light on the relation between these two functions.

3.4.3 Univocal functions and interpretation

Let us now move to Weber's point (2), namely that models represent only insofar as they provide a uniquely defined function that univocally assigns to each element of the model an element in the target system, a requisite failed by MOs because of the many ways in which they may be related to their targets.

I grant Weber's point that, in a very general sense, we need some form of interpretation of each part of the model, and then a function that relates parts of the model to parts of the target. In the account adopted in this thesis, this is done via an interpretation of the object as a model and the subsequent adoption of a key. However, first, a key only applies to the properties exemplified in that specific context. In this sense, the mapping does not need to be complete for all the properties and elements of the model. Second, as I will show now, both the interpretation and the key allow a reasonable level of flexibility, which Weber's univocal function does not.

Weber seems to suggest that the basic terms of a model always must have a precise physical or biological interpretation, while the basic terms of MOs do not. In the case of the Lotka-Volterra model, for example, each of the two differential equations describing the model system are endowed with a clear physical-biological interpretation – for instance, a variable is univocally the number of the prey, another indicates the number of predators, and so forth.

There are a number of considerations that show that the alleged precision that Weber sees in the interpretation of models is too simplistic. For, in fact, mathematical or formalised models sometimes employ terms endowed with a precise definition that nevertheless do not seem to refer to anything real in the world. For example, the variable denoting the prey actually describes a prey population that grows limitlessly if it weren't for the presence of the predator, and this physical-biological interpretation of V does not map precisely to anything in the biological kingdom. The Lotka-Volterra model is no exception in this respect: the very term F that Weber uses to make his point in the above quotation stands for a theoretical entity: nobody has ever observed Newtonian forces, only their effects. As Cartwright (1999, Chapter 3) pointed out, it is arduous to even specify what counts as a force in the first place.

All this means that the physical or biological interpretation of the theoretical terms employed in a model is far from being uniquely defined: it depends on the target system and the context of application. That's why we need a key in the first place. This becomes even more evident when we look at cases in which the "same" model is applied to represent different targets. This phenomenon of model transfer, or "model

migration" (Bradley and Thébault 2017), is ubiquitous in science.[109] Hydrodynamic models were used to represent electromagnetic phenomena, and mechanistic models of particles are used to represent stock markets. Even the chromosomal walk that Weber analyses in his book is a version of the so-called random walk, which was used to model the Brownian motion of particles.

Once we realise that Weber's demand of a unique interpretation of the terms and elements of our models is untenable, it is also clear that MOs are actually like other scientific models in this respect. So, for example, in *C. elegans*, the mechanisms of cellular programmed death (Ankeny and Leonelli 2020, p. 8) can easily be taken to be a simplified version of more complex mechanisms of cell self-destruction in other organisms. Here, the elements of the self-destruction mechanism in the worm's cell stands for the analogous mechanisms in the cell of the designated target organism. The same can be said about the genetic cross-over in the *Drosophila*, where its specific chromosomes, the fragments of DNA, and the process of chromosomal walking respectively stand for chromosomes, DNA, and chromosomal mechanisms in a large array of other biological systems.

Just like in the Lotka-Volterra model, some elements of the mechanisms that the MO exemplifies will not map onto elements of the same mechanisms of the target species. But this is just another way to say that we need a key to export our results to other biological systems, adapting the interpretation of the elements of the model to each specific context. In the case of our architectural proteins, this becomes manifest: proteins are not mapped onto other proteins one by one, but they are grouped according to their function in the mechanism.

In conclusion, Weber's preparative experimentation is compatible with MOs being used as representations, and his requirement for a uniquely defined function is too strict, because the use of a system as a representation of another is highly context- and target-dependent.

3.5 Summary of the chapter

By taking full advantage of the resources offered by the DEKI account of representation, I have shown how exemplification and the key play crucial roles in the inferences drawn from MOs about other organisms. I have also provided an account of the justification for MO-based inferences: an inference from a MO to a target system is justified insofar as the MO exemplifies a set of properties that are mapped onto the target via an adequate key.

I have then addressed two views of MOs that challenge the representation view. I have shown that Levy and Currie's (2015) arguments do not undermine my view of MOs as representations, and I have argued that Weber's (2004) preparative exper-

[109] On the phenomenon of model transfer in science, see Herfeld (Herfeld 2024 and references therein.

imentation view is compatible with my view. Specifically, my analysis shows that representation does not imply unmediated analogy, nor uniquely defined functions for each single part of the model. In contrast, representation is always mediated by, and embedded in, theoretical assumptions and empirical knowledge. At the same time, representation always remains local, insofar as it is context- and target-dependent. Therefore, the justification for our inferences from a representation always consists of an interplay between the justification internal to the specific representational framework (the key), on the one hand, and on the other the justification provided by knowledge (the repertoire) largely extrinsic to the single representation.

Chapter 4

Why we love pictures (for the wrong reasons)

4.1 The controversial nature of pictures

Pictures are ubiquitous in science. Astronomers study pictures shot by telescopes and probes to understand how stars form and dissolve, medics use X-ray and MRI scans to detect diseases and provide their diagnoses, and epidemiologists create heatmaps to explain and predict virus spreading patterns. In general, scientists use a vast variety of pictures, from scans and photographs to diagrams, graphs, and maps, in order to gain information about phenomena. A number of philosophical questions arise from this use of pictures in the sciences. How do pictures function in the context of scientific enquiries? What is the process through which we learn from pictures? And how do we justify our inferences from pictures to the world?

In the literature on depiction, both in aesthetics and the theory of images, there is an important philosophical tradition which focuses on the concept of similarity in order to explain how pictures represent the real world. The application of the similarity view to *scientific* pictures, however, has remained importantly unexplored. An important exception is Meynell (2013), who explicitly wants to clarify the role of similarity in the use of pictures and visual representations in science. Meynell's is for now the best attempt to make sense of the use of pictures in science by appealing to the notion of similarity, so I take her view as a point of reference for my critical analysis of the similarity account in this context.

Her account is built through a combination of, on the one hand, an attack to Perini's (2005) attempt to apply Goodman's (1976) conventionalism to scientific pictures, and, on the other, a constructive proposal inspired by the work of Willats (1997) in psychology. While Meynell accepts that Perini's Goodmanian approach can work well with linguistic or quasi-linguistic visual representations, like schematic diagrams, it remains insufficient for "dense" pictures, like photographs, scans, microscopic and astronomical pictures, and so on. Then, she argues that it is better to understand the

use of scientific pictures by employing a similarity view, combined with our knowledge of psychology, theory of perception, and a combination of geometry and optics.

I aim to show that, when it comes to understanding the epistemic use of pictures in science, a focus on similarity is the wrong way to go, even when it comes to dense pictures. I argue that, even if Meynell's account seems to work well with simple, non-scientific uses of pictures, it remains wanting when we move to more interesting cases. For my argument, I will mostly focus to the recent picture of the M87* black hole at the centre of the Messier galaxy produced between 2017 and 2018 in the context of the project Event Horizon Telescope. I argue that, on closer inspection, similarity is not really the essential concept on which to base the epistemic use of pictures as representation. Instead, I will highlight the fundamental role of interpretation and exemplification, which provide us with a better understanding of the semantic and epistemic features of scientific pictures.

There is already a rich literature on the epistemology of the picture of M87* and in general of black holes. Curiously enough, however, philosophers of science have not studied this picture *as a picture*. That is, they have not focused on the features of this picture as a representation, as an object allowing surrogative reasoning about its target system. This has important repercussions for the way in which philosophers have conducted their epistemological reflections: they primarily focused on the role of the picture as a piece of evidence and there seems to be little interest in analysing the use of the picture as an epistemic representation. I want to suggest that, while the epistemological analyses conducted so far are crucial, they remain incomplete. Indeed, a study of the picture as evidence presupposes an analysis of how the picture of a black hole is supposed to be "read" as a representation. In this sense, my analysis will also be a useful contribution to the general epistemological enquiries about black hole pictures.[110]

This Chapter aims at filling this lacuna and provide an analysis of the epistemic use of this image: an interpretation of the M87* picture, and indeed a general framework for scientific pictures in general. The crucial ingredients of the proposed approach are: an interpretation of the image, the fact that that image exemplify certain properties, and the use of a "key" to translate idealised or distorted properties into the ones that we want to attribute to the target. I then naturally identify these elements with the basic ingredients of the DEKI account of scientific representation, recently elaborated by Frigg and Nguyen (2020).

While the DEKI framework explains how we "read" pictures as representations of other systems, the account has two shortcomings. First, it is skeletal by design: it needs to be completed with the specifics of each case study. My analysis of the

[110] I did not mention here the metaphysical investigations that arise from the very peculiar act of creating a visual representation of something that is in principle invisible. These are fascinating investigations that may bear on the discussion of scientific realism. However, I put these issues aside because they remain tangential to my present argument.

Figure 4.1: Jacques-Louis David, *The Death of Marat* (1793). Oil on canvas. Royal Museums of Fine Arts of Belgium, Brussels. Wikipedia Commons. https://en.wikipedia.org/wiki/File:Death_of_Marat_by_David.jpg

black hole picture provides the relevant details on how to apply DEKI to this specific case study. Second, the account remains silent on the justification for our inferences from the picture to the target system. While the interpretation of a representation and the conversion of knowledge from it to the target are arguably conceptually distinct processes, I suggest that in the case of pictures the root of their justification is the same, namely the causal history of production connecting a given picture to its designated target system. This affords us a way to go beyond what DEKI provides and assess the accuracy of a picture, and it shows how we can do so without an appeal to similarity. Also, our discussion of justification highlights a crucial difference between pictures and other types of scientific representations like models. For, I will argue, model inferences do not seem to exhibit the same kind of dependence on the representation's history of production as pictures do.

4.2 The mirage of similarity

Historically, pictures have been taken to represent their targets by dint of similarity: by being similar, in a relevant sense, to the represented portion of reality. This appears to be strikingly apparent with photographs and realistic paintings. Take for example David's famous painting *The Death of Marat* (Figure 4.1). Different areas of the painting are similar to the intended target, namely the body of Marat just after being murdered by Charlotte Corday. Different coloured regions represent

different objects, in virtue of having the same colour of those objects. For example, different white areas represent Marat's exsanguinated body, a towel wrapped around Marat's head, and a sheet on the edge of the bathtub where the French politician and intellectual used to spend his days (due to a skin disease he got after spending months in the sewers of Paris, hiding from the monarchic regime). The areas painted in red represent the blood spilling from Marat's wounded chest. We can observe other objects, like the knife on the floor, close to Marat's hand still holding a quill. In the foreground we see a wooden trunk used as a desk with another quill and an ink pot. By recreating a shadow-effect, the depicted objects acquire depth and volume, making them appear three-dimensional and in specific topological relations with each other.

This explanation of how pictures represent also seems to work for photographs. A photograph represents their subject by presenting colours and shapes that are similar to those instantiated by the target system. The question now is whether this intuitive notion of similarity can be made more precise so that it can serve as an effective analysis of the use of pictures in science.

First, one may worry that similarity is symmetric and reflexive, while representation is not (Goodman 1976, Suárez 2003). More generally, representation requires (intentional) directionality from a representing object to a designated target system (Frigg and Nguyen (2020, p. 11). However, the literature on scientific representation has already proposed ways to overcome these semantic problems. Weisberg (2013, pp. 135-155), following Tversky (1977, 1978), has proposed a notion of similarity as a weighted feature matching function that is inherently asymmetric (but still reflexive) and that requires a specification of which similarities are the relevant ones. Alternatively, Giere (2004, 2010) has put forward an agent-based account of representation where it is the agent that chooses to use a system as a representation of another, thus giving a representation the required directionality and the specification of the relevant properties.

I am taking for granted these replies to the traditional attacks to the similarity view of representation. I am assuming here, for the sake of the argument, that the similarity account works well with realistic paintings and every-day, colour photographs. What I want to critically assess, instead, is whether this view succeeds when it comes to scientific pictures.

A further, traditional problem of the similarity view, as Goodman famously said, is that everything can be similar to anything else. For the concept of similarity just implies that two systems are similar with respect to a property P *iff* they both instantiate that property P.[111] If we are liberal on what count as a property of a

[111]They may be similar also in another sense, namely they possess two distinct properties that are in turn similar to each other. As the latter case is just one where the two systems share some second-order property (e.g., they respectively instantiate two different shades of red, and so they share the higher order, more general property of being red), this distinction is not particularly important here. The core of the concept of similarity is still reducible to the fact that the *relata*

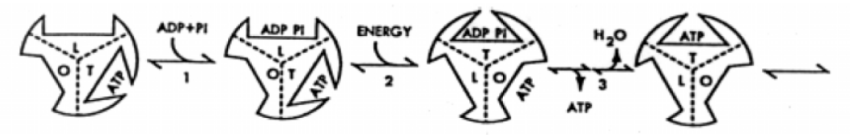

Figure 4.Mechanism diagram (from Jan Pieter Abrahams, Andrew G. W. Leslie, René Lutter, and John E. Walker (1994), "Structure at 2.8 Å Resolution of F_1-ATPase from Bovine Heart Mitochondria", *Nature* 370 (6491)). Reprinted with permission of *Nature* (http://www.nature.com).

Figure 4.2: Mechanism of the F_1ATPase. In Perini (2005, p. 276).

system, it is easy to find many similarities between objects. Which are the relevant ones?

In the next subsection(4.2.1), I look at a specific attempt to develop a similarity account of scientific pictures, proposed by Meynell (2013) and show that it gives an answer to the relevance problem: the relevant properties in which pictures are similar to their targets are their spatial and more generally visual properties. As I will show, Meynell also offers a way to relate in a systematic way the visual properties of the picture with the visual properties of the target. Let us see how.

4.2.1 Meynell's account

Meynell offers a similarity account of scientific pictures. Among many types of visual representations commonly used in science, she holds that her account appears particularly effective with what she calls "*pictorial* content", in contrast with what she calls "visual languages" (2013, p. 329; original italics). Roughly, the pictures of the latter kind are analysable as linguistic or quasi-linguistic representations. Paradigmatic examples are Venn diagrams, Peirce diagrams, and Perini's own example of diagrams used to illustrate biological mechanisms (e.g. Figure 4.2). The pictures of the former kind are those that are too "dense" to be parsed into atomic elements as one can do with language or diagrams. Standard examples of "pictorial" pictures are paintings, photographs, and scans. Because Meynell focuses on dense, non-linguistic pictures, I will do the same here and evaluate her proposal only with respect to dense pictures.

The definition of density that Meynell endorses comes from Goodman (1976, pp. 130-141), who gives a rigorous definition of it (what he calls "density throughout") in the context of representation and symbols more generally. For our purposes, it is important to clarify that density is here understood as just an extreme case of what Goodman calls lack of "syntactic articulation" or "syntactic differentiation". A symbol system is not syntactic articulated when it is (theoretically or practically) impossible to discriminate one symbol from another one. Imagine, for example, a simple symbol system consisting only of two symbols. A symbol is a vertical line not

share a property, first or second order is not really a problem.

longer than one inch, the other symbol is a vertical line as long as an inch or longer. With such a symbol system, we will often be unable to distinguish one symbol from the other, no matter how precise our measurement of the lines' length will be. For there will be cases where we cannot tell if a line is the former or the latter symbol. A dense system is a generalised non-articulated system, where is always possible to find a symbol between other two symbols.

The details here are beyond the scope of this Chapter, where I take the distinction between dense and non-dense symbol systems for granted. The take-home message is that the concept of syntactic non-articulation should capture the very intuitive idea that, with some types of representations, like in photographs and paintings, we are unable to clearly discriminate what a certain mark or inscription is supposed to mean. If the colour of a painting, for example, changes continuously from, say, black to grey to white, and we assume that different colours mean different things, we will not be able to clearly indicate where we move from a certain symbol to the next. This of course applies to scientific cases, like astronomical pictures, MRI scans, many instances of maps and heatmaps, and non-discretised charts. In what follows, I take my examples of scientific pictures to be paradigmatically dense in Goodman' sense, and thus the target of Meynell's analysis as well.

Meynell holds that, while it is intuitive to analyse diagrams along the lines of linguistic expressions and parse them into atomic constituents, we cannot do the same with dense pictures like Figure 4.1. The elements of the image are dense, so it is not possible to clearly distinguish each of them and specify their meaning in a conventional, stipulated way. It is better, Meynell holds, to abandon the idea to analyse pictures as if they were like languages, namely conventional symbol systems. It is better to look at what psychology tell us (in particular Meynell is inspired by the work of Willats 1997). She suggests thinking of the picture as constituted by "basic visual properties" (2013, p. 336), or "picture primitives" (*ibid.*, p. 338), which are still not endowed with an interpretation. While Meynell remains silent on exactly what counts as basic or primitive in this sense, it is widely assumed in psychology and semiotics that our perceptual systems can recognise and distinguish between very general visual elements: points, lines, and regions; relative spatial dimensions and topological relations; and different shades of colours.

What could these primitives correspond to in David's artwork? In the painting, we can identify colours and shades, as well as lines and regions, placed in a specific way relative to each other. At this point, it is not assumed that there is an interpretation of these picture primitives. Such an interpretation, Meynell submits, must be added in a next step:

> "These 'picture primitives' in turn represent the most elementary units of shape information in the scene – 'scene primitives', which can be 3D (lumps, sticks and slabs), 2D (surfaces), 1D (edges) or 0D (corners). These scene primitives then represent [real] objects" (*ibid.*).

It is important to note that, as Meynell explains, "'[r]epresent' is not used here in any familiar philosophical sense [...] the term 'represent' depends on the *psychology of perception* and should not be understood as an abstract or stipulated property or relation" (*ibid.*, p. 336, fn. 12, my italics). Hence, the core of Meynell's account is that the various picture primitives are associated with aspects of the world through our perceptual mechanisms, and not on our explicit decisions (or conventions that result from such decisions). This is meant to distinguish her view from that of her opponents, namely Perini (2005) and Goodman (1976), who defend a conventionalist view of symbols and representation, pictures included. However, it is important to clarify that, for Goodman, "convention" does not imply conscious choice: our conventions may come automatically from habit, as well as be subconscious or inherited via evolutionary selection or cultural learning. His point is just that the interpretation of symbols may be otherwise, depending on the symbol system of reference. Any symbol *qua* symbol is always part of a symbol system, that is, a conceptual framework, and it is the system as a whole that defines the meaning of the symbol. Here, the term 'conceptual' should not be confused with 'linguistic': for Goodman, even perceptual properties like colours are part of a conceptual, non-linguistic framework (the colour red is identified only against a conceptual framework that includes other non-red colours).[112]

As transpires from the quote, Meynell argues that the relation between picture and target comes in two steps. First, the picture primitives are related to scene primitives. Picture primitives are the traditional primitive elements of semiotics: points, lines, areas of colours, regions. One can think of them as the elements of a picture deprived of any referential connotation. A point is just a point, a region of colour is just a region of colour. A scene primitive arises when a picture primitive is interpreted already as a symbol, even though in a very minimal way: a line becomes an edge, a region becomes a surface, a point a vertex. Second, the scene obtained as a result of the composition of the scene primitives is in turn related to a real target system. The edges and surfaces of the scene are then interpreted as the edges and surfaces of, say, a real cube in the world. Let us have a closer look at these steps one by one.

First, let us consider the relation between picture primitive (the basic ingredients of picture) and so-called scene primitives (edges, surfaces, volumes). Scene primitives are basically an interpretation of the corresponding marks on the page according to our perception. A line becomes the edge of an object, a region of colour becomes a surface, differences in shadings become indication of depth and volume, and so on. All put together, the ensemble of scene primitives constitutes the scene.

We all have an intuitive understanding of what this means. When we look at realistic paintings and photographs, we immediately see something *in* the picture: not

[112] See more on this in Giovannelli (2017, section 4.1).

just a bunch of colours and lines but also objects, silhouettes, people, buildings, and so on. Meynell explicates this point by noting that her account builds on Willats' theory of perceptual representation, and that Willats explicitly acknowledges Wollheim's (1987) lesson about "the 'double reality' of pictures—our capacity to see a house or ship *in* what is [...] recognisably a set of marks on a page" (Meynell 2013, p. 336). More concretely, the spatial and visual properties of the primitives, combined with our psychology, make us "see-in" the scene primitives. What was just a line is the edge of an object, a coloured area becomes a surface, different shades of the same colour become a curved with depth, and so on. Therefore, I take Meynell to assume Wollheim's see-in as a way to bridge picture primitives and scene primitives.

However, how do we move from pictures primitives to the elements constituting the scene, exactly? Importantly, for Meynell it is important to show that this step is not stipulated or conventional in Goodman's sense. It is perception that grant us to move from picture primitives to the scene. But Wollheim's see-in seem to dangerously bring us towards some forms of subjectivism: the scene seen is just what the subject associates with the picture primitives. Therefore, Meynell needs to offer a notion of the relation between the picture and the scene that satisfy three desiderata: (i) the relation is still grounded on the notion of similarity, (ii) it depends on perception and not on convention or stipulation, and (iii) it provides some objective and systematic connection between picture and scene. Meynell gives us the answer: the systematic relations connecting picture primitives and scene primitives are geometrical projections.

These projections can be of different sorts. The most common are orthogonal, oblique, and perspectival (Meynell 2013, p. 336). Irrespective of the type, a projection geometrically translates 3D spatial properties of the scene into 2D spatial properties of the picture primitives in the picture. Concerning our first desideratum, we can see that, via projection, similarity is once again preserved: simply put, projections grant the picture to retain some spatial and visual properties from picture primitives to scene primitives, and in that, the resulting scene is similar to the non-interpreted elements of the picture.[113] For example, a perspectival picture of an object will present the occlusion shape of that object, given a certain point of view. Therefore, similarity is guaranteed, in the sense that the picture and the represented object share the same objective property, namely the occlusion shape they offer from a certain point of observation. This idea can be generalised to other forms of projections, where what is preserved is not necessarily the occlusion shape but other spatial relations. For example, a Mercator planisphere will preserve the angles between meridians and

[113]In the context of artistic depiction, this use of projection to preserve similarity is also employed by Hyman (2006, 2012). Interestingly, Hyman too holds that geometrical projections (and thus similarity) do not have to do with the relation between a picture and a target. Rather, similarity is involved in the relation between the picture and what he calls the *sense* of the picture, namely what the picture presents in terms of a content. I take it that Hyman's concept of a picture sense is basically equivalent to Meynell's scene.

parallels of Earth, but not the distances between points on Earth's surface.

Let us consider the second requirement, that is, that the relation connecting the picture to the target is not stipulated or conventional. This requirement, however, seems to fail, because all forms of geometrical projections are clearly conventional in Goodman's sense, including perspective. As Goodman (1976, pp. 15-19) convincingly argues, there is nothing objective or purely geometrical in perspectival representation. For example, in perspectival paintings and photographs, only horizontal parallels converge while vertical lines do not, but this does not follow from any law of optics' geometry: it is an a posteriori correction. Indeed, photographic lenses are adapted in order to correct their perspectival representation and make the parallel vertical lines parallel again.[114] Therefore, any choice of projection already involves some form of arbitrary convention in Goodman's sense.

This of course allows the chosen projection system to constrain our interpretation of the picture in relevant ways. Nevertheless, my main point remains that projections are conventional systems of translation, which retain or systematically convert certain properties and exclude others. It is just that there are types of translation that are more useful than others for certain purposes, and perspectival projections have become more entrenched with our depictive practices for cultural and historical reasons, not because they are more objective than other forms of projections.

This, however, may not be an insurmountable problem for Meynell. What she seems to have in mind is something among these lines. First, we usually do not need to even think about the projection system because we intuitively apply it. We are able to do it on the basis of how our perception works.[115] We always see, for example, occlusion shapes of real objects, and from that infer the three-dimensional features of those objects. Second, Meynell could also add that, in her framework, even if there is some arbitrary choice in the projection system applied, this only concerns the relation between the picture and the scene, and not between the picture and the target. Once the projection system involved has been set, the scene and the target are objectively similar to each other, because the projection system is already assumed as given. I will come back to this later when I talk about the relation between the scene and the target.

As regards the third requirement, this seems satisfied: geometrical projections give us a way to rigorously, systematically connect picture primitives and scene. Meynell seems also to suggest that it is the type of projection itself that determines which properties are relevantly similar, because projections objectively constrain the interpretation we can give of the picture. This should be exemplified by Figure 4.3.

[114] See also Feyerabend (2001, Chapter 4) for an insightful reflection on the rise of perspectival drawing.

[115] However, it is important to note that this automatism seems plausible only when one deals with perspectival projections. It seems instead quite a stretch to assume that human are able to read orthogonal projections and automatically understand the scene represented.

First, we identify the two objects as cubes because the picture primitives are in fact geometrical projections of cubes. The one on the left is a cube drawn in cavalier oblique projection – the front of the object is represented with its actual shape and the other lines preserve their actual length. The cube on the right is instead drawn in perspective. First, we will tend to perceive the right one as closer to us than the other, because we implicitly assume that they lie on the same plane. Second, while we can see the left one as either concave or convex due to the projection system employed, this cannot be done with the right cube constructed with the perspective projection, unless we assume that the "two remaining square sides have been cut down into smaller irregular quadrilaterals" (*ibid.*, p. 339).

However, while projections can be used to explain how spatial properties translate from pictures to scenes, they seem to not work well with many other properties of pictures. In pictures, we do not see only points, lines, and shapes, but also colours, shades, and shadows. Meynell does not talk about a systematic relation between colours in her paper, but it is not difficult to see how she would go. Colours of the picture can be systematically related to the colours of the scene by assuming a certain light intensity, a certain position of the light source with respect to both the picture subject and the perspective point, and of course the chemical and physical properties of the objects represented, and the consequent wavelength with which light would reach our sensorial apparatus.[116] This is very important, because it gives us a way to understand the pictures as a scene. For example, a difference of colour in the picture does not necessarily imply a difference of colour in the scene: a darker tone of a colour may indicate that a part of an object is in the shadow with respect to a light source, or that an object has volume, and so on.

It is important to notice that, for Meynell, this translation from the picture to the scene is non-linguistic. Even though geometrical projections can be expressed mathematically, observers are not required to translate elements of the picture in symbols in order to see the scene in the picture. The translation is also usually objective, in the sense that it is not a purely conventional procedure: it depends on physical properties of light and objects and the rules of optics. In this sense, Meynell[117] allegedly free representation from the burden of conventionalism (Goodman) or subjectivism (Wollheim): the relation of similarity is objective, not depending on arbitrary choices or purely subjective idiosyncrasies.

For now, we just talked about Meynell first step, namely from picture primitives to scenes. Let us now move to the relation between the scene and the real object depicted, the target system. At this point, it is actually not clear how Meynell would relate the scene with the actual target system. I take it that there are two alternatives. Either the scene represented *is* the target system itself, or the scene

[116] This way to define a translation of colours is inspired by Hyman (2006), whose views on artistic depiction are relevantly close to Meynell's ones on scientific visual representation.

[117] And Hyman, too.

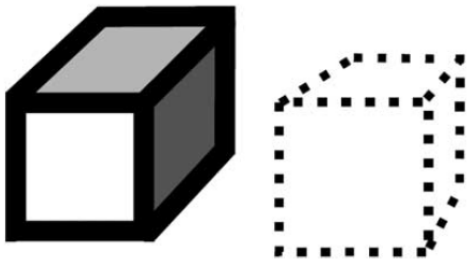

Figure 4.3: Two cubes drawn with oblique projection (left) and perspective (right). Meynell (2013, p. 338).

represents a target system insofar as they share the same spatial (and more generally visual) properties. And this is again intuitive: once we see a tree in a painting, if the tree we see in the scene is similar to a tree that exists in terms of perceptual properties, then the picture is representing that existing tree.[118]

For which alternative should one opt? For one could just say that picture primitives directly relate to objects by being projections of them. Meynell does not consider this issue directly. However, she mentions that the scenes represented can also be imaginary or fictional (2013, p. 339). Also, we can easily imagine the case of a scene being not entirely identical to the target system. Perhaps, the shapes, colours and in general the visual properties may not be exactly the same. In David's painting, the position of Marat's body was perhaps slightly different from the one presented in David's painting scene. Therefore, I take her distinction between scene and target to be useful, both because it allows pictures to represent scenes that do not have a corresponding target, and because it provides leeway in the case the scene itself is not exactly identical to the target system.

Summing up the basic elements of Meynell's account, we have a similarity relation connecting picture primitives and the scene, expressed in terms of projections, and then a similarity relation relating scene and target, expressed in terms of spatial and visual properties. Let us look once again at *The Death of Marat*. The picture primitives are to be understood as a perspectival projection of a 3D scene. Namely, what we see-in the picture is an occlusion shape of a body in a bathtub in front of us, with relative colours. The occlusion shape of the scene objects is also supposed to be the same shape that someone looking at Marat from the left side of the bathtub, a few meters away, would see.

Meynell also attempts to apply her account to a scientific case, namely an electron micrograph image (Figure 4.4). The micrograph is meant to be a representation of F1-ATPase complexes attached to mitochondrial membranes. In this case, Meynell tells

[118]This similarity can be between the scene and the target, or between our perceptions of them – like for example in Peacocke (1987). This distinction may result in some form of ambiguity, but I will come back to this later.

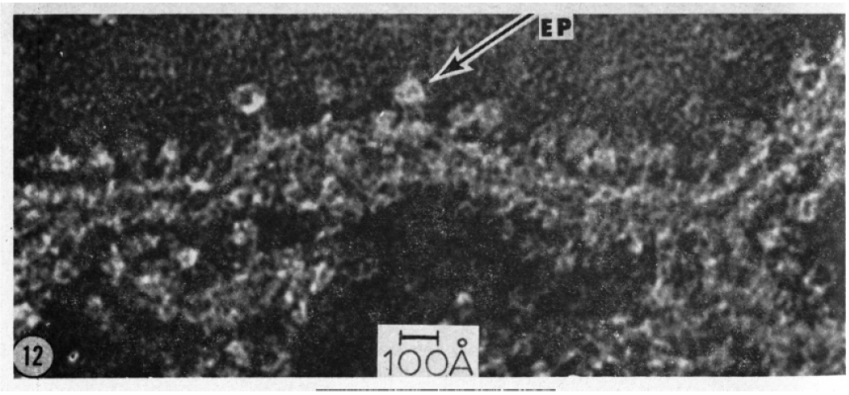

Figure 2. Electron micrograph (from Humberto Fernandez-Moran (1962), "Cell-Membrane Ultrastructure: Low-Temperature Electron Microscopy and X-Ray Diffraction Studies of Lipoprotein Components in Lamellar Systems", *Circulation* 26: 1039–1065).

Figure 4.4: Example of electron micrograph. In Meynell (2013, p. 329).

us, we have an orthogonal projection in place, where lines and surfaces perpendicular to the vision plane becomes points and lines respectively. This projection systems "determines which features are significant and what information can be gleaned from them, that is, true relative size, true shape and relative position in the two dimensions parallel to the picture plane" (*ibid.*, p. 342). This analysis is also supposed to allow one to assess whether the projection system "captures the spatial features of the cell membrane ultra-structure that are of interest" (*ibid.*).

In summary, according to Meynell, pictures are geometrical projections of scenes that resemble their targets spatially and visually. Also, the similarities between the picture and its target, mediated by the geometrical translations, constitute the ground for (a) our understanding of the meaning of the picture as a scene, (b) what we learn from the scene about its target, and (c) why we are justified in doing so. If we understand pictures as geometrical projections of scenes (and corresponding translating procedure for what concern colours), we interpret their meaning correctly. Also, what counts as relevant or not in the scene is constrained objectively by which geometrical projections are employed. Finally, we are justified in our inferences from the picture about the target system (via the scene) because projections preserve some of the objective features of the scene, and the scene is similar to the target with respect to those features.

4.2.2 The picture of a black hole

Let us assume, at least for the sake of argument, that Meynell's account succeeds in explaining how photographs, realistic paintings, and simple geometrical figures represent (although I note *en passant* that Goodman and others who are sceptical of similarity, me included, would still disagree). I want to offer a new argument

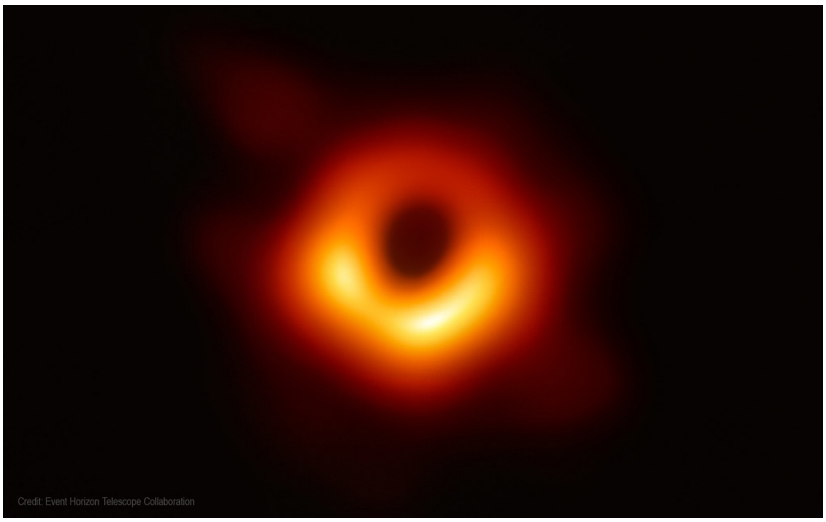

Figure 4.5

against the similarity view, in addition to more general considerations that have been already covered in the literature. The point I want to make here is that even if the similarity view as cashed out by Meynell were successful for more mundane cases, it is yet inadequate when applied to more complex examples of visual representations used in science. I will argue that any attempt to make it work will lead to an abandonment of any non-trivial notion of similarity and, in turn, to the abandonment of related intuitions about realism and the absence of conventionality in picture-like representations.

Let us try to apply Meynell's account to Figure 4.5.[119] There are basic visual elements in the picture that we can automatically identify with Meynell's picture primitives, namely some coloured regions. These regions are not clearly demarcated: they are expressed by a continuum-like shading that ranges from black to white, passing through red and yellow. Therefore, they seem to be a perfect example of dense visual representation in Goodman's sense, and as non-linguistic "pictorial content" in Meynell's terms.[120]

To what scene primitives can we associate the coloured regions of the pictures? Meynell's account tells us that the picture primitives are the result of a geometrical projection of a scene: we should see the scene *in* the picture, and what we see is constrained by the type of projection involved. Let us assume that we know that the projection involved is a perspectival projection, as it actually is. Then, the scene

[119]The absence of a caption is deliberate; it will turn out to be useful for my argument. Relevant copyright information for the picture can be found in footnote 121.

[120]For the sake of the argument, I am for now ignoring the fact that the picture is actually composed by a finite number of discrete units, namely the pixels, and that each pixels expresses a colour which, in turn, is the result of the combination of three different shades (green, blue and red) of three sub-pixels units, all the shades still being discrete types of shades. My argument against Meynell's view is independent of this assumption.

that we see is a sort of red-yellow doughnut against a black background.

This is all we can get from Meynell's account. Particularly, there is no guide about the target. Is there one in the first place? The picture may even fail to be a representation at all. For it may be an image created randomly by a computer, or the digitisation of an abstract painting. In both cases, it would not represent anything.

Let us assume, for the sake of the similarity view, that the reader knows what the target of Figure 4.5 is, by associating a caption to the picture that specifies that it is a picture of M87*, a supermassive black hole at the centre of the galaxy Messier 87, and its immediate surrounding.[121] Even assuming that, the difficulties of Meynell's account just started. For example, one may start looking at the picture of the black hole and say: the black hole is surrounded by a certain ring-wise shape and it is yellow on the bottom of the ring and red on the top, the borders of the ring are blurred, all around is black and thus there is nothing around the black hole, and so on.

However, all this information would be strikingly incorrect. Let us start from the interpretation of the colours. The colours in the picture are not to be interpreted as real colours, but as levels of intensity of radiation in an electromagnetic field, from white that expresses the higher intensity to red that expresses the lower (I will come back to this in more detail below). However, nothing in Meynell's account gives us a way to understand this translation of colours into levels of radiation intensity. But without such a translation, there is no similarity whatsoever in a pattern of coloured pixels and an electromagnetic field. Besides being both continuous and not discrete phenomena, colour shades and radiation intensity levels are not similar at all, they basically have nothing in common.

With the case of the black hole, we see that neither of the two interpretations of similarity obtain: the black hole does not have the colours of the picture, and, luckily for us, the picture does not exhibit the levels of radiation intensity of the black hole. But once we take the translation of colours into levels of radiation intensity into consideration, the initial appeal of similarity seems to vanish, because all the work is carried out by that interpretation, not by similarity. For there is no sharing of properties anymore: colours and levels of radiation intensity are just different sorts of properties.[122]

Besides the problem of interpreting visual properties as non-visual properties,

[121] Figure 4.5 relevant copyright information: Jason Major, "M87 Supermassive Black Hole", Flickr, uploaded on April 10, 2019. https://www.flickr.com/photos/lightsinthedark/47579266551/in/photostream/.

[122] Meynell could here take the very radical move to bite the bullet and argue that the picture of the black hole is not a picture in her sense. She could argue that this object does not represent pictorially, but rather in a much more indirect way. However, her account was proposed to deal exactly with this kind of scientific pictures, like electron micrograph and other sorts of "dense", pictorial representations. Moreover, I will show below that these issues extend naturally to other paradigm examples of scientific pictures. Biting the bullet and excluding the picture of the black hole from the relevant class of pictures, then, would just imply that Meynell's account fails.

other problems remain open. For even if we specify what type of projection is at place in a visual representation, this is insufficient to infer the actual spatial properties of the target. Let us look again at the picture of the black hole and remember that we are assuming a perspectival projection. For all we know, we may think that the black hole is flat, and we were just lucky we got it from the top (or the bottom). Otherwise, instead of a 3D doughnut, we would have just seen a horizontal, yellow-red band. In the end, perspective projections show us the occlusion shape of an object, and if the object is a flat doughnut, then the resulting perspectival shape will change accordingly to our position as more and more elliptical. In fact, this is simply wrong: we know from theoretical physics that the *shadow* of the black hole – the dark region at the centre – would appear from any point of observation, because of the gravitational symmetry of the black hole. But this is problematic for Meynell's account because it shows that while a type of projection somewhat constrains our interpretation of a picture, it still leaves open many possible readings of it.

More generally, the spatial properties of a designated target system are underdetermined under the properties of the picture even when the projection system employed is given. This means that the same picture with the same projection system can be read in different ways. This applies to any instance of geometrical projection: the reading of a photograph, for example, will depend on very trivial factors like how far away the target is from the camera's lens. This underdetermination is of course amplified in the case of the picture of a black hole, because of the highly complex geometry of the black hole region. All in all, not even the spatial properties are to be read in terms of similarity: what is similar and what is not, and to what degree, is always the result of an interpretation.

I will come back to the correct reading of the picture of M87* in the next section. Still, independently from what the right way to read the picture is, my point was simply to show that Meynell's account is wanting because, even when the type of projection involved is specified, it does not take into consideration how there is still a lot of choices to be made in how to read the picture.

So, first, Meynell's account seems to ignore the fact that visual properties in the pictures do not necessarily map onto visual properties of the target. Second, her account cannot deal with the issues related to the interpretation of what is projected, even once the type of projection has been specified.

The proponents of the similarity view could react to these problems by saying that some similarities are nevertheless preserved between the picture and what is represented. For example, the yellow-red region in the picture is still supposed to be in some way similar to the ring shape of the electromagnetic field surround the black hole. At this point, though, the problem is simply that the account turns out to be trivialised: given that projections do not suffice to specify how exactly the picture can be translated into a specific scene, and given that visual properties can be translated in non-visual ones, the concept of similarity becomes basically inert.

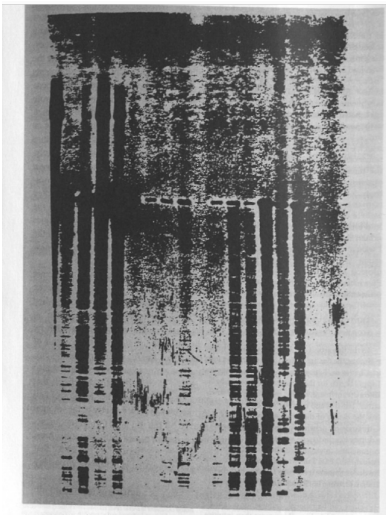

Figure 4.6: Example of autoradiograph. In Elkins (1999, p. 34).

For we don't know how to translate visual properties in non-visual ones anymore, because we cannot appeal to perceptual similarity, and we do not know what are the relevant properties on which we have to focus on.

The picture of a black hole is an extreme example of the untenability of the similarity view, at least in its insistence on linking visual representation to a combination of similarity and perception. It is important to stress though, that the picture of M87* is not at all a cherry-picked exception as concerns the importance of interpretation for the semantics of scientific representations. Indeed, the vast majority of scientific pictures require some form of sophisticated interpretation, informed by our theoretical understanding of how the picture was obtained and was it is supposed to do in our scientific enquiries. Look, for example, to Figure 4.6 (from Elkins 1999, p. 34). Without an explanatory legend, it is impossible to understand what it represents. One could easily take it to be, if anything, some sort of Japanese-like, minimalist representation of bamboo canes on the border of a water pond, where the canes silhouettes are mirrored on the rippling water. Instead, this is an autoradiograph scan. It reproduces radioactively labelled DNA or RNA fragments, separated in an electrophoresis gel, on which an X-ray film has been exposed.

Autoradiographs display the length of the RNA that results from a presumed enhancer or promoter effect through the position and size of dark "bands" (spots) in the relevant "lanes" (columns) of the transparency-like film. Most strikingly, from the position and the length of the horizontal bands we can infer properties of the DNA fragments under investigation. However, this is an interpretation of black lines as a representation of DNA, where the features of the former are systematically interpreted as features of the latter. No similarity is involved, and if geometrical projection is at play, it is not clear which one. Sometimes, visual representation does not involve projection at all: an electrocardiogram reports the beats of a heart as a continuous

line, but that line is no geometrical projection of the pumping activity of the heart in any meaningful way: the Cartesian coordinates of lines are to be translated in thumps of a heart. But lines and thumps are not similar in any meaningful sense until we specify a way to translate the former into information about the latter.[123]

As we have seen, similarity does not make us proceed much in the semantic analysis of scientific pictures, consequently endangering our inferences from them about what they are supposed to represent. With everyday pictures, though, this does not seem to be a problem: how so? I take it that the translation of properties from, say, a family photograph or a journalistic picture is simply more automatic to us. We already know, from previous use and habits, that colours are supposed to be the same as the ones of the portrayed subject (unless there are some recognisable filters; of course, sometimes it is not so easy to detect them). Also, we automatically know that proportions are usually correct, but dimensions are not, some converging lines are instead to be read as parallel, and so on. The distinction, then, is just a matter of degree: with every-day, more "realistic" pictures, we are simply more familiar with the translation required to correctly read them.

Furthermore, nothing implies that even in the case that properties are shared, and thus similarity is in fact in place, we can reliably trust pictures. Consider the case of deepfakes,[124] where pictures and videos portray famous people saying things they have never said or doing things they have never done. From these cases it results as evident that similarity is far to be *enough* to warrant our inferences from pictures to their target. Even when you see what looks like a photograph or a documentary, you can only accept its content as justified if you know that the picture has been produced in a certain way (namely by pointing a lens at scene that records the incoming light etc.). Indeed, the best advice one receives from experts about deepfakes is to always check the source of the picture or of the video. If similarity were the issue, we would have been recommended to focus on the details of the image instead (which of course can be revelatory, but they are also the reason for which we fall prey of these fakes in the first place).

In summary, similarity accounts seem in principle inadequate to deal with the semantics and the epistemology of pictures, the problems becoming more acute in the scientific contexts. Let us now build up a positive alternative, starting from the various elements that were identified as missing in Meynell's account and from the acknowledgement of their relevance.

[123] If my choice to mention a fever graph seems partisan, as it should not be considered a picture at all, it is relevant to mention that such sorts of visual representations are also the target of Meynell's account. So, at least, her strategy seems to be wanting for this type of more "schematised" pictures. As I will show below, the framework I am proposing deals well with all the examples I am mentioning in this section.

[124] See e.g., Dan Milmo and Alex Hern's article "'Inceptionism' and Balenciaga popes: a brief history of deepfakes", published online in *The Guardian* on Monday 8th April 2024.

4.3 One step back: imaging a black hole

In this section, I offer some general background on black holes and how they relate to radiation measurements, as well as the general functioning of heatmaps. I then give a more detailed reconstruction of the production of the picture of M87*.[125] This will provide a helpful basis for the development of a philosophical account of this picture based not on similarity but on interpretation in section 4.4.

Imagine an object, far away in space (ca. 54.8 million light years from the milky way) with the mass equivalent to around 6.5 billion times the Sun but compressed so that its size is comparable to our Solar System. This physical object, which exhibits some of the most extreme conditions in the entire universe in terms of mass, speed and temperature, as well as in terms of effects on the texture of space-time by the effect of gravity, is what scientists think that probably lies at the centre of the Messier 87 galaxy (Gebhardt *et al.* 2011; Walsh *et al.* 2013): a supermassive black hole that the astronomers call M87*.[126]

Black holes are astronomical objects predicted by the general theory of relativity (Einstein 1915; Penrose 1965), and they are still central to crucial issues concerning the unification of GR with quantum physics (Hawking 1976; Giddings 20017). The very definition of a black hole is a matter of dispute, as it strongly depends on the disciplinary area of reference (Curiel 2019). Many of the technical aspects concerning the precise definition of "black hole" are irrelevant for my analysis, which intends to be general enough so that it is compatible with the main features of black holes irrespective of their exact theoretical definition.

The gravitational pull created by the black hole is so strong that, if something gets actually too close, it is irremediably swallowed into the black hole: not even light escapes (Schwarzschild 1916), and that is what gives the black hole its name. There is then a boundary beyond which even photons cannot escape and are inexorably attracted towards the centre of gravity. We call the line delineating this point of no return the *event horizon* of a black hole. Nothing escapes, so nothing can be observed[127] when it is beyond the event horizon: we can only make theoretical hypotheses on what happens beyond that line, as no trace is left.

[125] This is of course a very simple reconstruction of the entire process that made the production of the picture of M87* possible. The reader can find all the details in the six articles published by the EHTC team reported in the bibliography. A detailed but concise analysis, understandable also by non-experts, can be found in Muhr (2023). Important insights for my following analysis of the M87* picture as a scientific representation in DEKI's terms come also from the work of Doboszewski and Elder (2024), who however focus on the dimension of robustness.

[126] Supermassive black holes, with masses from millions to tens of billions of solar masses, distinguish themselves from the far smaller, non-supermassive black holes originating by the implosion of a star. Supermassive black holes are thought to exist in the centres of nearly all galaxies (Lynden-Bell 1969; Kormendy and Richstone 1995; Miyoshi *et al.* 1995), including in the Galactic centre (Eckart and Genzel 1997; Ghez *et al.* 1998; Abuter *et al.* 2018).

[127] Here, I am using the term observation in a technical sense, so that it encompasses any form of measurement – it should thus not be restricted to human vision alone.

Fortunately for us, the black hole attracts all sort of matter and energy from its surrounding. The first observational confirmations of the existence of black holes were due to the very fast, very small orbits of stars around a centre of gravity where no observable object was reported (Harms *et al.* 1994). Furthermore, and more importantly for our purposes here, there is something that we can observe in the external proximity of the event horizon. There, orbits at incredible speed what is called an *accretion disk*, namely a tremendous amount of matter, mostly ionised gasses, burning at a temperature ca. from 1 to 10 billion degrees Kelvin. Because of its high temperature, the accretion disk irradiates many forms of radiation among which light. Most of this radiation, of course, travels at wavelengths that cannot be perceived by the human eye but can still be measured by our interferometric devices.

When observing a black hole, what the scientists aimed at was a picture representing the distribution of radiation intensity of the electromagnetic field produced by the accretion disk. Interferometry is the main methodology employed in astronomy to measure the radiation intensity (also called "brightness") of an astronomical source.[128] The intensity of radiation of an electromagnetic wave is defined as square modulus of the wave's amplitude. If several waves interfere at a point, the resulting amplitude depends on the relative phases of the waves. Hence, the total intensity is a function of the relative phase differences. An interferometer is an instrument that make two or more electromagnetic waves interfere in order to observe their interference fringes.

What we obtain is a Fourier transform of the original wave, namely the sum of separate monochromatic components of the original electromagnetic signal. These components are waves with a defined frequency which interfere to produce the final signal. This decomposition is important in two senses. First, it is useful to clean the signal from possible informational noise deriving from the interference of other electromagnetic sources. Second, one is able to distinguish different frequencies of the radiation emitted by the same source.

By this measurement procedure, we can obtain interferometric data about the monochromatic electromagnetic components: the amplitude of these waves, their frequency, and as a consequence, the phase difference between them.

So, the basic functioning of an interferometer is to decompose the original electromagnetic wave and then calculate the phase difference between the various resulting decomposed waves in order to calculate the total resulting brightness of the source of the original wave. In this case, the source is the accretion disk surrounding the black hole M87*.

From the data about the electromagnetic waves emitted by the accretion disk, one can interpolate the (admittedly sparse) data and obtain a relevant distribution of the radiation from the accretion disk. This is basically the only thing we can measure

[128]The standard reference here is Thompson *et al.* (2017). More details about the specific methods that the EHTC employ are provided in the Science and Technology sections of the Event Horizon Telescope website.

from the immediate surrounding of the black hole. These measurements are then crucial, because the distribution of radiation can give us insight on the dimensions and variations of what lies within the event horizon. Furthermore, the data are useful to understand the structure and features of the accretion disk itself, its dynamic and composition, and its consequent interaction not only with the gravitational field produced by the black hole, but also with what surrounds the accretion disk itself. Altogether, these interferometric measurements are then a very important way to test the predictions provided by the general theory of relativity and the derived theorising on black holes.

In order to obtain an intensity distribution of a source, one has to infer it from the recorded data, that is, from the electromagnetic waves that reached the measuring device. This of course means that many different distributions are compatible with the data. This problem becomes even more serious in cases, like this one, where the data are very sparse and noisy.

The result of an interferometric measurement is a function of the source brightness, fringe separation, and orientation of the device. Therefore, increasing the number of, and distance between the detectors increases the accuracy and precision of the resulting data. Thus, for four days in April 2017, seven (systems of) telescopes in different places on the globe were pointed towards the centre of the Messier 87 galaxy and measured the radio signals coming from that region. The idea was basically to synchronise all the telescopes so that they could be used as one single telescope. The resulting "lens" of this composite telescope, even though fragmented, had the width of the entire planet Earth. The rotation of the Earth also allowed a less sparse sampling.[129] One and half petabytes of interferometric data were collected for each night of observation, that is, the greatest amount of data in the history of science for a single experimental measurement.

Once interferometric data are collected, the further goal is to assemble and interpolate them so that we have a visual representation of their source, namely the electromagnetic field of the region surrounding the event horizon. The result is our picture of the black hole. In this sense, the picture of the black hole is no photograph: the mechanism of production involved is completely different, and thus also our interpretation of the picture qua representation – and consequently, as evidential source for our potential inferences about the target. Photographs are usually meant to reproduce visible colours of the source, whereas here colours are just a way to indicate intensity of radiation reconstructed by data of the Fourier transform of the original signal.

Particularly, it is important to highlight both the reconstruction aspect (interferometric data concerning the phase differences of the monochromatic components of the original signal reconstructed as a radiation distribution) and the underdetermination

[129] Doboszewski and Elder (2024, p. 7) metaphorically call this a "sweeping" effect. Muhr (2023) also mention this aspect and refers to Thompson *et al.* (2017, pp. 31-34) for a theoretical explanation.

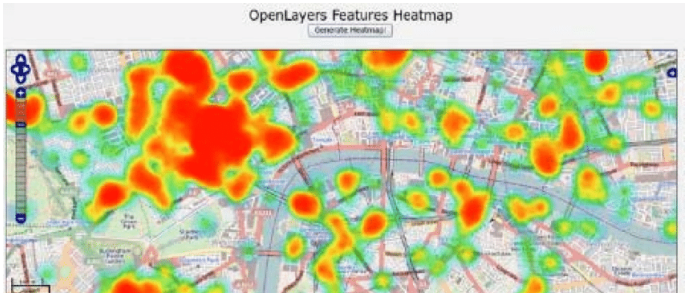

Figure 4.7: Example of heatmap showing pubs and restaurants in central London. In Bee *et al.* (Sept. 2012).

aspect (many distributions could result from the same data). Putting underdetermination aside and focusing on the reconstruction bit, it is useful to think of the resulting picture as a heatmap.

Technically, a *heatmap* is a visual presentation of data where colours are used to express values. For example, in Figure 4.7, we see colours expressing the density of pubs and restaurants in central London. Red areas indicate highest concentrations of pubs and restaurants, and the more me move towards the green (passing through orange and yellow), the less pubs and restaurants will be present. Furthermore, the spatial arrangement of the elements composing the heatmap usually translate in some properties of the represented phenomenon. The simplest case is when spatial properties of the heatmap just translate in other spatial properties of the target. In these cases, the translation will usually be some sort of geometrical projection. In our toy example, the location of a point on the heatmap is to be geometrically projected on the actual terrain of centre London, interpreting the map as a perspectival representation of a territory observed from above.[130]

The spatial properties of the heatmap, however, do not need translate into spatial properties of the target. For example, they may express logical, conceptual, or causal relations or properties of elements of the target system.

In the case of the picture of the black hole M87*, we try to obtain a heatmap similar to the one showing concentration of pubs and restaurants in London. In the black hole picture, though, we look at the intensity of the radiation emitted by the electromagnetic field surrounding the black hole. In this case, colours express the distribution of radiation intensity, going from black (low levels of intensity) to yellow-white (high levels of intensity). The spatial coordinates in the picture correspond to spatial position in space, though the geometrical projection in place is further complicated by the complex geometry of the black hole – more on this below. Let us look now at how this heatmap was produced.

[130]The geometrical translation may involve some approximation. For example, the surface of Earth is curved, and the heatmap in Figure 4.7 approximates the curvature to zero, as if the represented space was flat. This is because the difference is negligible when the purpose is to represent pubs and restaurants concentration in such a circumscribed area of the Earth's surface.

First, the interferometric data were fed to a supercomputer which integrated the data of each single telescope. The data were then further calibrated.[131]

Then, four teams of researchers were created to independently produce a visual output from the data. The four teams worked autonomously and were not allowed to talk with each other. Two teams ended up using the so-called CLEAN algorithm (EHTC 2019, p. 4). The other two teams used two different versions of the so-called Regularized Maximum Likelihood (RML) family of algorithms: the algorithm SMILI and the algorithm EHT-imaging, the latter created specifically in the context of the Event Horizon Telescope measurement (*ibid.*, pp. 4-5).

CLEAN is an instance of so-called inverse modelling and starts from the assumption that the image consists of point sources. Then, the areas with highest intensity are subtracted from the "dirty" image and added again as delta functions[132] to the "clean" image. This procedure is then reiterated until the removal of all points with intensity above a certain brightness threshold. In addition, "CLEAN typically convolves the many-point-source image model with a 'clean beam.' This beam is obtained from matching a Gaussian to the central component of the dirty beam, and it approximates the point-spread function of the interferometric data" (*ibid.*, p. 4).

The RML algorithms are an entirely different family of algorithms, and they go with the name of forward modelling. These approaches represent an image "as an array of pixels and only require a Fourier transform of this array to evaluate consistency between the image and data" (*ibid.*). The algorithm then tries to minimise the difference between the data and the image that we expect to find on the basis of previous theoretical models, informed by what the authors of the experiments call *regularisers* (*ibid.*, Appendix A). These regularisers are basically a set of parameters informing the theoretical models.

The details of how these two families of algorithms work are very technical and are not central for my analysis.[133] What is crucial is that CLEAN and the RML algorithms are very different in nature. This was important in order to show that the output image was not just an effect of the specific algorithm employed: algorithms that exhibit very different techniques and modelling assumptions had to produce the same picture. I will come back to this aspect of robustness in section 4.5.1.

Once the algorithms produced their output images, the four teams compared them and confirmed that all exhibited two important structural features: a ring shape with

[131] Cf. The Event Horizon Telescope Collaboration (EHTC) *et al.* (2019).

[132] A Dirac delta function is a generalised function on the real numbers, whose value is zero everywhere except at zero, and whose integral over the entire real line is equal to one. Very roughly, it is a way to treat events of probability 1 as if they were still a Gaussian distribution, but where the standard deviation tends to zero. This is relevant here because CLEAN starts from the assumptions that the image is composed by point sources, each of which is best described probabilistically by a delta function.

[133] The reader can find the specifics in EHTC (2019). Doboszewski and Elder (2024, pp. 15-16) provide a concise description of all three algorithms. For a general introduction to CLEAN, see Thompson *et al.* (2017, Chapter 11).

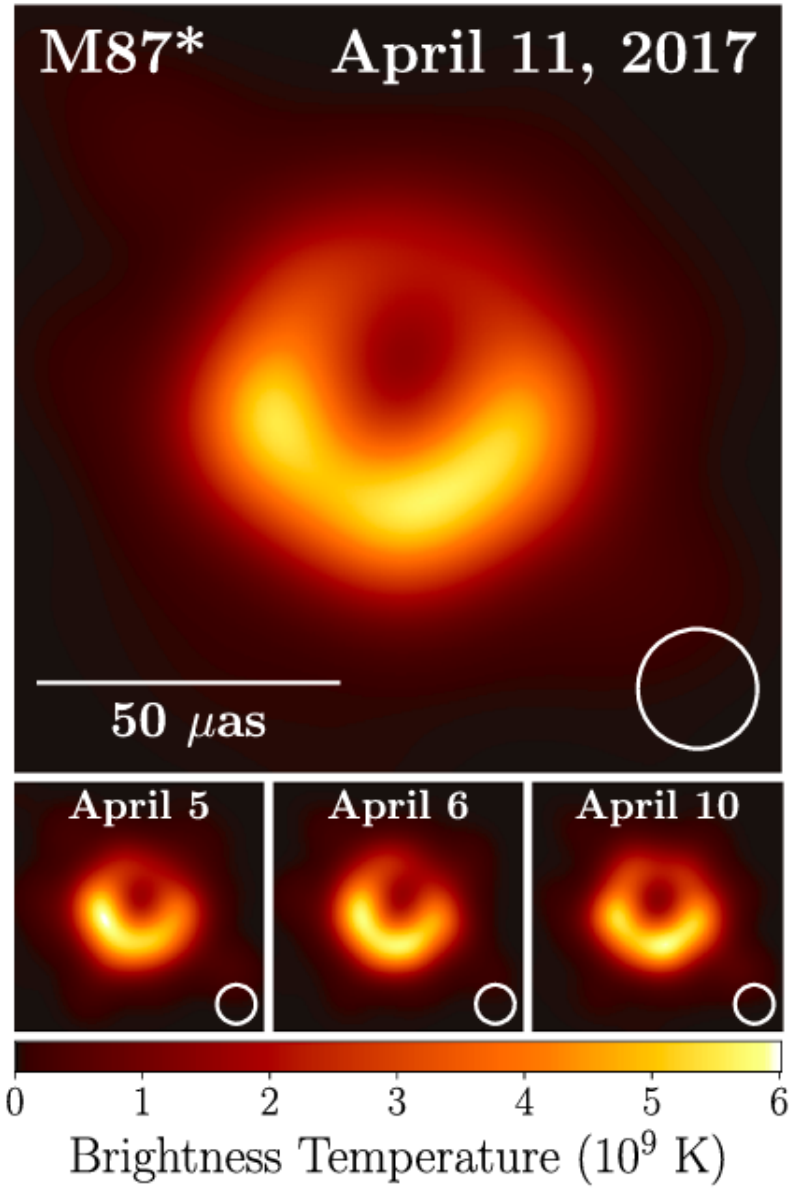

Figure 4.8: Picture of the black hole M87*. In EHTC (2019, p. 5).

more intense brightness in the south region of the black hole, and the diameter of the ring estimated around 40μas (EHTC *et al.* 2019, p. 9). The three algorithms were further tested against synthetic images, showing different geometrical shapes, which the algorithms had to reconstruct as with the picture of the black hole. This was done by surveying a large range of combinations of parameter values. By this parameter survey on synthetic images, the researchers obtained two results. First, they identified the fiducial parameters, that is, those parameters that allowed a more faithful reconstruction of the original image. Second, they proved some robustness of the algorithms by showing that they were sensitive to the input image: the outputs were really different for each synthetic image and the black hole picture, showing that there was a relatively strong counterfactual dependence of the visual output on the original source.

Four images were then produced from each algorithm pipeline, one for each night of observation, with a total of 12 pictures (four pictures for each algorithm). As a further step in making these pictures more reliable, all these images were further blurred to obtain a "common, conservative resolution" of each of them (*ibid.*, p. 20). Finally, to further emphasise the common features of the images produced by the three different pipelines, the scientists produced an average picture for each of the four days (*ibid.*, p. 21).

The top picture of Figure 4.8 is the resulting blurred, average picture for the observations on 2017 April 11 and was then chosen as general "representative example of the images collected in the 2017 campaign" (EHTC 2019, p. 5). From now on, I will refer to the top picture of Figure 4.8 as the target of my analysis.

4.4 An interpretation-based account of scientific pictures

In section 4.2, I have illustrated the failures of the similarity view, and in section 4.3, I delineated the basic information of how black holes relate to interferometry, how heatmaps generally works, and how specifically the picture of M87* was produced. Both these parts will now be the basis on which to elaborate a better philosophical analysis of the picture of the black hole. Specifically, the account aims at explaining in general terms how such a picture is to be read, or understood, and what epistemic value it has. In other words, I want to elaborate both a semantic and epistemological analysis of this picture. This will be done by applying once again the conceptual toolkit of the DEKI account of scientific representation, the elements of which will then emerge through my analysis.

So far, DEKI's authors have never fully engaged with cases of scientific pictures with real world targets. Frigg and Nguyen (2020, pp. 181-182) briefly discuss a visual representation of the Mandelbrot set, which is however a visual representation of a mathematical object. This lacuna concerning pictures of empirical target systems is quite general in the context of philosophical analyses of scientific representation.

Relevant exceptions are Perini (2005), who nevertheless focuses on diagrams (see Figure 4.2) and aims at a semiotic account of scientific pictures more than an epistemic one, Meynell (2013), and Tufte (1997), who focuses though on specific practices to enlighten different ways in which the visual can help a lot or do a lot of damage. Here, I show that DEKI has the resources to deal with the case of the black hole, and that this approach can be easily generalised to other examples of pictures used in science for epistemic purposes.

4.4.1 Denotation, interpretation, and the *Z*

First, as we have noticed already when criticising the similarity view, we need to specify that Figure 4.8 is indeed a representation *of* the black hole M87*. Following Goodman (1976, p. 5), I suggest that the relation in question, at least minimally, involves denotation, that is, the sort of relation connecting a name to its bearer, or more generally, relating a term to a class of objects named with that term. Therefore, we need to specify a designated target system that Figure 4.8 is supposed to denote. This target system can be identified as a certain region at the centre of galaxy M87. More specifically, the target is now beyond doubt thought to be the black hole M87* and the area immediately surrounding it. The target is now expressed via a proper caption associated with the picture itself.

Here, there is nothing intrinsic in the properties of a representation system *qua* symbol that determines its designated target. For what we know, the exact same picture may have been the picture of another black hole. Worse, it may well be the result of some coloured ink spilled on a black surface, or the random production of an artist performing in abstract painting. So, how the image was produced is a relevant information for its interpretation.

In the case of a mechanically produced picture like Figure 4.8, denotation is not just a matter of arbitrary stipulation: it depends on the causal process connecting the target system with the visual outcome of the picture. We are not here concerned yet with the level of epistemic reliability of this causal process: this will be the focus of section 4.5 below. What matters for now is that this picture, used as a picture of M87*, denotes M87* at least in part[134] in virtue of the causal relation between the two. This already seems to set this representation and in general mechanically produced pictures apart from the examples of model representations we have encountered

[134]I say "in part" to express caution: there may be space of philosophical debate here about intuitions on what suffices to establish a denotation relation. Suppose I take a picture aiming at a target X and, like in Antonioni's movie *Blow Up*, I accidentally capture a detail – perhaps, even the potential author of a murder – that I did not intend to. Is the face of the murderer the target system denoted by my picture? Suppose that, like the movie's protagonist, I don't realise this fact immediately, but only after many days. Does my awareness of this fact make a difference in what the target is? I have strong intuitions that intentions and awareness may be pivotal for establishing a denotative relation in this case, but some philosophers may disagree. For now, given that nothing of my analysis hangs on this specific problem, I suspend the judgement and postpone a thorough discussion to a later work.

in Chapter 1, where the denotation involved seemed more a matter of stipulative association between different parts of the models with elements of the target.

As we have already seen in previous chapters, the specification of what counts exactly as the target system denoted by a symbol is often an arduous task (Frigg and Nguyen 2020, pp. 13-14), and the result of a dynamic process that involves a continuous re-evaluation and adjustment of a denoting system. What exactly is the target of the picture of M87* depends on what we know about M87* and on the interpretation that we give of the picture itself. However, this should not be surprising by now: the problem of identifying the target systems of representations is a general one, from models to experimental specimens – to even names and their bearers.

Insofar as it refers to another system, a picture is referential: it is about that system. However, the aboutness of the picture of M87* can be hardly expressed purely in terms of denotation. The picture is no mere name, like all we could say about the name 'Napoleon' is that it denotes the person Napoleon. Like with scientific models, thought experiments and experimental specimens, the black-hole picture also allows us to perform surrogative reasoning about its designated target system. For example, the picture is supposed to show us the structure of an electromagnetic field surrounding the shadow of the black hole M87*. But how does it do that? For, we need to remember here, what we see are for now just colours. How do we move from colour patterns to information about a black hole?

It seems that we need some interpretation that translates information about the pictures into information about the target. Let us generally call this function an interpretation I, and let it be the classic interpretive function in place with heatmaps. As we have seen in the previous section, a heatmap associates the intensity of a quantity (such as the density of pubs in area) with a colour of the part of the map that represents that area. This is what happens in the picture of the black hole as well: coloured tonalities from black to white, passing through red and yellow are to be interpreted as levels of the intensity of radiation of an electromagnetic field coming from the accretion disk of the black hole to the telescope. The continuous gradation of colour gradation maps on the continuous values of the radiation intensity.[135] More precisely, the EHTC authors (2019, p. 5) specify that "the image is shown in units of brightness temperature" and indicate the units of brightness temperature scale, with a max. temperature of 10^9K.

Here, I agree with Meynell that the translation involved in the interpretation does not necessarily involve language: we do not have to put into words that colours are to be interpreted as levels of radiation intensity in order to actually perform such interpretation. Still, I take it that in most scientific cases a linguistic or at least mathematical translation offer us a higher level of specificity and rigour, as well as the

[135] Of course, the colours of the picture are only approximating a continuum, because they are the result of pixels that cannot in principle express an actual continuum of gradations.

applicability of mathematical formalism and derivation. For this reason, it is often welcome. In this specific case, for example, colour shades translate in mathematical values of radiation intensity. What is clear is that nothing of what we have said so far can be cashed out in terms of similarity between the picture and the target, not even passing through Meynell's scene: the relation between the picture (as a visual output) and the target is established by a proper interpretation and the actual causal processes of image-production.

Let us go back to the picture of M87*. We could now be tempted to cut the chase and say that the properties of the interpreted picture as a distribution of electromagnetic intensity are also the properties of the actual black hole *simpliciter*. However, this may be too quick. First, it is possible to represent the same object in many different ways. As Curiel (2019) has eloquently shown, black holes themselves can be theorised in many different ways, each way implying a different definition. Second, and more importantly, even when properly interpreted, representations often involve abstractions, idealisations, and approximations.[136] So, the properties shown by the representation may not be *exactly*, or even remotely, the ones possessed by the target. The very story of how the picture of M87* was produced, particularly the many steps necessary for its creation, should make us cautious about both the factual correctness of the picture and the realism of the picture (understood here simply as lack of distortions): the picture of M87* is the result of a complex computer elaboration and theoretical interpretation of sparse interferometric data via different models obtained from general theory of relativity.

More generally, if we had been interested in a purely true description of the target system, we should have stuck to the interferometric data alone. But truth here is not the only epistemic value that we are interested in: we want to see if those data, for example, are compatible with the predictions of our most updated scientific theories. Therefore, we need to create a system, a picture, out of those data that fits the predictions of those theories.

Given the I function that we have specified above, we can see that the picture represents the black hole *as* a heatmap in the sense illustrated in section 4.3. More specifically, the I makes the picture a spatial-electromagnetic-intensity-representation, which functions as a heatmap where the relevant quantity indicated by the colour is the electromagnetic intensity. If we employ once again Goodman's concept of Z-representation, we can say that the picture is a radiation-heatmap-representation of M87*. What we conceptually achieve by insisting on the distinction between representation-of and representation-as is that the heatmap in question may still be understood as literally different from the denoted target with respect to the exemplified properties in some relevant way.

[136] As noted in the previous chapters, I appeal to the taxonomy proposed by Frigg (2022), p. 312.

4.4.2 The properties exemplified by the picture of M87*

The epistemic function of representation-as we have amply illustrated in the above chapters, is exemplification. For the sake of clarity, I report the technical definition again: a system X exemplifies a property A *iff* X instantiates A and X also refers to A. This reference of an object to a property, as we have already seen in the previous chapters, usually makes that property salient or highlighted. In a scientific context, a property is usually exemplified by being epistemically accessible for an expert user.

The picture of M87*, when interpreted as a radiation heatmap, definitely exemplifies certain properties. For example, when one looks at the picture, one will easily notice the characteristic "shadow" at the centre of the radiation distribution. The diameter of the ring appearing in the visual reconstruction was further estimated to correspond to an actual diameter of 40μas. Furthermore, the picture exemplifies an asymmetry in the distribution of radiation, with a higher intensity in what is usually called the South region of the accretion disk with respect to the North region of disk.[137]

Most importantly for our analysis, by exemplifying certain properties, a representation also acts as a selective system: in order to make some properties salient, other properties will be ignored or overshadowed. The focus on the accretion disk, for example, required the elimination of informational noise produced not only by objects and radiation sources present between the black hole and the Earth, but also by radiation sources in the external surrounding of the accretion disk. Electromagnetic radiation from the surrounding regions is then discarded as informational noise in order to make us focus on what we are here more interested in: the shape of the shadow, its dimensions, and the distribution of radiation immediately around it.

The illustration of the informational noise that I have just offered is in fact a simplification. While it is true that a great amount of matter and thus radiation emitted is present around the accretion disk, the geometry of the spacetime surrounding the black hole is very complex, so spatial designations like "in the proximity of the disk" are only approximations of what may actually be the case at the centre of the galaxy M87. As an implication, what I call noise here was surely eliminated so that the picture could emphasise the properties of the accretion disk, but this elimination does not have the clear-cut spatial distribution that my simplified reconstruction may suggest.[138] Nevertheless, my simplified reconstruction does not invalidate my more general point that the picture in Figure 4.8 exemplifies certain properties (e.g., the characteristic shadow, and as we will see in a moment, an asymmetry in radiation

[137] I don't want to suggest that these are the only properties exemplified by the picture. The spatial information that the picture provides is very rich, with a lot of fine-grained detail about the distribution. For the sake of simplicity, I focus here on the most crucial, macroscopic features exemplified by the picture.

[138] I thank Paula Muhr for pointing this out and clarifying the simplification to me in private conversation.

intensity between different areas of the accretion disk) by at the same time bracketing or neglecting other features of the designated target system.

As we have seen when discussing Walton's idea of principles of generation in the context of scientific representation, we also know that exemplification usually works dynamically: on the basis of certain exemplified features and certain theoretical assumptions, we can draw inferences that lead us to further exemplified properties.[139]

For example, the picture exemplifies the asymmetry in intensity of radiation between the North and the South regions of M87*. From our physical theories, we also know that the gravitational pull of the black hole makes the material composing the accretion disk rotate in a vortex. From these two premises, we can infer that the radiation we observe in the South region comes towards us, while in the North region moves away from us. So, the South region would appear brighter not because there actually is a higher intensity of radiation in that region than in the rest of the accretion disk. The difference in radiation intensity recorded is then just the result of the Doppler effect.

We can go further and, coupling the picture properties with our models of black holes and "information on the inclination angle, [...] derive the sense of rotation of the black hole to be in the clockwise direction, i.e., the spin of the black hole points away from us" (EHTC 2019, p. 9). Also, the EHTC authors explain the brightness asymmetry in the South region of the black hole "as relativistic beaming of material rotating in the clockwise direction as [...] moving toward the observer" (*ibid.*).

Other interesting properties of the system can be inferred by looking at the picture. For example, from the scaled dimensions of the pictures, astronomers can infer that the dark object at the centre cannot be a naked singularity or a wormhole, but it is more likely to actually be a supermassive black hole as it is theoretically predicted by the general theory of relativity (Bouman 2020).

This active productivity of the picture of M87*, in terms of offering a means of developing new hypotheses about its target system, has not passed unnoticed. Muhr (2023), for example, compares this astronomical picture to a map and, referring to the terminology used by Krämer (2013, p. 276), talks about a "cartographic impulse" of this type of images. The picture thus, instead of simply being a passive reproduction of data or a mirror of the target system, it also allows to explore it epistemically, like a map allows exploration in the spatial physical sense. I perfectly agree with Muhr's reading, but I would insist that all pictures that function as epistemic representations can in principle and often do exhibit this cartographic impulse. Even a portrait from the past, or a photograph from the present, can be used to epistemically explore

[139] It is important to remind here that representational inferences are ampliative in nature and thus not necessarily producing correct conclusions. Also, they may lack the required justification, given a specific context (also non-epistemic factors like risk may play a role in setting the threshold of accuracy necessary for a certain investigation). Below, I talk more in detail about the justification for inferences from mechanically produced picture in general, and from the picture of the black hole in particular.

its target system, the period in which the picture was produced, and so on. The difference comes in degree, if any, and not in the cartographic, epistemic nature *per se*. In this sense, I don't think that the picture of M87* is an exception or a special instance among pictures: all representations that exemplify properties are legitimate candidate for cartographic use in Krämer's sense.[140]

In her paper, Muhr also shows how the many different visual outputs of the EHTC project have been used in a complementary way. Not only different pictures and visual outputs were produced to answer different questions, but also different pictures, produced for example via computer simulations of theoretical models (Muhr 2023, p. 18) have been used in combination with each other. My focus on one single picture should not be taken as in sharp contrast with this multi-picture practice. My analysis takes single pictures as a valid, if of course incomplete, level of analysis of a complex scientific practice. Muhr's emphasis on the plurality of pictures is still compatible with my analysis. One can say that different pictures exemplify different properties, and a set of pictures can constitute a more complex representational system. Together, all the pictures of the black hole will then capture different aspects of it by individual contribution and mutual comparison, and different interpretations and keys will be required depending on the specific features we want to impute to the target system.[141]

Exemplification, as we have seen in the previous chapters, comes in many ways. Since in this Chapter I am focusing on pictures, it is natural that the exemplified properties will be made salient by the means of visual properties: colours, shapes, topological relations, and relative dimensions. However, it is important to insist that these visual properties, when appropriately interpreted, have nothing to do with what we would "see" if we could observe the target directly. In the case of the picture of the black hole, this is manifest because what we observe in the picture is not the actual black hole, but a coloured heatmap of the radiation intensity of the electromagnetic field surrounding the black hole looked at from a specific point of view.[142]

Here, we can see how problematic any appeal to similarity can be, because it can easily lead us to assume that the representation is similar to the target by sharing properties with it, while this is often not the case, or only *modulo* a given interpretation of the representation's properties in a more or less precise conceptual

[140] In private conversation, Muhr has explicitly agreed with me on this point, and emphasised that even everyday photographs need to undergo interpretation: they are not simply transparent copies of the external world, but results of technological practices that require hermeneutic activity.

[141] An interesting question arises about the actual target system denoted by the picture. I take it that different pictures of the black hole, while exemplifying different aspects of it, can and should be interpreted as all pictures of the same spatio-temporal region, so that we can make sense of comparisons between them. Also, nothing rules out the possibility that the boundaries of the target system be vague or changing by the very activity of measuring, studying and representing them. This is in fact the case for most target systems, not only of pictures but also of models, experimental specimens, and so on.

[142] Also, notice that different visual setups would be more or less useful depending on different epistemic and pragmatic purposes. On this, see Tufte (1997).

framework. In other words, only if visual properties are interpreted as non-visual ones we can start using the picture of M87* as a representation of M87*. Thus, not only the two systems do not share the relevant properties that we focus on to use the picture as a representation of its target: it is also the case that the representational use of the picture *requires* us to interpret the two systems as *not* sharing the properties we focus on.

4.4.3 Imputation and de-idealisation

Of course, as we have seen with thought experiments in Chapter 2, the aim of our epistemic representations is primarily to impute some of the exemplified properties to the denoted target system. I remind the reader that imputation here is simply property attribution: as we have seen in the previous chapters, representations may just misrepresent their target. So, imputation does not conceptually rule out the possibility that this attribution be mistaken.

As discussed before, we want to distinguish between the derivational correctness of a representation and its factual correctness (see above, section 1.6). We now focus on what is derivationally correct, namely which properties we are supposed to impute to the target system on the basis of what is exemplified in the Z-representation. As I mentioned before, though, we should be cautious: the process of creating a surrogative system will almost inevitably result in some forms of distortions or idealisations.

In the previous chapters, we have then introduced Frigg and Nguyen's (2020) concept of a key, namely a mapping function associating properties exemplified by the model to the properties we eventually impute to the target system. We have already encountered many different types of keys used in the sciences: scale factors, geometrical projections, limit keys, counterfactual and susceptibility keys, approximation keys, functional identity keys between biological mechanisms, and phylogenetic keys.

Is a key, or a set of keys, involved in the picture of M87*? Some of the classic examples of keys that I mentioned in their work come definitely in play. For example, a scale factor will be required to translate the picture's dimensions into the real ones. Also, a geometrical projection is in place. Indeed, exactly when interpreted as a radiation heatmap, Figure 4.8 is two-dimensional: we need a geometrical projection that translates the 2D image in a 3D object. Again, we have a distortion of the actual spatial properties that allows the picture to exemplify certain important properties that would have not been made salient otherwise. For instance, the asymmetry in radiation intensity between North and South. We need then to specify that the projection involved is perspectival, and only by this specification, coupled with the radiation asymmetry in the picture, and our knowledge of theoretical physics, we can then infer the Doppler effect as the best explanation for the asymmetry itself.

Another important key at stake, here, is approximation: the colours of the picture are not really expressing a continuous gradation, because they express this gradation

by the use of discrete units of information, namely the single pixels composing the visual output. The levels of radiation intensity, instead, are supposed to be continuous. We then need a key that translates the discrete values exemplified by the picture into ranges of continuous values actually imputed to the target system.

There is, however, another interesting aspect of this picture that concerns a step in the process of its production that we have seen in section 4.3, namely the blurring intervention that the scientists performed before averaging the four images into one. As Doboszewski and Elder (2024, p. 21) emphasise, this process of blurring makes "the images lose resolution and (potentially informative) structure", but it also "increases the security of the results, resulting in images whose features can be considered more trustworthy that their unblurred predecessors". They then conceptualise this blurring process as "analogous to adding larger error bars to a weakened conclusion – sacrificing the precision of the result to improve the confidence in the fidelity of the final images" (*ibid.*).

I think that it may be useful to consider the blurring action as implying the use of a key, which re-translates the blurred properties of the image into the more exact properties resulting from the applications of theoretical models to the interferometric data. Here, we can see that the re-translation performed by the key depends on the epistemic value we give more emphasise too: either precision an accuracy, or, if the blurred image is preferred, to robustness and reliability. This type of deblurring key also seems the converse of another type of keys used, for example, in modelling the climate. There, experts often downgrade the likelihood of a certain result trying to factor in the high level of uncertainty about the model ensemble's reliability (IPCC 2010).[143]

In summary, I have shown how the DEKI account can enlighten our understanding of the picture of M87*, specifying the interpretation involved, the properties that it is supposed to exemplify, and the keys at lay to impute those properties to the target system.

4.4.4 Beyond the black hole: pictures are no photographs

While being an extremely manifest case of scientific picture where the relation of representation and similarity breaks down, my focus on the picture of M87* and its treatment as an epistemic representation in DEKI's sense is not the result of cherry-picking, but it enlightens a general pattern, which is applicable to many other examples of mechanically produced pictures.

For instance, let us consider the autoradiograph shown in Figure 4.6. This picture has a designated target system which the picture denotes, namely a certain DNA fragment under investigation. Furthermore, the visual properties of the picture are

[143]The issues concerning how this practice can be epistemically justified are numerous and deep. See Harris (2021, particularly pp. 245-261).

interpreted, so that the various little black bands, particularly their length and position, are associated with a specific pattern of decay emissions from a distribution of the radioactive substance in the DNA fragment of interest. This interpretation allows us to look at the visual output as a Z-representation of our DNA fragment, which we could call a decay-emission-pattern representation of a DNA fragment. The shape, disposition and length of the little black bands are then interpreted as properties of this decay process and exemplify certain properties of the observed material. Finally, certain keys of approximation will be at work in order to convert the properties of the Z-representation into the ones we want to in fact impute to the target system.

As one can immediately recognise, similarity here has no role to play, particularly if one focuses on visual properties. A DNA fragment does certainly not appear visually like the picture shown in Figure 4.6. Both the step bringing us from the picture to the Z, and from the Z to the actual target, are interpretive activities, where the sharing of properties is neither sufficient nor necessary. What is relevant is that some properties of the picture are interpreted so that they are also exemplified, and then mapped onto the target via a key.

Only because of their visual, pictorial nature, we may think that the picture of the black hole and the visual outcome of an autoradiograph both function like photographs do. Given that photographs have for long constituted the emblematic example of faithful, trustworthy representation (think of their use in trials, for example), one may think that we effectively use pictures in science for the same reasons. But this is a mistake based on a wrong generalisation. In the case of photographs, the exemplified properties of the representation (colours, topology, proportions) usually are the visual properties displayed, so the Ifunction mostly maps visual properties onto themselves. These properties are then simply mapped onto their targets via an identity key. Surely, many pictures distort their targets: black and white pictures change the colours, and most pictures reduce the dimensions of their targets. But we are so used to such distortions that the translating keys come to us automatically. The mistake is to think that simply because in photographs the I and the key are apparently doing little work, merely relating shared properties between picture and target, then *all* visual representation would function in the same way. But this is not the case, as just illustrated with the picture of the black hole and the autoradiograph. Instead, one can see that the case of photographs is just a special case of representation in DEKI's sense, where most of the exemplified properties (visual, perceptual properties) are preserved through the interpretation process, and the I-mapping and the keying-up are mostly done automatically. But this is a psychological fact due to repetition and habit: most scientists are often able to automatically translate visual properties of scientific pictures into the interpreted ones just as automatically as we do daily with photographs. This does not mean though that some translation is in fact involved.

In general, imaging techniques abound in the sciences and are particularly frequent

in the medical context: X-rays, computerized axial tomography (CAT), ultrasound and Magnetic Resonance (MRI), and Positron Emission Tomography (PET) are all good examples of processes to obtain pictures of the human body. The visual outputs of these various techniques are then used to make inferences about the represented parts, organs and tissues and further guide practitioners' diagnoses and treatments. They are therefore all perfect instances of epistemic representation.

One can apply the DEKI's schema to all these cases. The details for each of these types of imaging go well beyond the scope of this Chapter. The aim of these brief remarks is, first, to remind one crucial take-home message, namely that, like with the picture of M87* and the autoradiograph, all these pictures should be understood as photographic representations. In other words, a correct interpretation of a medical scan usually requires some form of interpretation, mapping visual properties into non-visual ones, and requiring a key to translate simplified or approximated or even idealised properties into the ones we impute to the target system. Second, that the DEKI account can provide a general guide for a philosopher interested in the semantic and epistemic use of such pictures. What is the designated target system? How does the interpretation of the picture function exactly? What are the exemplified properties? Are there keys at work, and if so, which ones? While I leave the specific answers for each different type of medical scans to future investigation, my analysis of the picture of a black hole will exemplify a general strategy to apply to other instances of visual representation.

I have illustrated DEKI with respect to the picture of the black hole M87* and suggested that this can be applied to other instances of mechanically produced pictures. What remains to be said, of course, is how we move from the picture as a representation to the picture as piece of evidence for our inferences. To do that, we have to clarify what justification we can offer for the inferences we perform on the basis of the picture about its designated target system. In other words, what justifies the attribution of a certain set of exemplified properties via a specific set of keys? This is the task for my next section.

4.5 From semantics to epistemology

4.5.1 Production, interpretation, and justification

From what I have said in the previous section, one can see that the very interpretation of Figure 4.8 as a representation of the black hole M87* is inevitably dependent on its history of production. We can notice this dependence in at least three steps of the DEKI framework where the model system, namely denotation, the interpretive function I, and the keying-up.

First, as I have already anticipated, the picture denotes its target not just as a result of simple stipulation. It could have been a pure act of stipulation. However,

the relation between the symbol (the picture) and the target system will be justified on the basis of the mechanism of production of the image on the basis of observations of the target. Again, we are not here concerned with the reliability of this process yet: I am just acknowledging that the denotation relation between the picture of M87* and that black hole will be at least in part motivated by the fact that there is a story we give about the causal process connecting the target with the final visual outcome.[144]

Second, our interpretation of the colours and shapes of the picture is substantially informed by our knowledge of how the picture was produced. Namely, sparse interferometric data about the phase difference of electromagnetic waves emitted by the accretion disk of M87* were collected and then used as a basis to produce a visual output. This provides us with a function I that associates the colours shades and distributions on the picture to levels of radiation intensity of the electromagnetic field.

Third, the keys will be again chosen on the basis of what we know about the production of the picture: the scale factor will depend on the estimated distance of the Earth from M87*, the width of the "lens" resulting from the coordinated and simultaneous use of the seven telescopes as parts of a single, huge telescope, and the translation of the interferometric data into pixels. Similarly, the approximations and the deblurring key are both motivated by the way in which the picture was produced. The approximation will depend on the robustness of the imaging algorithms and the consequent reliability of the visual outcome to be faithful to the interferometric data. This faithfulness comes in two ways: the visual output is not only compatible with the data, but it would also change were the data different. Second, the key translating the blurring image into a deblurred one, is just the converse of the blurring action performed by the scientists at the end of the picture production.

For now, I restricted myself to talk about DEKI's elements, thus remaining within the scope of an analysis of the derivational correctness of our inferences from the picture to the actual black hole. Indeed, the DEKI account remains silent on the factual correctness of a representation, as we have already seen in the previous chapters. The justification for our inferences from a representation in the sense of factual correctness will inevitably go beyond the single representational system.

However, one can notice that, like our interpretation of the picture, our justification for its factual correctness as a representation will also hugely depend on the history of production, and specifically on the causal mechanisms connecting the

[144] Again, caution is here to be recommended. An abstract painting of the same black hole will denote the black hole without requiring any reference to causal processes of image production. While I agree that denotation can be obtained in many different ways, I would like to insist that the way in which denotation realises here seems interestingly different from cases of pure naming by ostension or by stipulative fiat. In cases of measurement acts, the measurement output denotes the measured system, at least in part, by the means of some story about a causal process relating the latter to the former. This story, it goes without saying, may be completely wrong.

target with the final visual output. Our inferences will be justified, then, to the extent that we can prove the causal relation reliable or stable. In other words, we want to establish a counterfactual relation between the state of affair constituted by the target system and its behaviour, and the resulting visual output of the image.

Along these lines, Doboszewski and Elder (2024) analyse the picture in terms of robustness analysis,[145] by showing that the multiple algorithms employed for the imaging converged on similar results even if taking different procedures and assumptions, and they also exhibited reasonable sensitivity to data – namely if the data had been different, the resulting visual output would have been different in a consistent, systematic manner. The robustness of the algorithms was assessed as follows. The three algorithms were tested against synthetic images, showing different geometrical shapes, which the algorithms had to reconstruct as with the picture of the black hole. This was done by surveying a broad range of combinations of parameter values. By this parameter survey on synthetic images, the researchers obtained two results. First, they identified the fiducial parameters, that is, those parameters that allowed a more faithful reconstruction of the original image. Second, they proved some robustness of the algorithms by showing that they were sensitive to the input image: the outputs were really different for each synthetic image and the black hole picture, showing that there was a relatively strong counterfactual dependence of the visual output on the original source.

This was necessary to secure a reliable counterfactual relation between the data and the visual output obtained from applying the algorithm, and consequently, the accuracy of the latter with respect to the former.

The causal relations between target and pictures that I have illustrated so far are supposed to express counterfactual relations between the state of affair represented and the visual output we are presented when the imaging process is completed. This counterfactual relation, in turn, substantiates the inferential stability[146] of our inferences from a picture to its designated target system. The more numerous, complex, and (more importantly) weaker are the steps in the causal chain of producing the picture from the target, the more difficult it will be to justify the inferences we draw from the former to the latter.

Here, I am not arguing that the picture of M87* is, in fact, epistemically reliable for all the possible inferences one can draw from it. Something in the process of production may have gone wrong, or that process is just unable to justify some of the inferences that we think we can draw. My point is that, in the case of pictures, the correctness of the imputation is built into the interpretation of the picture. If

[145] The philosophical literature on robustness is vast, the origin of which can be tracked back at least to Levins (1966, 1968), Wimsatt (1981, 1987). The reader can see Weisberg (2013, Chapter 9) and Hudson (2014) for discussion and references.

[146] Here, I am importing the concept of *inferential stability* from Roskies (2008) who employs it to analyse the case of inferences about brains from scans obtained via magnetic resonance imaging.

our imputation fails, and the target does not in fact possess the properties we want to attribute to it, we have made an error somewhere; but in principle the picture is constructed and interpreted so that "T possesses the property Q" is true. As a consequence, if one wants to assess the reliability of our inferences from a picture, one has to look at the process of production, and how this relates with the interpretation of the picture and the interpretation (I-function and keys) employed.

The dependence of both the interpretation of a picture, and the justification for the inferences that we draw from them, on the mode of production of the picture similarly applies to the other examples of pictures I mentioned in this Chapter. Even traditional photographs and their apparent preservation of visual properties is justified via an appeal to the way in which they are produced, namely by the effect of light and their impression on a photographic film.

The importance of the mode of production becomes starkly clear when such counterfactual relation between target system and visual output does *not* support our interpretation of a picture and thus our inferences from it. Let us consider the case made by Klein (2010) concerning *functional* MRI scans.

In order to understand what functional MRI is, it is first necessary to briefly illustrate the general functioning of MRI, specifically in diagnostic scenarios. Generally speaking, MRI is a technology that allows to produce scans of soft tissues in the body, especially the brain. In this type of scans, the patient is placed within a machine which produces a magnetic field that make the protons of hydrogen atoms in our tissues align. Then, a radiofrequency current is emitted through the patient, making the protons spin dis-align from the magnetic field. When the radiofrequency is turned off, the protons reacquire their original alignment to the magnetic field, and some sensors of the MRI machine detect the time for them to realign, as well as the energy released by the protons for the realignment. These two dimensions (time for realignment, and quantity of energy) are interestingly associated with the different types of tissues the atoms of hydrogen are parts of. Given that different types of tissues will take different amounts of time to realign, or release different levels of energy in the process, the MRI is a very efficient technology to mark the presence of different tissues in a body region. The most common way is, roughly speaking, by comparison: if a region which should be uniform with its surrounding instead exhibits important different values from it, then we plausibly have an indication of a pathology – for example, a tumour.

In contrast with the use of MRI in medical diagnosis, functional MRI, or fMRI, is used to associate the neural activity of certain region of the brain with specific cognitive, psychological tasks or function.[147] Functional MRI constitutes one of the most employed technology in cognitive sciences and neuroscience, with most of our attempts to provide a general theory of cognition based on empirical results obtained

[147] Useful introductions to functional MRI can be found in Huettel (2004) and Buxton (2009).

through it.

However, Klein argues that functional MRI scans should not count as evidence for such associations between the brain and the high-level cognitive functions. The reason, Klein argues, is that the MRI apparatus only detects the level of blood oxygen in the brain tissues. As an indication of such value, MRI are excellent sources of evidence. The problem is that this level of blood of oxygenation is causally related with cognitive activity only in a very weak sense. This is because, as Klein (2010, p. 269) argues, the brain is a causally dense system: when a certain cognitive task is carried out, many parts of the brain activate, even if they are not causally responsible for that specific task. Therefore, only because a certain region of the brain shows higher level of blood oxygen during the performance of a certain cognitive task, this fact does not necessarily imply a causal role of that region for that specific task.

Again, we see that when we want to interpret an MRI scan and justify our inferences from it, one is required to look at how the picture was produced and how this mechanism causally relates the target system with the visual output. In this case, until some further justification is offered, the causal mechanism connecting the target system to the picture does not legitimise strong inferences about the causal role of specific areas of the brain with respect to certain cognitive tasks. Therefore, it seems that the interpretation of a functional MRI scan as a map showing associations between brain regions and cognitive activities lead us to unjustified inferences: in other words, we obtain a divergence between interpretation and justification. Of course, nothing in DEKI rules out the possibility of interpreting functional MRI scans as images of cognitive activities. However, an analysis of the mode of production of the picture gives us the necessary resources to argue that that interpretation is not well supported.

We can then see how looking outside the single representational system one can find strategies to confirm or rule out certain interpretations of pictures (and in general, of representations *tout court*). Our Goodmanian framework results then flexible enough to allow pictures to be interpreted freely (and sometimes, in an unjustified way), without thus implying that anything goes: once we turn to justification, only some interpretations (and some consequent inferences) will be warranted and thus legitimated.

In summary, any sort of interpretation of a mechanically produced picture, and the consequent reasoning that converts information about the picture into information about a designated target system, needs to be grounded on the production of the image. This production has to show some form of robustness or counterfactual stability holding between the visual outcome in the picture and the original target system.

4.5.2 A substantial difference between "pictures" and models

The dependency of interpretation of a representation, as well as the justification of the inferences we draw from it about its target, on the history of production of the representation, may not surprise us. In some disciplinary areas of philosophy of science, like what is usually called Science and Technology Studies (STS), much emphasis is always given to the history of production of scientific tools, as well as to the social, cultural, and political factors that influenced that process of creation. I want to suggest, however, that this close dependency of interpretation of a mechanically produced picture from its mode of production is more interesting than it may look at first sight.

In this section, I want to argue that the way in which we justify our inferences from the kind of scientific pictures I have considered so far is different from the way we justify inferences from other instances of scientific representations.

Nothing of the sort of what I have said about the picture of M87* applies, for example, to models, thought experiments and diagrams – which I will call models now for short. I want to argue now that the process of justification for these representational systems is different: in these cases, justification usually comes from theoretical assumptions and empirical evidence informing our interpretation.

A model system is usually constituted by a set of assumptions on that system (an abstract object or a material one), often in interaction with each other. Let us take the simple case of an assumption that is expressed by a certain functional relation between two quantities. How do we justify such an assumption? There are many ways in which this can be done. First, if the target system actually instantiate that functional relation and we can ascertain this by direct testing, we have an optimal reason with which justify our assumption. Of course, direct testing on the target system is often an unavailable path, and we have to justify our assumptions in a more indirect way. For example, our assumption in its current form may directly derive from more general theory in the relevant discipline, or being a model of that theory, with model here understood in the logical sense of a structure about which the axioms and theorems of the relevant theory are true. Or, our assumption may be a simplification of a more general functional relation that however is intractable in its current form (e.g., an equation with no analytic solutions). Here, the justification follows from our reason to hold the original formula, plus some further reason to consider the simplification acceptable. Alternatively, our assumption may boil down to a hypothesis abstracted away from data. For example, it may be the result of an abductive inference on the basis of current observations. Here, some further justificatory analysis is required for the inference to the best explanation.

In all these cases, the justification for the assumption will be more or less provided on the basis of previously acquired knowledge. However, the assumption could also be something completely new, detached from theory and experiments. The justification for that assumption will then solely depend on the success of the model

as a whole. Success can take many forms: empirical adequacy, unification, explanation by providing an underlying mechanism. The more the model proves itself successful, the more we can justify its further application as an epistemic surrogate system.

Of course, interpretation is crucial also in this epistemological analysis. Each relevant piece of formalism, or material element in the case of a material model, has to be endowed with a theoretical interpretation in order to then investigate the justification for such interpretation. Furthermore, various assumptions will interact with each other in the model. That interaction will also need some further justification, at least in terms of how plausible it is that different assumptions can be taken to hold simultaneously. From the set of initial assumptions in the model, there will be further properties that we can derive, and the justification for the latter will depend on the justification for the former.

This is, I take, what we can say about justification for model inferences without entering the details of each specific model. I take this analysis to naturally apply to scientific thought experiments and diagrams as well. In contrast, I previously showed that the process of justifying an inference from a mechanically produced picture is crucially different. Pictures of the sort I have focused on in this Chapter offer an alternative avenue to justification, namely an appeal to the causal processes that eventually produced the picture, and the associated, more or less strong counterfactual relation between what is true according to the picture and what is true about the target. The correctness of our interpretation of the picture itself will actually depend on this causal relation between picture and target, in a way that does not seem to occur with models and diagrams.

In this sense, at least some pictures are much more assimilable to measurement outputs. The latter can also count as representations according to DEKI – indeed, one of the favourite examples of the authors of the account is the litmus paper used in chemistry. But as one can see now, the justification for inferences made on the basis of measurements and pictures depends on considerations about causal mechanisms and counterfactual dependence with the target system. This does not seem to be the case for most of our theoretical models. On this basis, I distinguish two substantially different types of representation, according to their characteristic justificatory style: measurement representations, on the one hand, and model representations, on the other.

The causal relation that I am interested is at the root of the inferential stability from pictures to targets. Nevertheless, it is important to remind the reader that the characterisation of this causal relation is still based on theoretical assumptions – particularly our theories of measurement devices – and previously acquired empirical knowledge. This follows from the general understanding of knowledge, particularly scientific one, as a holistic phenomenon in line with the suggestions contained in Elgin (1996). Therefore, I do not want to undermine the theory-ladenness of our interpretations of images: whether an image is causally linked to its target and

how accurately so can be and often is a matter of dispute, even among experts. So, I do not want my focus on causation here to foster the suspicion that I am considering pictures somewhat more "direct" or reliable representations than models or other types of representations. For even the assessment of the hypotheses about the causal relations in play in our production of images will strictly depend on the theoretical framework we are assuming in the first place.[148] My point is just that we employ different epistemological strategies to justify our inferences from these types of representation with respect to what occurs with models.

This is in the end basically the same result that we obtain when we compare the epistemic status of models or computer simulations with experimental specimens. Both count as representations in DEKI sense. And the inferences we draw experimental specimens about their extrapolation class is not always stronger, or more objective, than the ones that we draw from models or simulations. For what matters is just the stability of our inferences, not whether we are directly intervening and measuring, or instead building a fictional or virtual surrogate system.[149]

Moreover, in a specular way, I do not want to deny that how the way in which we interpret models also depend on how we construct them. And of course, the way in which this production takes place may causally depend, in a loose sense, on the target system: modellers often build and construe their models in order to be effective representations of some pre-determined target system. So, I grant that models can be motivated, heuristically, by the target, but there is no mechanism like a telescope that turns a target into a model. My point simply concerns the systematic justification for our inferences. For justificatory purposes, we can see that the causal relation connecting the target with a model is negligible: the justification for our inferences depends on the theoretical and empirical knowledge supporting the plausibility of our assumptions about the possible similarities between the model and the target. There is no investigation on the causal process that from the target brings us to the representation. With pictures, this is simply different: we are not looking at a way to justify our assumptions about the relation between the picture and the target in that way. Rather, we tell a story about the causal relations instantiated by the mode of production of the picture.

One could doubt of my thesis, particularly when it comes to material models. For example, model organisms like *Drosophila* and the common mouse, commonly used in biological research, undergo considerable selection and genetic intervention in order to be used as representations of other organisms.[150] Yet, the causal, counterfactual relations holding between the target system and the model here is still different from

[148] As it may be already clear to the reader, this is the same line of argument that many philosophers and historians of science have taken, according yo which mechanical scientific pictures are not necessarily more immediate, direct, and therefore objective representations of reality (see e.g., Daston and Galison 2021).

[149] Cf. also Parke (2014), and section 3.3 above.

[150] Cf. Chapter 3 above.

the ones holding between pictures like Figure 4.8. Sure, the interpretation of all models, and the justification for the inferences we draw from them, will depend to a certain extent on the modification we make of them, material or, say, theoretical. Still, there seems to be no causal mechanism that (at least in principle) explains how features *of the target* have been preserved and translated in the representation. The modification scientists perform on an organism to make it a standard model organism does not causally depend on how the designated target system is in the same way as, say, a painter tries to reproduce the skeleton of a human being, or a machine reproduces the distribution of different tissues in an MRI scan.

Anyway, my distinction should be taken as a useful one even if it allows in principle for some grey areas, where it is not clear exactly whether there is a causal relation between representation and target grounding our inferences. If such cases were discovered, the distinction would still be enlightening.

An interesting consequence of my distinction is that it seems to fly in the face of our intuitive ways to categorise representations. According to my proposal, some apparently pictorial, non-linguistic representations, like orbital diagrams in chemistry, protein structures generated by theory-driven predictions (or nowadays by alphaFold), and crystallographic pictures, would instead fall into the category of models.[151] While I acknowledge that my result seems counterintuitive, I want to suggest that it is not in fact problematic but insightful. First, my proposal does not aim at a faithful description of our intuitions. Second, it provides a principled reason to go against our intuitions, namely a difference in the justificatory strategy employed for the inferences we draw from them. This may be the starting point to reconceptualise, or at least integrate, our intuitive taxonomies of representations. If the dense vs. non-dense, or pictorial vs. linguistic distinctions are quite accepted among the participants to the debate,[152] my epistemological distinction is a novel contribution and is fruitful insofar as it cuts across previous dichotomies in a way worth of further exploration.

Perhaps, my distinction between measurement representation and model representation will open new lines of investigation and will prove more useful than the more traditional ones between pictorial and linguistic. Finally, my distinction is not trivial, as it goes against the assumption, shared in STS circles, that the epistemic import of *any* representation, models included, depends on its mode of production in analogous ways. For, if my analysis is correct, we can distinguish interestingly different justificatory roles of the history of production of a representation, and particularly of the causal relation between model and target, in our representational inferences.

[151] I have to thank Jonathan Birch for suggesting me these cases in private conversation.

[152] For a reconstruction of this very articulated debate, see Mößner (2018). Goodman (1976) too identifies syntactic and semantic distinctions as the basis for more general, high-order peculiarities of images with respect to languages. On this, I disagree with him, and my application of DEKI to specific cases indirectly shows the limitations on focusing on syntactic and semantic elements alone.

In this sense, my argument and the consequent novel distinction between measurement representations and model representations enrich the present debate on epistemic representations, particularly visual representations and provide good reason to explore new philosophical taxonomies, which help us better understand the relevant epistemological features of pictures, models, and representation overall.

4.6 Summary of the chapter

So, why do we love pictures? If the first part of my analysis is correct, their epistemic use in science has nothing to do with their similarity with their targets. And if it were because of similarity, we would just love them for the wrong reason. Indeed, in most cases of scientific pictures, similarity does not play any role either in the semantics of those pictures, nor in their epistemic value. This holds for both the relation between the picture and its target, and for how we are supposed to interpret the picture content (or sense).

In contrast, the way we use pictures as epistemic representations can be understood only if we focus on interpretation. Like other instances of representations, we "read" through them by interpreting them as symbols. This interpretation can be further unpacked in terms of denotation, I-functions, exemplification, and keying-up. Denotation "hooks" a picture to its designated target system, the I associates elements of the picture to what will constitute a Z-representation, exemplification will select the salient features of such a Z, and finally the key will translate, when needed, the idealised properties exemplified by the interpreted picture into the ones imputed to the target system.

Each of these steps is dynamic, open to re-consideration and change. Representation is an activity, an interpretive process that works in interaction with the specifics of the given context and the holistic nature of our knowledge.

Moving from the semantic level to the epistemic one, and from representation to accurate and justified representation, one can see that similarity has again no role. Accurate representation simply means that the properties we eventually impute to the target system are in fact possessed by the target. The justification will depend on whether our interpretation was warranted by factors extrinsic to the visual output itself, namely the process with which that visual output was produced. If the causal mechanism that relates the represented target with the visual output secures counterfactual relations between the former and the latter, and these counterfactual relations support our interpretation of the picture, then our inferences from the picture to the target are justified. The level of justification required will depend on what we want to do with our inferences. But whether a certain amount of justification will be sufficient or not is not subjective: given a certain question or goal, there will be more or less justified reasons to trust a specific inference.

In summary, we can trust pictures insofar as we interpret them in the light of

what we know about the counterfactual relations between the state of affairs the picture intends to represent and certain features of the picture itself. So, if we shall love pictures, it should be for this reason: an alignment of our interpretation and our production of the picture. So, we may have ended up loving pictures for the wrong reason. For, like any other instance of epistemic representation, they are not faithful mirrors of reality. Still, we do have good reason to love them, at least sometimes, insofar as we interpret them correctly, that is, in the light of how they were produced. If we do that, we obtain an inferential stability between what is true about them and what is true about their targets.

Chapter 5

Who's afraid of representation?

5.1 The vast anti-representationalist camp

In the previous chapters, I have illustrated DEKI and its application to different types of representation – models, thought experiments, experimental organisms, and pictures. In each case, I tried to draw theoretical consequences that helped us solve or at least improve some of the specific issues at stake for each of these representational types. One more general recurrent theme, which roused prominently in my application of the DEKI account to model organisms and scientific pictures, has been a clear separation between epistemic representation in all its various instances and the concept of similarity. By now, my analysis should have clarified that treating representation in terms of similarity is a dead end. Instead, it is much more promising to focus on interpretative activities of various forms (the I-function and the key) and on the resources we can get from the philosophy of language (denotation and exemplification in particular). DEKI's combination of interpretation and reference seem able to shed light on representational reasoning, and also make sense of what we intuitively call similarity: what counts as a shared property and whether this property is relevant for our reasoning is the result of the interpretation of a certain system as a symbol, which then is taken to be related to its target system via denotation, exemplification, keying-up and imputation.

The contrast between DEKI and the similarity view is of course no news in the literature, but the previous chapters should have provided substantial support to the former against the latter. However, the acknowledgement of the failure to ground representation on substantial or intrinsic properties, like similarity, has led a large group of philosophers of science to scepticism about the very concept of representation in the context of modelling, and surrogative reasoning more generally. This Chapter is devoted to face some most representative examples of this sort of anti-representational scepticism.

Under the label of "anti-representational sceptics", or "sceptics" for short, I group considerably different views. This grouping is nevertheless far from being arbitrary:

as it will become clearer through the Chapter, many of the authors that I call sceptics raise very similar or at least related doubts about representation. In general, all these authors share the idea that the focus on the concept of representation in the context of scientific surrogative reasoning is unhelpful at best or mistaken and misleading at worst.

As a substantial and systematic defence of representation from all these sceptical attacks is still missing in the literature, my work contributes to the current state of the debate not only by illustrating the shared arguments and affinity that all these positions have in common, but also by showing that the general suspicion that they raise against representation can be addressed in an organic, non-patchy fashion.

Within the sceptics' camp, I drew a further classification, with three resulting groups. I call them the success-first view of models, the artifactualist view of models, and the pragmatist-inferentialist view of models. For each of them, I chose one or more most representative views.

Concerning the success-first view, I critically assess Isaac's (2013) attack to representation in section 5.2. Isaac's concerns relate to the fact that representational talk intrinsically implies forms of scientific realism, and an excessive focus on explanatory value of our models to the detriment of other epistemic features of models, and more general of forms of models' success that are not immediately connected to their accuracy. While I take Isaac's scepticism towards representation not to be very troublesome, I believe that the intuitive suspicions of many philosophers of science of representational talk are often related to the points raised by Isaac. It will then be useful to dispel Isaac's criticism against representation and show that his concerns are premature. This will hopefully help other philosophers of science recognise that representation is not as problematically intertwined with other controversial philosophical views (scientific realism, priority of explanation with respect to other forms of epistemic success...) as it may appear at first glance.

Section 5.3 focuses on the artifactualist view of models and surrogative reasoning in general. The artifactualist attempts to loosen the relation between representation and models, proposing to look at the latter as tools or artefacts. The artifactualist camp is broad and diverse: starting from a general artifactualist approach to models conceived as tools put forward by Morrison and Morgan (1999), relevantly artifactualist proposals can be found in Knuuttila (2005, 2011, **Knuuttila:2021**), Currie (2017), and Sanches de Oliveira (2021, 2022). In the section, I first focus on Knuuttila's account because it is currently the most articulated artifactualist view of models developed in the literature so far. Then, I take into consideration Sanches de Oliveira's radical artifactualism, because explicitly distinct from Knuuttila's traditional version in a number of interesting respects. While my analysis will leave some artifactualist proposals aside, my critique of these two artifactualist accounts views should cover much of the artifactualist spectrum.

Finally, in section 5.4, I turn to a family of views that originated from the combi-

nation of pragmatism in epistemology and philosophy of language, and inferentialism about surrogative reasoning. After a brief introduction to the pragmatic roots of this view, I focus on a recent proposal by Khalifa, Millson and Risjord (2022), which tries to give a purely inferentialist account of representation, devoid of any appeal to reference in general and denotation in particular. My analysis of their specific proposal is relevant in two senses. First, it clarifies the extent to which an account based on interpretation and referential relations like DEKI is substantially distinct from an inferentialist approach, while retaining or being compatible with most of the positive consequences of an inferentialist view. The second aim of the section is to highlight the advantages of my account with respect to their proposal.

5.2 Success-first view of models

As a representative of this first subgroup of the anti-representational sceptics, I opted for the view expressed by Isaac in his paper "Models without representation" (2013). There, Isaac expresses doubt about whether representation is the best concept to understand and analyse scientific modelling. He also proposes a pragmatic turn in order to abandon a standard view of models that over-emphasises the importance of representation. Thus, Isaac has both a *pars destruens* and a *pars construens*: in the former, he criticises a representational view of models insofar as it would depend on a problematic ideal of realism and an excessive focus on truth and explanation with respect to success in other dimensions (prediction, production of testable hypotheses, and policy guidance). In his positive proposal, he suggests that in order to evaluate a model we should focus on success, understood not only in representational terms but more broadly. First, I will show that the DEKI account is safe from the threat of realist assumptions, and then show that the success he talks about is still captured by, or at least compatible with the concept of representation I am endorsing in the present analysis.

Isaac argues that the main epistemological problem arising in the context of scientific models is that models usually make many false assumptions, but we keep using them anyway. This puzzle, he suggests,

> rests upon a tenuous assumption, one entrenched in the realist perspective, but unnecessary and unwarranted in the context of modeling. This assumption is that successful science depends upon successful representation. On this view, the justification for modeling as a scientific practice must ultimately rest upon an analysis of how models represent: representation is conceptually prior to success. (Isaac 2013, p. 3612)

This realist view of models is then contrasted with a pragmatic understanding of models, where models' success can be relative to functions other than representational realism: examples of these other functions that models serve are "(i) generating

testable predictions; (ii) offering a policy recommendation; or (iii) demonstrating how an unexpected phenomenon is possible" (*ibid.*).

From this characterisation of the realist view of models, I take it that when Isaac talks about realism, explanatory adequacy, and explanatory success, he refers to the truth of a model's assumptions, and not about the model's results, namely the testable predictive hypotheses we generate from the model about a target system.

A first, clear problem with this line of criticism is the presumed relation between representation and realism. This relation, assumed by Isaac from the start, betrays a very radical understanding of representation, namely what we could call a mirror view of representation. According to this view, a system is a representation of another if the former provides a replica, a copy, or mirror-image of the latter. However, this seems rejected not only by DEKI, but also by most of the other views of representation proposed in the debate. Everybody will agree that the success of a model will not be assessed against the truth of its assumptions. Still, the representation camp will all agree that some, if not all of the success of a model will depend on its success in representing some relevant features of the target system.

Specifically, the DEKI account allows the Z-representation to be very idealised and not at all a veridical description of any real-world target system. What is relevant is that *some* of the properties of the model, derived from our initial assumptions and exemplified by the model, can be actually imputed to the target system. Even these derived properties are sometimes too ideal or distorted, as we have seen in the previous chapters, and that's why we will have to use a key to convert them in properly attributable features of the target system. In fact, a testable hypothesis or a testable prediction is exactly what we expect to obtain in DEKI. So, the first function (i) that Isaac emphasises is paradigmatically representational in the DEKI framework. A similar reasoning applies to Isaac's (iii), that is, that models demonstrate how an unexpected phenomenon is possible. This is just one case of an unexpected property of the target system that is exemplified in the model, or because the model is also a model of a theory, then showing that something is accepted as possible by that theory.

Concerning (ii), it is clear that DEKI does not build any prescriptive force into the concept of representation. First, according to DEKI, a representation can also be a misrepresentation, so it definitely does not imply any guidance for action. Second, even in the case of a model that accurately represents its target system in DEKI's sense, there is no immediate normative valence. Remember that representational accuracy is defined as the case where we impute a property exemplified by a model to a target (once properly translated by a key) and the target actually possesses that property. Then, even if the model accurately predicts how a system functions, this does not give us any *normative* ground for intervention. It is just correct in the description of how the system would behave.

Both these results, however, are clearly positive outcomes of the DEKI account.

We certainly don't want to say that a model, even if representationally inaccurate, can provide guidance for policy and action. And, as concerns models that are representationally accurate, what the source of the justification for our prescriptions is depends at least in part on what we want to do with the target system – in other words, from specific practical and moral considerations, not (only) from our models' results.[153] Certainly, though, representational accuracy is often useful for elaborating a policy or a line of action. Designing policies on the basis of predictions that are overtly false seems a clearly bad strategy for a policy maker.

One can also easily see that, even abandoning Isaac's characterisation of representation in terms of realism, the original puzzle still persists: how is it possible that we get correct predictions from false assumptions? And, more importantly, even if this is the case, how can we justify the validity of such predictions?

Inspired by Levins' (1966) and Friedman's (1953) views about modelling, Isaac suggests turning pragmatist and focus on the model's predictions only. But this does not answer the puzzle, it just restates it: once we have good predictions, we still don't know (a) how it is possible that we obtained them from false assumptions or (b) how to justify them. Therefore, it seems that the puzzle is still there, troublesome as before, and the pragmatic turn does not help us gaining understanding of the source of the success of the model. So, even in the case Isaac's argument worked for a representational account of models, his positive account does not seem to avoid it either.

Here, it is useful to once again recap what the DEKI framework allows us to say about the epistemic value of models. This epistemic value is understood through the lens of representation, understood as denotation, exemplification, keying-up, and imputation. This gives us a strategy to answer to (a). We get from false assumptions to good predictions because the model system, once properly interpreted, exemplifies certain properties that either are also possessed by the target system, or can be systematically imputed to it via a translating key. The value of idealisation is nevertheless recognised, because only thanks to those idealisations we can actually emphasise some previously unnoticed aspects of the target system. Of course, the model as a Z-representation could be useful also to shed light on certain properties of our theories – for example, by highlighting certain functional relations between quantities. This is a non-representational function of a model, and indeed the model functions as a model of a theory. But this is compatible with its distinct function as a model of phenomena, as it should already be clear.

What can we say about (b), namely how we justify our inferences from the model to the target? Here, again, it becomes clear that an account of representation cannot give us a complete answer to this question. How we justify our inferences from a representation cannot be exhausted simply by focusing on the single representational

[153] On the normative source of normative models, see Beck and Jahn (2021).

system in use. But this is true about anything besides tautologies: any fundamental theory is justified on its empirical adequacy, and any interpretation of an observation or measurement depends on more general theoretical understanding. These limits of an account of representation are once again acknowledged and emphasised by the DEKI account.

Of course, we can describe the inferential process from a representation to its target, and we can individuate general common features of this process (exemplification, interpretation, keys). We can further come up with a definition of representational accuracy in terms of factual correctness of our inferences. However, this is all we can do from a general perspective. What remains to be done, for each specific case, is to look at what properties end up being imputed to the target system and how well this set of imputations serve our more general purposes. These purposes can be purely descriptive, explanatory, or predictive, but also practical, for intervention and action guidance. In both cases, representational accuracy as defined above in DEKI's terms seems to be a powerful tool for scientists as a first step before the further required need of justification.

In summary, the DEKI account remains entirely safe from Isaac's concerns about realism, and it furthers gives a plausible story about models' success, instead of taking it as a brute fact.

5.3 The unbearable lightness of artifactualism

In this section, I focus on artifactualist views of modelling and more generally of surrogative reasoning. For the sake of space, I concentrate on two proposals: Knuuttila's standard artifactualist account, which has by now become the main point of reference for the entire artifactualist camp, and Sanches de Oliveira's radical artifactualism, which tries to develop the original artifactualist tenets to their most extreme consequences.

5.3.1 Knuuttila's anti-representationalism

Knuuttila's (2011, **Knuuttila:2021**) work results in both a criticism against what she calls representationalism about models, and a positive proposal of analysing models as epistemic tools or artefacts. While I will mainly focus on the *pars destruens*, arguing that DEKI remains safe from it, I will also show how the main tenets of her artifactualism are either compatible or complementary to the DEKI framework.

5.3.1.1 Representationalism and representation

Let us start with Knuuttila's concerns about what she calls *representationalism* about models (2011). Generally, representationalism is the view that to understand the epistemic use of models in science, one has to understand their relation to specific

target systems, a relation understood in representational terms. We can immediately stop here and clarify that it is quite a matter of consensus among philosophers of science that there exist multiple ways in which a model can be used, representation being just one of them. However, Knuuttila's attack here targets representational accounts of model *tout court*: the very notion of representation, she seems to claim, does not work well with actual use of models in science, even in the standard cases that we have seen in Chapter 1. It will be useful then to show that her criticisms are ineffective when we understand representation in DEKI's sense.

Knuuttila (2011, p. 264) identifies two questions about models and representation that have to be answered for any representationalist account of models: what the relation of representation exactly consists of, and what makes a model an accurate or successful representation of its target (*ibid.*, p. 264).[154] Let us focus on the definitional problem first.

Knuuttila considers two general families of responses to this problem: the strong accounts and the deflationary approaches. The first group tries to reduce the concept of representation to some form of morphism between model and target. These views of representation thus "revert solely to the properties of the model and its supposed target system" (*ibid.*). Furthermore, the sort of property which philosophers mostly focused on are morphisms between mathematical structures instantiated by the model and the target (which we have briefly mentioned above). Knuuttila then proceeds to criticise these reductionist attempts. As DEKI does not reduce representation to similarity or any sort of mathematical morphism, I ignore this line of Knuuttila's attack.

The second family of responses, Knuuttila suggests, take representation to be (at least) triadic, instead of dyadic as the first group, in that not only they consider the model system and the target system, but they also include an agent or user. These views, Knuuttila argues, normally assume that the representational relation cannot be defined if not in terms of the intentions or purposes of the user. The problem with these accounts, Knuuttila argues, is that they end up putting all the burden on the user's intentions and purposes:

> When representation is grounded primarily on the specific goals and representing activity of humans instead of some specific properties of the representative vehicle and the target object, it is deprived of much of its explanatory content: if one opts for a pragmatist deflationary strategy, not much is gained in claiming that models give us knowledge *because* they represent their target objects. (*ibid.*, p. 266)

All the work, so to speak, would be done by the specifications relative to the user, and the concept of representation would become conceptually inert.

[154] The explicit formulation of this distinction can be traced back to Suárez (2010).

First, it is important to immediately notice that DEKI is not deflationary in this extreme sense. Of course, it requires users because of the role played in the account by interpretation, denotation and also exemplification and keys. However, exemplification requires instantiation. And even though sometimes the properties are instantiated by the model once interpreted, and so only I-instantiated, the specification of the I-function and the key is not completely random or arbitrary: both functions must connect actual properties of the carrier with properties of the Z, which will then be imputed to the target system via a key.

Furthermore, in DEKI, the context importantly transcends the single user: particularly in the scientific context, what counts as relevant in the interpretation or what counts as exemplified or not usually depends not on the single user but on the disciplinary field of enquiry, the relevant programmes of research, and the more general theoretical background.

More generally, Knuuttila's point seems to be uncharitable also for other user-based accounts. For even if one acknowledges that users are playing an important role in determining which features of a model are relevant in order to investigate a target system, this does not imply that the concept of representation becomes completely inert. In fact, it only means that representation *involves* such a role played by the user. In other words, Knuuttila's argument seems to assume still the "strong", substantial conception of representation mentioned above. There, representation consists of an objective sharing of property between a model and a target, distinct from the user's intentions, interpretation, and purposes. But this conceptualisation seems to simply deny the spirit of what she calls deflationary approaches, which instead acknowledge from the start the importance of contextual and user-based aspects in their understanding of scientific representation.

Second, Knuuttila has clearly shifted from the definitional problem to the accuracy or success problem, because her argument hinges on the appeal to a question of knowledge and consequent epistemic success of a model as a representation of a target system. However, according to DEKI, misrepresentations are included in the class of representations, so the definition of representation cannot entail (by definition) representational accuracy or success. This is quite uncontroversial, as in everyday life and in science there are many examples of representations that are just mistaken about their targets, even if one focuses only on the features a competent user focused on – for example, models used in the past and no longer deemed as accurate, but that are models nevertheless.

Knuuttila is aware of this distinction between representation and representational accuracy, and she indeed moves on to this latter issue. In her 2011 paper, Knuuttila suggests that the deflationary accounts, while acknowledging the importance of the user, tends to ground models' accuracy again on similarity (e.g., Giere 2010)[155] or

[155] And, we can now add to Knuuttila's analysis, also Weisberg (2013).

mathematical morphism (such as van Fraassen 2008). Here, Knuuttila argues that "it is the user who identifies the relevant respects and the sufficient degrees of similarity" and the relevant isomorphism (2011, p. 266), so again what really counts is the user's interpretation. As a consequence, the concepts of similarity and isomorphism risk here to be trivialised or screened off: what counts is the interpretation, not the similarity or the isomorphism *per se*. The only alternative is some form of radical deflationism à la Suárez (2004), which remains very liberal about representational accuracy and success.

Here again, Knuuttila's attack as it is does not seem troublesome: the account I have proposed also includes a notion of representational accuracy, provided in section 1.6, that does not require any reference to similarity or isomorphism. In my account, a representation is accurate if the properties that we impute to the target are in fact possessed by the target. This, however, is far from being uninformative: if the model turns out to be representationally inaccurate in this sense, we better change something in the model, or change the model all together.

In the end, I agree with Knuuttila that interpretation is the crucial aspect in order to understand representation and, in turn, representational accuracy. Our disagreement concerns the fact that the concept of representation does not help us to shed light and analyse these interpretive activities. The DEKI account, with its conceptual articulation of representation via the concepts of interpretation, *Z*-representation, denotation, exemplification, keying-up, and imputation, seems very useful to disentangle the complexities of both the semantics of a model and its epistemic use for surrogative reasoning.

Here, I am not suggesting that representational accuracy is always necessary for any "successful" use of a model. A model can be successfully used in many ways that are not representational in the sense expressed by the DEKI account – for example, when a model is used to develop know-how to be imported in new contexts. This does not necessarily imply a representational use of the model. In the same way, only because I learn how to use hammer and nails with tables and then I use the same techniques with chairs, the tables in question are here no representation of the chairs in the DEKI sense. For they do not denote them, nor we necessarily impute to chairs some properties exemplified by the tables.

Therefore, I simply want to suggest that representational accuracy be a goal of the representational use of a model. Nonetheless, representational accuracy in the sense I propose, even if not necessary for other forms of model success, will surely be helpful. For example, if the property ascriptions that we make on the basis of the model are correct, this will definitely be useful for the export of know-how.[156]

In general, success is a very general term, which, in accord with Knuuttila (*ibid.*, pp. 264-265), I take to be much broader, and incorporate other epistemic values, than

[156] See above, section 3.4.2

accuracy in my strict sense. As examples of such values, Knuuttila (2011, p. 265) mentions empirical adequacy, truth, reliability, and explanatory power. While I take all these values be at least related to accuracy, I agree with Knuuttila that the last two values, reliability and explanatory power, are not entirely reducible to accuracy. Furthermore, I add to her list external consistency, unificatory and predictive power, and precision.[157] Still, all these values seem perfectly compatible with accuracy, if not also requiring it. So, *pace* Knuuttila, it is manifest how representation, and representational accuracy, while far from providing the entire story about scientific models, constitute a very crucial part of it.

Knuuttila may want to insist that my definition of representational accuracy is still inert, as it is not accompanied by a way to assess it, or in other words, a theory of the justification for our model inferences. This seems in line with her later remark that "[p]ragmatist approaches [among which DEKI] tend to be deflationary in that they do not explain how and why models give us knowledge" (Knuuttila **Knuuttila:2021** p. 4).

In response to this criticism, first, the DEKI account and my definition of accuracy are not vacuous to answer the question of knowledge. By exemplification, models afford us epistemic access to properties that we may not notice in the target systems; and so by studying the former, we can increase our knowledge and understanding of the latter. This will normally take the form of inferences where we produce testable hypotheses about a target on the basis of the results obtained in the model.

Second, there seems to be no universal recipe for the justification of an inference from a surrogative system.[158] Each case will have to be analysed with its specifics and details. More generally, it seems just unfair to ask for a general, one-size-fits-all method for model success, because success is something we discover only a posteriori and is highly context-dependent. If it works, we know. And we will continue to employ the model until we will have reasons to doubt its effectiveness.

If this answer to the question of knowledge from models does not satisfy the artifactualist, so be it. But, as I will try to show now, the artifactualist alternative does not fare better in this regard. Exactly like the representationalist, they will have to get their hands dirty and focus on the specific details of each model. Yet, their conceptualisation of models as tools, if any, is even more vague and general on success than the representational view I have defended so far.

5.3.1.2 The artifactualist alternative

As a positive alternative to representationalism, Knuuttila (2011) proposes to look at models as artefacts that are used as epistemic tools. This perspective, she holds,

[157] This more comprehensive idea of the success of a model is consistent with the concepts of model adequacy developed in other accounts of models (e.g., Alexandrova 2010; Parker 2020).

[158] An analogous argument for what concerns the justification of our reasoning through scientific thought experiments can be found above, section 2.3.4.

allows one to focus on five aspects of modelling that she argues are normally ignored or neglected by philosophers of science:

> (i) the constrained design of models, (ii) non-transparency of the representational means by which they are constructed, (iii) their results-orientedness, (iv) their concrete manipulability and (v) the way their justification is distributed so as to cover both the construction and the use of models. (Knuuttila 2011, p. 267)

Concerning the design of models, Knuuttila emphasises that models are objects in their own right, at least partially independent of their targets, and as such they possess specific affordances and limits that depend on how they were constructed. In this context, Knuuttila also emphasise that, in her account, we can appreciate the value of idealisations: far from being mere shortcomings, they are valuable in their own way because part of a system that as a whole afford us new knowledge and understanding.

However, the DEKI account acknowledges the partial independence of models from their targets, by recognising the properties of both the carrier and the Z, and explicitly distinguishing them conceptually by any potential target system. Moreover, by highlighting the role of exemplification, a user of DEKI will have to clarify how certain properties become highlighted in particular contexts by a certain model system, giving a more detailed account of how idealisations of the models positively partake in this process. Finally, via the proper translating key, idealisations can become truly informative: a map that distorts the distances between places, for example, can become accurate if we employ the proper function that translates the distances on the map into distances on the territory, like what happens in the case of maps built with the Mercator projection system (cf. Nguyen and Frigg 2022).

Knuuttila's second aspect concerns what she calls representational *means*, which she later explains are the result of a combination of a certain representational *medium* (paper and ink, digital pixels, living tissues...) with a certain representational *mode* (which can be linguistic or symbolic, pictorial, or diagrammatic) (2011, p. 269). Representational means and their potential, Knuuttila argues, are often non-transparent to scientists from the start: they have to explore those means and find out what they can be used for. Given its skeletal nature, DEKI does not specify the characteristics of each single carrier or resulting Z: it is the work of a philosopher to illustrate each case in detail. While I concede to Knuuttila the importance to remember the non-transparency of representational means, there is nothing in DEKI that leads us to underestimate this factor: on the contrary, the account gives us all the motivation to appreciate it.

With *result-orientedness*, Knuuttila means the fact that "the starting point [of modelling] is often the output and effects that models are supposed to produce" (*ibid.*, p. 268). This is also in line with her point that models "are constrained in view of answering a pending scientific question" (**Knuuttila:2021** p. 5). But this is

compatible with DEKI, as the logical distinction between denotation, exemplification, keying-up and imputation does not imply any sort of temporal order: as I argued at the end of section 1.7, these four elements develop together, can change in time, and dynamically interact with each other.

Furthermore, according to the representational account defended here, one can further specify that models often help us answer a *specific type* of questions, namely questions about target systems in the world. A problem with Knuuttila's emphasis on the question- and result-orientedness, though, is that it describes the activity of models as quite too constrained. Certainly, we start with some questions about the target, but it may be the case that the model we designed does not answer those questions but opens and possibly solves new puzzles, or frames the problems in different ways: modelling is much more an open-ended process than Knuuttila seems to suggest. The account of epistemic representation presented in this work easily avoids this issue by not building specific questions into the representation activity, letting them arise in a dynamic way.

The concrete manipulability of models has sometimes been neglected by philosophers of science, mostly interested in, and working on abstract or mathematical models. However, the DEKI account is born first of all as an account of material models,[159] then adapted to non-material ones,[160] and nothing in the account seems to privilege the latter type of models with respect to the former. And even in the case of abstract models, there is no problem with saying that models are manipulable, granted that we have a clear sense of what manipulation is in that case (for instance, changes in the relevant assumptions, logical derivation, and so on). But if this is a problem, it is shared by all accounts of scientific modelling, the artifactualist included.

Finally, Knuuttila emphasises how focusing too narrowly on a single model and one specific target system deprive the philosopher of a proper understanding of actual scientific modelling, whereas a model should always be understood as the result of a dynamic activity, related to a wider epistemic framework, and used in synergy with other models. Particularly, our justification for a model inferences must refer to such a broader perspective.

However, these points are in harmony with what I have said above about justification, which is always (at least in part) extrinsic to the single representational system. And more generally, while a concept of representation may seem to require a focus on a specific relation of a model with a target, nothing in the account rules out that the effective epistemic use of a model fundamentally depends on its relation to the rest of our theoretical, empirical, and modelling background. The focus on the single model-target relation is just a practical requirement of philosophical analysis: we start from there and then recognise the web of interdependencies with other bodies of knowledge, methods, and representations. Moreover, while the DEKI account

[159] Cf. Frigg and Nguyen (2016, 2019, and 2020, pp. 159-184).
[160] Cf. for example Frigg and Salis (2017) and Salis Frigg and Nguyen (2020).

still focuses on one model and one target, it allows the simultaneous use of multiple representations for the same target (as we have seen in the case of the pictures of the black hole in section 4.4.2), or even the same model system to denote different targets via different keys (as we have seen with Galileo's thought experiment in section 2.3.2).

Now that we have looked at Knuuttila's positive suggestions, it seems fair to ask in what sense her artifactualist account explains how and why we gain knowledge from models. Saying that models are artefacts designed, built, and manipulated as tools to answer scientific questions seems to simply rephrase the problem. How do they do that, *exactly*? And on what basis? There seems to be no one-size-fits-all answer available for the artifactualist. However, DEKI at least can distinguish *one* particular use of models, namely epistemic representation, from the rest other possible uses of a model in science. This is possible by defining representation in terms of denotation, exemplification, keying-up and imputation. By focussing on these elements, we give a further articulation of the interpretive activity that makes a carrier become a symbol, an epistemic surrogative system for a designated target system. In contrast, in an artifactualist framework, there is no difference between how we learn from a model and how learn, say, by using a hammer. Surely, we learn things by hammering nails in a piece of wood. But this seems simply different from the way we learn from models.

5.3.1.3 The modal dimension of models

Knuuttila (**Knuuttila:2021**) insists that there are cases where an artifactualist approach fares better than a representational one. These are cases of models that are built to answer only modal questions, and thus cannot be taken as representations, because they lack a target in the real world. She presents two case studies. One is Tobin's (1970) ultra-Keynesian model to show that the resulting cycles between monetary policies and inflation were the same exhibited by the completely different monetarist models of Friedman (1961, 1970) and his co-workers. In this way, Tobin showed that an entirely different set of theoretical assumptions would have produced the same observational pattern, discrediting Friedman's hypothesis of a causal priority of monetary decisions over inflation: the correlation (and the temporal cycles) did not imply causation.

The second case study is the construction of so-called repressillators, that is, bio-synthetic oscillators that mimic the oscillations of physical systems but where the circuit is constituted by a set of proteins that repress the protein production of their neighbour gene (Elowitz and Leibler 2000; Gao and Elowitz 2016). The point of these models was to show the biological realisability of intracellular feedback oscillatory mechanisms. For both cases, Knuuttila holds that "it is difficult to point out any actual target systems for either of these models, as they instead appear to address objective possibilities" (**Knuuttila:2021** p. 9).

As a first answer to this modal challenge, there is nothing in the DEKI account

that rules out the utility of targetless models. They are Z-representations, and they can increase our knowledge or understanding in different ways by still exemplifying certain properties. Sometimes they function as intra- or inter-theoretical tools by being models of a certain theory (or a certain set of assumptions), like in Tobin's case of an ultra-Keynesian model of money-inflation cycles. Sometimes, they are projects to build something completely new, like the mathematical and abstract models used to build the repressillators, which are not only a possibility: they constitute an actual case of biological oscillator.

However, one can go further and show that, by exemplifying certain characteristics (money-inflation feedback and synthetic, protein-based oscillations), these models are also representations in DEKI's sense. They can indeed exemplify very general mechanisms that we end up imputing to actual economic cycles and biological oscillators, if at a remarkably high degree of generality. Similarly, in Chapter 2, I have suggested that something similar is the case with Galileo's thought experiment of a ball rolling in a V-shaped cavity. There, the principle of inertia, actually instantiated by Galileo's toy system, can be attributed to any moving object in a counterfactual formulation. Moreover, Nguyen (2020) has argued that many very idealised and simplistic models can be understood in the same way by using a key that translate an actual property of the model in a related susceptibility-statement: the actual target is susceptible to the behaviour factually exemplified by the model.

My two answers – models as Z-representations, and models as exemplifying only possible properties of actual target systems – explain why we are interested in *possibilia* in the first place: we want to know about our theories, but also about what we can do and how the actual things may be. And, of course, on how these two levels interact with each other. Therefore, once the DEKI framework is in place, Knuuttila's examples are far from counting as inexplicable instances of non-representational models. On the contrary, through a fine-grained and articulated account like DEKI we can make sense of such cases of modal models in a fruitful way.

Finally, it is not clear how the artifactualist account fares better in this respect. How do we learn from these targetless models, even if we assume that they should be understood as artefacts? At this point, Knuuttila may point out that these models were meant to answer modal questions, that is, about what is possible and what is not. But this reply is slender at least. In the end, it just states the obvious fact that models are useful for something because they answer a question they aim to answer. However, the account does not say *per se* how and why these models provide an answer to a possibility question. But this seems to be exactly what Knuuttila was requesting the representationalist to do in the first place.

Without going in further details, I gave a multidimensional answer to Knuuttila's modal challenge. According to the DEKI framework, modal-questions models inform us about our theories and techniques, but also about real target systems in the actual world, or at least about some (potential) properties of these targets. In contrast,

until proven otherwise, Knuuttila's artifactualist positive proposal stops far before and, remains far too general about the specific epistemological features of models with respect to all the rest of the vast universe of artefacts and tools that populate our world and our scientific practices.

5.3.2 Radical artifactualism

Within the artifactualist camp, Sanches de Oliveira (2021, 2022) has put forward what is supposed to be the most radical formulation of artifactualism, which would distinguish itself from the other artifactualist accounts of modelling by renouncing any vestige of representationalism. His ideas can be reconstructed as a two-move strategy, consisting of a paper (2021) that focuses on a *pars destruens*, and another paper (2022) that focuses on a *pars construens*. The former describes the inevitable problems that any representationalist account of models must face and cannot solve (2021). The latter argues that an alternative, radical route without representation altogether is viable (2022). Let us look at the two arguments one by one.

5.3.2.1 Representation as a "dead end"

Let us first focus on Sanches de Oliveira's criticism against representation applied to models, which he develops in his 2021 paper, which has the very explicit title "Representationalism Is a Dead End". His attack comes in two steps. First, the concept of representation cannot work for intrinsic difficulties of a representationalist view of models in general. Second, we can well do without representation, so we don't need it in the first place.

Let us see the first line of attack. Sanches de Oliveira argues that any representationalist account of models will not work because of the inevitable tension between two methodological assumptions, which however irremediably informs any representationalist account of models:

> *Ontological component of representationalism (OC)*: models are representations; i.e., models stand in a representational relation to target phenomena.
> *Epistemological component of representationalism (EC)*: modeling is epistemically valuable because of its representational nature; i.e., the representational relation between model and target is what secures the epistemic worth of modeling. (Sanches de Oliveira, 2021, p. 213)

He then argues that the tension can be shown both in the case that we assume representation to be mind-independent relation (for example, simply a matter of similarity or isomorphism) and in the case that it is instead conceived as mind-dependent. Given that basically all current accounts of representation acknowledge that representation is not mind-independent but derives from an interpretation or use on the part of an agent, I will focus on the second case only.

Sanches argues that mind-dependent accounts of representation acknowledge that models are representations (thus endorsing OC), but in order to do that, they inevitably allow the possibility that models misrepresent their targets in relevant ways. This, as we saw already when reviewing Knuuttila's argument, is an expected outcome: from the start, we wanted a definition of representation that allows for inaccurate representation, as we are plenty of models that inaccurately represent their targets. However, Sanches de Oliveira argues, this move turns out to undermine EC: if models can misrepresent, it is not clear how representation helps us understand models' epistemic fruitfulness. Therefore, if we get OC, we end up losing EC. But this is problematic of course, because by renouncing EC we also remove the motivation of theorising about models in representational terms in the first place. This is because, Sanches de Oliveira suggests, what we are fundamentally interested in is models' epistemic success, and representationalism, by sacrificing EC, does not explain this success.

To reply to Sanches de Oliveira, I first want to highlight that his second, epistemological component of representationalism seems unreasonably strong. Particularly, in the case of DEKI, a representational relation between the model and the target does not completely secure the epistemic worth of a model as a representation. First, the model must also be accurate in the sense expressed in section 1.6, that is, the properties imputed to the target on the basis of the model are correct about the model itself.

Second, as we saw talking about Knuuttila, a model's epistemic success is a broader concept than simple representational accuracy. Most importantly, there is no intention to give a once-and-for-all answer to the question of success. This is because success comes in various ways, changes over time, and it concerns here a form of ampliative reasoning.

At the same time, it is clear that DEKI retains the idea that a representational treatment of models is still relevant epistemically. Even though representation does not by itself secure our inferences from a model to a target, it is not the case that representation is a useless concept. If a representation is epistemically successful it is because it succeeds in exemplifying certain properties, and we manage to impute these properties accurately to the target via a proper key, and this imputation results in an epistemic advancement regarding what we want to learn about the target. This of course is not the end of the story: we are not satisfied with lucky guesses in science. We will find reasons to justify our success, and also hypothesise that the strategies employed in successful cases will be successful in other cases. Of course, this justification and reasoning will always be tentative and conjectural, but this is to be expected by a form of ampliative reasoning like the one in play with modelling practices.

So, we can see that representation, while not being a sufficient condition for model success, is a prerequisite of the epistemic use of a model when employed to draw

inferences about some portions of reality. DEKI does not even take representation to be necessary for model success: there are other uses of models, and consequently also successful instances of these uses, which are not representational. We will talk about this in a moment. But this is not a problem for DEKI; the important point is that representation is still important to understand a specific way to use models, which is to make inferences about a certain target system. Particularly, while being a representation is no guarantee of a model's representational accuracy, a representational account allows us to frame the question clearly and puts us in the position to answer clearly whether a model represents accurately or not.

Sanches de Oliveira is aware that his representationalist opponents will make a similar move to the one I have just performed, shifting from representation *simpliciter* to representational accuracy. However, his reply on this point is extremely fast. He says that "the same problem would arise of whether accuracy should be defined in mind-dependent or mind-independent terms" (*ibid.*, p. 225). Sanches de Oliveira does not explain this problem further, but it is not clear in what sense it would be a problem at all. In the account that I have defended so far, representational accuracy is mind-independent, as it simply depends on whether the target in fact possesses the properties we are imputing to it. The pencil in front of me, even if taken to be a model of the universe, is still not an accurate model of the universe: it is actually a bad one, no matter what my purposes are, or how much I believe otherwise. The same holds for success: whether or not an imputation was precise enough, or general enough, or reliable enough for a specific purpose is an objective fact, whatever notion of objectivity one wants to endorse. Of course, success will always be relative to a certain purpose of an agent or an epistemic community, and purposes are of course mind-dependent things. But whether an epistemic result matches successfully those purposes is nevertheless a question independent of our intentions.

Sanches de Oliveira reports further problems deriving from the move from representation to representational accuracy.

First, focusing on accuracy in order to give a sense of models' success "would entail that idealization and abstraction (i.e., inaccurate representation) *cannot* make models better epistemic tools" (*ibid.*, original emphasis). This, he adds, would not only go against an important growing philosophical view,[161] but would also "fly in the face of scientific practice and make the widespread reliance on idealization and abstraction into a mystery – something that, somehow, is more and more used for epistemic purposes but is not epistemically valuable" (*ibid.*).

It should be clear by now that these preoccupations are unjustified once one acknowledges the complex articulation of representation. In representation-as accounts, DEKI included, idealisations and abstractions are essential in order to make some properties more salient than others. They are an incredibly useful part of the process

[161] See Bokulich (2017) and references therein. Cf. also Battermann and Rice (2014), Potochnik (2015, 2017), and Elgin (2017).

through which exemplars exemplify. Sometimes it is only through those idealisations that we can reach an understanding of those properties that then we will impute to the target system. In addition, DEKI has also a way to make use of the exemplified properties that are also idealised themselves: we apply a key to translate them in imputable properties. Again, this doesn't exclude the possibility of being wrong, but it gives a plausible story about why distortions are necessary and even fruitful part of the picture.

More generally, it is clear that Sanches de Oliveira's conception of accuracy is too simplistic: a model can be perfectly accurate, in the DEKI framework, even when replete of distortions. This is because the selectivity of exemplification and the translation carried out via the keys: what we measure accuracy against is the specific set of properties that we end up imputing to the target, not the entirety of the model's assumptions (which may well include distortions with respect to the target system).

Second, Sanches de Oliveira argues that the motivation of representationalism is again lost: "we no longer need an account of scientific models in representational terms because the real epistemological heavy-lifting would be done by the model-target correspondence or informational, two-place relation (which is not representational)" (*ibid.*). But this is simply incorrect, at least in the case of DEKI. This is because representational accuracy does depend on the specific properties exemplified and the keys and in general the interpretive activity involved. Without all this information, the "representation bit", we cannot even start addressing the "accuracy bit". The former, while not sufficient, is not redundant. Once DEKI's conceptual framework is in place, one can look at specific instances of model practice and understand what goes well and what wrong in each case.

In his 2021 paper, Sanches de Oliveira also argues that we can provide a good account of scientific models even without representation. He gives two arguments for such redundancy of the concept of representation.

First, scientists use models all the time, but they do not have nor need a theory of representation. Instead, what seems really important for scientists are all "other factors, such as the disciplinary, theoretical, methodological, technological, erotetic and purposive aspects" (*ibid.*, p. 227). Now, Sanches de Oliveira's point concerning scientists not having a theory of representation is not extraordinarily strong. Only because scientists do not theorise about models in terms of representation, this does not imply that a theory of models as representation cannot be very instructive about modelling practice.[162] Furthermore, the same argument would also apply for a theory of epistemic artefacts, which scientists do not seem to have developed yet. As regards

[162]In a later paper, Sanches de Oliveira himself argues that "if we really want to understand how any practice works, the key thing to understand is *what people do* when participating in that practice, and that it's important to distinguish what people do from what people *say about what they do*" (2022, p. 22, original emphasis).

the factors that scientists would instead focus on (erotetic, disciplinary...), Sanches de Oliveira's list is so general that it would be difficult to deny it. In fact, most representationalists would be happy to grant it. Surely, no account of representation can give us all the details of every single aspect of our modelling practices. However, all the different aspects mentioned by Sanchez seem to find natural space and better conceptual organisation in DEKI's articulation.

The second reason that Sanches de Oliveira gives is simply that one can have non-representational views of models. In the paper that I have followed so far (2021), the author does not advance a specific positive proposal, limiting himself to drawing on and combine existing literature, particularly pragmatism and artifactualism about models. As regards pragmatism, I will deal with it in more details in the next section (5.4). Concerning standard artifactualism, we have already looked at Knuuttila's arguments and showed that there is no strong quarrel between her artifactualism and the DEKI account, and that the latter just performs better. More interesting, then, is to look at Sanches de Oliveira's original and more articulated proposal developed in his 2022 paper, where he delineates what he calls *radical artifactualism*. This would be an account of models that completely reject any sort of representational assumption and provide a viable alternative to representation in order to philosophically understand models.[163] Let us look at this positive proposal more in detail.

5.3.2.2 An Heideggerian perspective on models

The basic tenet of artifactualism, Sanches de Oliveira holds, is that models have to be understood as tools, thus embodying "the conviction that models aren't informative about target phenomena in and of themselves" (2022, p. 18), but because we use them in a certain way. In addition, they cannot be disentangled from an "agential dimension": models "are things that get designed, built and interacted with" (*ibid.*). One can see that all this seems perfectly compatible if not entailed by the DEKI framework: interpretation is an action performed by agents with certain purposes, and the building and design of the model are just part of the process to make an object become an exemplar for certain interesting properties we then want to impute to a target system.

However, Sanches de Oliveira aims at an artifactualist conception of models that "entirely bypass[es] talk of representation" (*ibid.*), thus "without assuming that the relevant epistemology is one in which models act as *sources of information* about some target or other that they represent" (*ibid.*, p. 19, original emphasis). The idea is instead to focus on models just as tools or as artefacts. The intuition is that, when we use hammers, forks, and needles, we learn something not only about those objects,

[163] In this later paper, Sanches de Oliveira also ends up arguing that *all* current versions of artifactualism, Knuuttila's included, presupposes some of the assumptions of representationalism (pp. 14-17). As this line or argument is benefiting me and the representationalists, I do not repeat it here.

but also about the use we can make of them, their relation to us, and their role within the practices of which these objects are a constitutive part (*ibid.*, p. 23). Sanches de Oliveira here appeals to Heidegger's (2001) idea that objects, when conceived as tools (or as "things in-order-to"), possess a basic referentiality to the practice in which the tool is involved and furthermore about us, our goals and purposes, and our relation to those tools and practices.

According to Sanches de Oliveira, the Heideggerian perspective changes the direction of aboutness and reference with respect to the traditional representationalist view. Models are not about their targets, but about the practices they are involved in and about us. It is then wrong to cash out the aboutness of models in terms of their target systems, like all representationalist accounts tend to do (2022, p. 24).

However, it is not clear why this Heideggerian referentiality of tools towards their practice should exclude their referentiality as representations of their targets. In fact, any artefact *qua* tool refers to us and our use of it in a certain practice. This is Heidegger's lesson. But this does not rule out representational reference. A painted portrait is an artefact and as such refers to the practice of painting, artistic movements, the culture and market of artworks, and so on. But it also refers to the portrayed person. So, what stops us taking the specific practice we carry out when we use models (and to which they refer as Heideggerian tools) as a form of *representational* practice, where the tool is not only a tool, but also a symbol that stands for another target system? Sanches de Oliveira himself seems to leave this option open:

> Models are, of course, typically used for guiding how we think and talk about some phenomena, but this does not necessitate analyzing the model itself as being 'about' a given phenomenon (in the sense of being a truth-evaluable description or representation of some 'target') any more than as about ourselves and our projects, goals, and concerns. (*ibid.*)

From this formulation, it seems that both senses of aboutness are legitimate. However, he says that representation is not necessary in order to analyse a model. Certainly, some models are used in a non-representational way, as I have clarified many times. But if the model is used, as Sanches de Oliveira says, to guide how we think and talk about some phenomena, and particularly when we describe and predict and explain certain phenomena, then some reference to those phenomena is already in place, whether we want to talk about it or not. For, if Sanches de Oliveira had to get rid of referentiality towards a target altogether, there would seem to be no reason at all of why we take a model to be informative about some parts of the world: they would just be informative about our practices and us. But this is in overt conflict with what scientists do: exactly from an artifactualist point of view, they design and build and interact with models in order to make inferences about some real system they are actually interested in.

In this sense, Newton conceived the model of the Solar System, Watson and Crick designed the model of DNA, and Phillips and Newlyn built they hydraulic model of a national economy. They wanted to use these models, of course, but use them in a specific way: to exemplify certain properties that could then be imputed to actual target system via a proper translating function. This is one specific way to use a tool. Hammer and nails are not used in this way, for example. And models can be used, as tools, in ways that are not representational in DEKI's sense. For example, when they are used to develop new methodologies and techniques, or generally know-how, which will then exported in other modelling or experimental contexts. I have already illustrated this point in section 3.4.1 by looking at Weber's (2004) point concerning model organisms as instruments for preparative experimentation – with the specific case study of the use of *Drosophila melanogaster* to develop new genetic engineering techniques. I also argued in section 3.4.2 that, while not representational in DEKI's sense, this use of models as instruments of production and exportation of know-how is tightly related to representation.

Sanches de Oliveira' account thus risks to equate all tools, models included, in a way that fails to appreciate interesting and relevant epistemological distinctions between not only models and other tools, but also between different uses we can make of models themselves. One thing is to use a model to make inferences about its target, another thing is to use it to develop know-how that will be exported somewhere else – like when we export our know-how of hammering nails from, say, chairs to tables. Another thing is to use it to test our theories. And so on.

So, while I agree with Sanches de Oliveira that "*learning about something by interacting with something else does not require the one thing to a representation of the other*" (2022, p. 25, original emphasis), there is no reason to then imply that there aren't *different ways to learn* from a system about another, and that models allow this sort of learning in a peculiar way, that is, in a representational manner. Only because we don't learn from hammers representationally, it does not follow that we do sometimes from other artefacts, namely models, (thought) experiments, and pictures.

All in all, once one acknowledges the liberality of DEKI regarding the possible properties imputable to its target system and the high level of conventionality allowed by its interpretation-based framework, one could well start questioning the motivation behind all the anti-representationalist aversion assumed in the radical artifactualist framework. Given that Sanches de Oliveira's own concerns about representationalism, as we have seen just a moment ago, seem misplaced, particularly when applied to DEKI, it is legitimate to ask whether this move towards radical artifactualism, while in principle a viable view among the many allowed in the possibility space, is worth taking in the first place.

5.4 Pragmatic-inferentialism

5.4.1 Main themes of pragmatic anti-representationalism

Representation, particularly in pragmatist circles, sounds like a bad word. Since Dewey and moving to Rorty (1980, 1982), Brandom (1994) and Price (2010),[164] pragmatists have often tried to make a point that the very idea of representation applied to thought and language is in itself problematic, marking an unnecessarily fundamental distinction between what represents (mental states or linguistic objects like words and sentences) and what is represented (the world, or the phenomena). I call this general target of the pragmatists' attacks *Representationalism* with a capital *R*, to distinguish this cluster of ideas about language and thought from the distinct family of theories that analyse models and other forms of surrogative reasoning as instances of epistemic representation. Let us look at some of general characteristics of the pragmatist attack to Representationalism.

First, Representationalism is also often taken to assume a specific idea of representational relation. Namely, representation is here usually understood as a synonym of mimicking, copying, replicating, or mirroring: a (linguistic, mental) representation would then be a copy, or a replica of what it represents. As Godfrey-Smith (2017, p. 155) has argued, John Dewey, usually considered one of the first anti-representationalist voices in the pragmatist sense, "often does not treat 'representation' as his target: he sees 'copy' and to a lesser extent 'correspondence' as guiltier parties".[165] Rorty's (1980) book carries on the word "mirror" in its very title, and in his collection of essays on pragmatism (1982, p. 164) he explicitly groups together "vision, correspondence, mapping, picturing and representation". Price (2010, p.270) seems also to interpret the fundamental core of the pragmatist's anti-representationalism as the "challenge [to] the assumption that language has a single core function, viz., to 'represent how things are'".

Second, and connected to the previous point, the pragmatist sees representationalism as distorting the *goal* of cognition and knowledge and, consequently, of scientific investigation. As Dewey himself puts it:

> The business of thought is not to conform to or reproduce the characters already possessed by objects but to judge them as potentialities of what they become through an indicated operation [...] Knowledge which is merely a reduplication in ideas of what exists already in the world may afford us the satisfaction of a photograph, but that is all. To form ideas whose worth is to be judged by what exists independently of them is not a function that (even if the test could be applied, which seems impossible) goes on within nature or makes any difference there. (Dewey 1929, p. 110)
>
> Any instrument which is to operate effectively in existence must take account

[164] See also Macarthur and Price (2007).
[165] See Godfrey-Smith (2017) for textual support.

> of what exists, from a fountain pen to a self-binding reaper, a locomotive, or an airplane. But "taking account of," paying heed to, is something quite different from literal conformity to what is already in being. It is an adaptation of what previously existed to accomplishment of a purpose. (*ibid.*, p. 165)

Third, and following from the points above, the pragmatist's alternative view to Representationalism is often deflationary in what truth and knowledge consist of. As what is true is not really a matter of mimicking an independent reality, and it instead is at least in part shaped by our lively interaction with that reality, there is not much to explain or analyse about the relation between mental and linguistic representations and the world in order to define what is true and, consequently, what knowledge is. Most of the pragmatist attempts have indeed provided a deflationary analysis of truth,[166] and in section 5.4.3.2 we will look more in detail at Brandom's (1994) account of both truth and reference.

Representationalism is then taken to express a general passivity of our epistemic practices: the aim of science would be to reproduce reality as it is. The pragmatist sees this as a distortion of actual scientific (and more generally of any epistemic) endeavour: knowledge is intrinsically a matter of action and intervention, not only because guided and informed by purposes and goals, but also because the only way to know and understand things is by actively interacting with them. Knowledge, for the pragmatist, is the result of this living interaction with the world, and not simply a passive recording of a phantom, static and immutable reality that we, as visitors in an aquarium, observe and study as though we were completely detached from it.

The pragmatist attack to representationalism can then be illustrated by these three main claims. First, mental and linguistic representations are not copies or mirrors of reality. Second, it is mistaken to take science and in general any epistemic endeavour as a passive recording and reproduction of a static and independent reality, because they are action-based and action-guided. Third, more generally, there is no need to impose, and even less further analyse and explicate, a fundamental distinction between reality and our mental or linguistic representations of reality.

If I am correct in my reconstruction of these three core theses of pragmatic anti-representationalism, it should be clear that the DEKI account is immune to all of them. As regards the first, the previous sections should have clearly illustrated that DEKI's concept of representation does not imply any sort of mirroring, mimicking, copying or even imitating. Denotation, exemplification, and keying-up require an interpretation activity that is completely independent from the concept of similarity, and *a fortiori*, of copying, mimicking, or mirroring. Actually, these interpretative practices are much closer to the idea of inferences put forward within the pragmatist camp.

[166] For an introduction to different theories of truth, see, for instance, Kirkham (1992). For different ways in which this deflationary spirit has been interpreted in the pragmatist camp, see also Price (2010).

Concerning the second issue, DEKI does not deny that epistemic representation, as an instance of scientific endeavour and more generally, is a dynamic and active process. In contrast, as I argued in section 1.7, it suggests the dynamic aspect: the different elements constituting representation are themselves actions and processes in constant feedback with each other and of course sensitive to changes in our more general epistemic framework, constituted by the rest of our theoretical and empirical knowledge. Moreover, DEKI does not only allow, but in fact requires user to actively interact with the representation system (and this can occur theoretically, by the means of an interpretive activity, but also materially, by manipulation). In DEKI, a representation system is of course also sensitive to an interaction with the target: our interpretation of a map may change once we put it in use and start navigating the mapped territory. Yet, the account holds that a direct interaction with the target itself is not a necessary aspect of representation as such: we can conceive and create a map of a territory and later never use it, and we can model a black hole system without having ever observed it. Nevertheless, even though the account does not require an interaction with the target for epistemic purposes, it does not deny the importance of that interaction either. So, there seems to be no grounds to attack the account because it is too static or because it provides a too passive picture of science.

Finally, and more generally, DEKI is not concerned with linguistic or mental representations, but with epistemic representations, that is, surrogative systems used to make inferences about a target system. Therefore, the account does not imply that language and mental representations couldn't be cashed out in pragmatic terms, and it is compatible with deflationary theories of truth and knowledge. However, from the pragmatist recognition that there is no actual relation to be analysed between the world and the propositions (or mental states) describing it, it does not follow that the same can be said about the relation between epistemic representations like models, maps, and pictures and what they represent (say, pendulums, territories, and galaxies). It just follows that our beliefs and propositions *about* those representation systems (models, pictures...) are not really fundamentally distinct from those representation systems themselves, and our beliefs and propositions *about* their target systems (pendulums, galaxies...) are not distinct from those target systems. But the two sets of mental states and linguistic descriptions – one of the representation systems, and one of the target systems, respectively – are still to be kept conceptually distinct.

So, even renouncing Representationalism about mental and linguistic objects, it is not clear that we need to renounce also representationalism in the sense used in this work, or that we directly obtain a pragmatic solution to the epistemic issues of how and why we employ surrogate systems to study other target systems.

5.4.2 From language and thought to representations

While anti-Representationalism, as I have just tried to argue, seems just orthogonal to our discussion of epistemic representation, there is a historic and conceptual con-

nection between pragmatist or deflationary theories in philosophy of language and epistemology, on the one hand, and inferentialist accounts of scientific representation, on the other. The most representative attempt to employ the intuitions from pragmatic, deflationary theories of truth in the context of epistemic representation is the inferentialist view of representation developed by Suárez (2004, 2024).

While other accounts of representation attempt at explaining how and why surrogative reasoning is possible by appealing to some definition of representation, Suárez's inferentialism inverts the order of explanation between representation and surrogative reasoning. A system is an epistemic representation of a target system *only if* it allows informed and competent users to draw inferences about that target system. Another necessary condition is that the representational force of the model points to its target.[167] Suárez's specific proposal and other possible improvements of the inferentialist account (in particular, cf. Contessa 2007 and Díez 2020) have already been discussed at length, particularly by the authors of the DEKI account.[168] Therefore, instead of repeating their arguments there, I point the interested reader to the already existing literature for the details. I just want to flesh out that, while a viable way to deal with representation, the main problem of Suárez account is that it remains terribly minimal. As Contessa puts it:

> On the inferential conception, the user's ability to perform inferences from a vehicle [e.g. a model] to a target seems to be a brute fact, which has no deeper explanation. This makes the connection between epistemic [e.g. scientific] representation and valid surrogative reasoning needlessly obscure and the performance of valid surrogative inferences an activity as mysterious and unfathomable as soothsaying or divination. (2007, p. 61)

Recently, Khalifa, Millson and Risjord (2022) have put forward a new version of the inferentialist account that attempts an explicit importation of the expressivist strategy employed by Brandom (1994) in the philosophy of language. Given that this proposal has not been critically analysed yet,[169] and the authors explicitly present it as an alternative to the DEKI account and other examples of representationalist accounts, I will focus on it in the next section.

[167] Together, these two conditions form a necessary but insufficient set of conditions for a system to be an epistemic representation. Even if they were jointly necessary and sufficient, they would not do any explanatory work (Suárez and Solé 2006, p. 41, and Suárez 2015, p. 46).

[168] Cf. Frigg and Nguyen (2020, Chapter 5), Nguyen and Frigg (2022b, Chapter 3), and references therein.

[169] Suárez (Suárez 2024, p. 279, fn. 7) briefly mentions Khalifa, Millson and Risjord's account and limits himself to suggest that it may fail to address Price's (2008, 2011 and 2013, p. 36) distinction between an internal and an external sense of mental representations.

5.4.3 Smugglers of reference

5.4.3.1 "Thoroughgoing" inferentialism, and the smuggling objection

Let me start by introducing KMR's definition of scientific representation. Given a model M and a target T:

Thoroughgoing inferentialist representation$_{\textbf{(def)}}$: M is a scientific representation of T *iff* M has scientifically justified surrogative consequences that are answers to questions about T. (KMR 2022, p. 265)

The account clearly provides an inferentialist definition of representation, by including the ability to make inferences from models to target as the constitutive element of representation that (allegedly) does not need to be analysed in terms of similarity, isomorphism or denotation.

I want to first highlight two main problems with KMR's very definition, and one clarification about their criticism against DEKI. Then, I will proceed with the illustration of the account.

The first problem, *pace* what KMR say in a footnote (*ibid.*, p. 281, fn. 9), is that their definition seems in principle unable to deal with targetless models. The authors may reply that their focus is (somehow) successful representation, which implies an existing target system. Targetless models would then simply perform a qualitatively different sort of function with respect to scientific representation.

Given the generality of their definition, they could also argue that it may be quite difficult to find examples of targetless models at all: because a model is a scientific representation of a target T *iff* it can be used to draw some justified inferences about T, then one can reconceptualise the target so that a model always has a target system. A model of phlogiston, for example, may just become a (quite poor) representation of combustion, and a model of electromagnetic ether would be an extremely incorrect representation of electromagnetic fields.

However, this generality of the definition of representation bring us to the second problem. For one may worry that KMR's account may become far too broad. To make my argument clear, I will make use of an analogy. Let us consider an everyday tool: a hammer. Now, the shape and the use of a hammer provide the basis to draw justified inferences about the hand that wields it, and in general the limbs and articulations and cognitive system an organism needs in order to use it. It also allows to learn something about what materials are commonly used to build a hammer and thus are available to the manufacturers. Furthermore, if I study the hammer when in use, I can infer many interesting things about some of the properties of the nails and other objects that an agent can hit with a hammer. In a word, the hammer affords many justified inferences about a lot of things concerning its users, the way in which it is used, the practical contexts in which it is used, and even the society that created

it.

However, this type of reasoning can be applied to model representation too, if the only requirement, as in KMR's definition, is the ability to provide justified inferences. Similarly to the hammer, a mathematical model allows inferences not only about its target, but also about our mathematical symbolism; about us as cognitive agents able to perform certain types of mathematical reasoning; about other theorems and theories assumed in the background, and so on. If studied in its specific context of application, one can learn much from the model about many interesting aspects of the relevant epistemic traditions and practices. So, in this general sense, a model can provide justified answers to questions about many things (the users, the practices...) besides the target system.

But then, according to KMR's definition, the model represents, by definition, not only its target system, but all these other things. In the case of the Lotka-Volterra model, we would have the implication that the model does not only represent the fish population in the Adriatic Sea, but also it would represent the users of the model (first of all Volterra himself!), the practices in which it is employed, and it would even represent mathematics itself. But this seems to trivialise the concept of representation altogether.[170]

In other words, if one remains liberal about what kind of inferences one carries out from a model, as KMR seem to do, and no further criterion of identification of the target is provided, one ends up trivialising the concept of representation, because a model would end up representing (according to KMR's definition) far too many things.

Notice that the DEKI account does not face this problem. If one takes denotation as a constitutive element of representation, we have already some way to restrict the representational scope to a single target system, as vaguely defined it may be. Exemplification further restricts the type of inference understood as representational. And indeed, exemplification is not involved in the inferences we can draw from the model to, say, its user, its embedding epistemic practices, and so on.

KMR have to give us some story to distinguish the target from the other objects and systems about which the model could provide some justified inference. Alternatively, they could bite the bullet and conclude that models do represent, beside the target system as traditionally conceived, also all those other sorts of things, about which we can draw some justified inferences from the model. However, KMR would then risk trivialising the concept of representation entirely, as there would be no interesting way to distinguish the inferences from a model to its target (in its original sense) and all other types of inferences that a model afford (e.g., about us as its users, about the epistemic practices in which the model is integrated, and so on). And this consequence may be quite hard to swallow.

[170]The reader may here recognise basically the same objection that I raised against Sanches de Oliveira (2022) in section 5.3.2.2.

These problems of scope of KMR's definition, namely that it is both too narrow and too broad, are serious ones. Here, I am not suggesting that they are inescapable. However, the second issue of excessive broadness seems tightly related to KMR's commitment to avoid any appeal to denotation in the context of representation. If one accepts denotation to play a role in representation, the problem would immediately dissolve. However, KMR do not want to go that way, as we will see below.

Before doing that, there is a clarification to be made about KMR's specific concerns about denotation. KMR claim that "[f]or representationalists, surrogative inference can only be *justified* if the representation relation [similarity, isomorphism, or denotation] holds between model and target" (*ibid.*, p. 265, my emphasis). This way to put it may be confusing, at least as concerns denotation. Let us unpack this issue.

First, appealing to the distinction I have drawn in section 1.6 between derivational and factual correctness, it is not clear whether KMR's concept of justification concerns only the derivational correctness of our inferences (that is, correctness with respect to the model's assumption), or also their factual correctness (that is, the target is in fact how the model predicts it to be). Later, I will show that KMR's position is ambivalent on this issue. For the time being, it is just sufficient to specify that, in DEKI, denotation is simply independent from the derivational correctness of our inferences, and it plays a role for what concerns factual correctness only in a very loose sense. Namely, denotation by itself cannot contribute to the justification of the factual correctness of our inferences. At the same time, we cannot even start imputing properties, and thus making inferences, from a model about a target, if there are no forms of denotational relation between (parts of) the model and (parts of) the target. In this indirect way, denotation is a necessary semantic condition for any subsequent epistemic use of the model, including the justification for factual correctness.

Still, the source of justification for the factual correctness of our inferences, as I specified above, is always at least partially extrinsic to the single representational system under investigation. Such justification is either obtained via a direct observation of the target or on the basis of further theoretical and/or empirical results.

All this to say that, if KMR are concerned simply with the idea that denotation does not provide justification for the factual correctness of our inferences, there is simply no disagreement here between them and me as proponent of the DEKI account: denotation does not provide such a justification. However, I understand their thesis to be stronger: according to them, we should give an account of representation without *any* appeal to denotative relations. This is then the line of argument that I will critically analyse in the rest of the paper.

When it comes to what counts as justification, KMR list five epistemic "entitlements" (*ibid.*, p. 269) that would constitute the inferential pedigree of a model with respect to a target: (i) no defeaters, (ii) derivation, (iii) relevance, (iv) measurement,

and (v) characterisation. Let us now look at these entitlements one by one.

The *no-defeaters* entitlement means simply that there are no facts to support reasonable objections to an inference: until proven otherwise, we have no fact of the matter to counter a certain line of reasoning.

Derivation corresponds to the fact that a certain inference actually follows from the model: an inference has the derivation entitlement if it follows from the model assumptions.

Relevance is the most flexible and general of the entitlements: the model's inferences are answers to questions we are actually interested in about a given target system.

As the name suggests, *measurement* simply indicates some form of measurement aimed at justifying a certain inference for a certain target.

Finally, *characterisation* is a plausible physical (or biological, psychological, economic...) interpretation of the model mathematical formalism or of its material, non-interpreted components.

I take all the entitlements to be either compatible with or already present in DEKI. Relevance seems a general requirement about any epistemic practice: if we are making inferences of any worth, we will have some questions in mind we want to answer. Of course, these questions may change with time depending on the practices we perform: if a certain material model used as a surrogative system does not allow to answer a certain class of questions but it is useful to solve other interrogatives, the relevance entitlement may shift. A question however immediately raises for KMR, as they take for granted that a model will be relevant to answer certain questions about the target. But this relevance should be further justified, because nothing in the model yet specifies that the model's results are relevant for those questions about the target. For, remember, we are not employing any form of reference or denotation. In DEKI, we logically "start" with denotation, thus identifying the relevant target, and then we ask whether what we learn in the model is relevant for our questions about the target.

The derivation entitlement, as I have already noticed, is an integral part of DEKI, and corresponds to part of what I have called derivational correctness of our inferences from a representation. KMR's derivation expresses what follows from the model's assumptions, while the concept of derivational correctness also includes the key and the specification of what is exemplified.

Of course, derivational correctness strongly depends in DEKI on a certain interpretation of the model, which I take to be equivalent to KMR's characterisation entitlement. Specifically, DEKI uses the I-function to interpret the carrier by associating elements and features of the latter to interpreted elements of the model system M. KMR's characterisation entitlement carries out an analogous function by giving an interpretation of the elements of mathematical variables or physical constituents of a material model. To be noticed, however, is that this characterisation does not

in itself grant a solution to the problem I raised just above talking about relevance – I will explain further below.

DEKI does not explicitly talk about measurement, because there are clearly models that were built without any direct measurement. Interestingly, this applies exactly to the case study chosen as paradigm by KMR: Volterra had casual observations about fisheries from his son in law (and maybe one wants to call this measurement), but those observations were exactly what the model had to explain. So, it would be odd to say that the model was justified by such measurements. Also, the model itself never gave rise to any measurements, nor was it ever tested against data.

Here, there is a question whether *every* inferentially relevant aspect of a model has to be connected to measurement. For instance, the fact the planets are spherical in the Newtonian model of the Solar System isn't measured, but it is relevant to exemplify the properties that we have to then impute to the target system. Maybe KMR want to say that any aspect that is inferentially relevant also relates to measurement, at least indirectly, but then the idea becomes almost trivial: everything in a model that is about a target is "somehow" related to measurement. In any case, this disagreement about measurement is not an unsolvable conflict: while DEKI does not recognise measurement as a fundamental dimension of representation, it does not deny its importance either. So, I have no quarrel with KMR on this point.

Similarly, the no-defeaters requirement seems pretty innocuous in its generality. It is simply another justificatory strategy: we go on with our inferences until we have reason to doubt. One can see why in DEKI there is no such requirement, though: in modelling and in representation more generally, we may often have some grounds to doubt about our assumptions, particularly when they involve distortions. In the end, there may always be some defeaters, weak as they may be, which oppose our models' assumptions. Thus, it seems to me that the no-defeaters entitlement may end up being too strong and rule out many potentially useful hypotheses connecting a model to a target. I would then weaken it and require that, if there were objections, they should not be too strong, depending on the context. However, this is not my main issue with KMR's account, and I put this aside.

As I hope these few paragraphs illustrate, there is no conflict between the epistemic entitlements of the expressivist-inferentialist and the DEKI account. One main problem for KMR, in fact, is that these entitlements do not solve the fundamental issue of any inferentialist account of representation, namely that they treat surrogative reasoning as a black-box system (Nguyen and Frigg 2022, pp. 36-45). This is because the five entitlements listed by KMR do not characterise model inferences in any way, as they arguably would hold for many instances of inferences that do not seem representational at all. As an extreme example, consider a deductive inference, say a universal syllogism. Now, syllogisms of this sort seem to satisfy KMR's five entitlements. For such inference should be relevant to a certain question (is Socrates mortal?), it definitely includes a derivation (a deductive one), some of its premisses

may be derived from ot at least related to measurement (that Socrates is a man, that humans seem to be mortal), it works if no defeaters of its premisses emerge, and it may well involve a characterisation of mathematical symbols or variables (S for Socrates...). However, there is no surrogative system involved, and it is hard to say that the syllogism represents Socrates. At least, it seem very different a concept of representation than the one involved in scientific modelling.

In contrast, and perhaps ironically for an inferentialist, DEKI gives a more precise characterisation of representational inferences, connecting them to denotation, interpretation, exemplification and keying-up, and thus distinguishing them from other types of inferences. In this way, representational inferences acquire their own peculiar features in comparison to other sorts of inferences. All in all, while KMR's entitlements seem perfectly legitimate, they do not seem to suffice for a detailed analysis of model inferences specifically. However, that was the aim that KMR were supposed to pursue in the first place.

This is the first problem with KMR's inferentialism: it does not make much progress in improving and detailing the inferential black box of representation. In this, they share the issues signalled for Suárez's account.

The second main problem concerns only KMR's account, insofar as it is not simply inferentialism, but inferentialism *throughout*: they want to reduce scientific representation to inferences and deny any further referential relation between models and targets. Specifically for my proposal, it intends to deny any role to denotation. However, this seems an arduous task for KMR. First, consider the characterisation entitlement, which is explicitly understood in referential terms: a certain variable in the mathematical equations of the model has to be interpreted as a physical quantity. This is a paradigm case of denotation. Of course, it is denotation by stipulation, but still denotation remains.

Second, take relevance: how do we know that the model's results, i.e. what is true *in* the model, are relevant in any sense to answer our questions *about the target*? If there were no characterisation, or interpretation, of the model as in some way relevant to our target system, then there is no way to even think of moving from the model to the target. What we obtain from the model would concern the model alone, and the model's results are not useful as intra- or inter-theoretical tools either, except maybe for very general mathematical or logical theory. This is because a model without characterisation is not endowed with a theoretical interpretation in the first place.

These considerations have a cascade effect on other entitlements. If neither characterisation nor relevance are in place, then it is not clear how (or even whether) the process of measurement would be performed. So, in general, the entire inferential pedigree of our inferences would collapse. Therefore, we need some way to cash out characterisation and relevance without appeal to denotation or in general any sort of referential relation.

About this point, KMR first admit that characterisation "[c]learly... depends

on the resources of language, and thereby on linguistic representation", granting the objection I have just raised. But they also immediately add that "scientific representation is *sui generis*", thus "such reliance on linguistic meaning must be uncontroversial" (*ibid.*, p. 271). The authors do not clarify this point further. However, their claim sounds suspicious: if scientific representation is really *sui generis*, then its reliance on linguistic meaning and denotation should be, if not controversial, at least in need of explication.

Moreover, if all the inferentialist has to say against denotation is that it is not the model that denotes its target, but it is the terms constituting the model description that denote things in the real world, then we still have denotation exactly in the sense of DEKI. Denotation then would be still in place and remain a necessary requirement for scientific representation.

However, I do not think that KMR would be satisfied with this reading. They seem to want to ban denotation *tout court*. Indeed, their entitlement of characterisation does not involve denotation of existing entities, but just a theoretical interpretation of terms and variables, or material elements in case of material models. This theoretical interpretation, though, as I will also show in more detail below, does not imply a relation with real target systems (think of how the theoretical interpretation of models still imply very abstract, idealised systems with respect to actual phenomena).

So it seems that, for the account to function, some sort of denotative or referential notion has to be assumed at least implicitly, even in thoroughgoing inferentialism. But KMR anticipate this objection: they call the general strategy that I have just employed the *smuggling objection* (*ibid.*, pp. 274-275): as I just did, a critic of KMR's proposal would still complain that, even though denotation does not explicitly appear in the account, it still survives in some form, and the entire inferentialist framework would still depend on it.

Let me summarise what we got so far. First, KMR's definition of representation risks to be too narrow (namely unable to deal with targetless models) and at the same time too broad (given the many inferences one can carry out from a model besides the ones concerning its target system). Second, KMR's five entitlements seem either already included within the DEKI account, or if not (like the measurement one) for good reasons. Third, the five entitlements do not seem to shed any further light on the inferentialist black-box of representation, as they do not differ from many other instances of (scientific) inferences. Fourth, because of the characterisation and relevance entitlements, KMR's account seems to still require some implicit notion of reference, namely a relation connecting the terms of the model with their theoretical interpretation and then, in some way, to the target system of interest. This is what KMR's call the smuggling objection: reference and denotation, perhaps implicitly, are still retained in their account.

Let us now look at how KMR try to respond to this fourth problem by employing Brandom's expressivism. Notice though that this is the only objection that their

paper addresses, leaving the other three I have just listed without a response.

5.4.3.2 KMRandomian expressivism

KMR's strategy to reply to the smuggling objection involves an explicit appeal to Brandom's (1994) expressivist theory of truth and reference and an exportation of the same intuitions to the case of scientific representation.

Here, there is no space to deal with Brandom's inferentialist programme in its entirety: this would require a monograph in its own right. Therefore, in the remainder I will solely focus my criticisms on KMR's illustration and application of Brandom's theory – I shall call it *KMRandomian* expressivism.

The expressivist strategy, applied in the philosophy of language, proceeds in two steps. First, it holds that propositional content of a proposition can be explained by its inferential role, which "is captured by the (semantic) commitments and entitlements undertaken by affirming the proposition" (KMR 2022, p. 277). Second, it gives an expressivist account of semantic vocabulary, including "true" and "refers", so that any referential notion is explained via previous uses and inferences.

It is important to notice from the start that Brandom's (original, in this case) inferentialism does not *identify* the semantics of propositions and operators like truth and reference with their inferential role. The semantics of language is then still kept conceptually distinct from its use (inferences included). The point made by the inferentialists in the philosophy of language, as clarified by Murzi and Steinberger (2017) in their introduction to this topic, is meta-semantic and concerns the *explanatory* direction of our theory of language: while the traditional referential theories of language explain the (inferential) use of terms and propositions on the basis of their semantic content, the inferentialist reverses the explanation order, and takes inferences and use in general to explain truth and meaning (and the rest of our referential relations). So, truth and meaning do not disappear, and remain conceptually distinct from the inferences that would explain them.

However, the challenge for KMRandom seems relevantly different in nature. They want to show how one can *reduce* both meaning (first step) and semantic operators like truth and reference (second step) to inferences.[171]

For the sake of KMR argument, I will take the first step for granted and assume that we could in principle reduce the content of a proposition, or its meaning, to the inferences we can derive from it. Let us focus instead on the second step, concerning reduction of truth and linguistic denotation to inferences and use, as this is the move that should allegedly save KMR's inferentialism from the smuggling objection.

[171] There are some, like Rosen (2010), who tend to think of reduction in terms of grounding, or other sorts of metaphysical explanation. If this were the case, my distinction here between explanation and reduction would become much more opaque. However, this association of reduction and explanation is a controversial one to say the least, and the burden of proof for such a parallel lies on KMR's shoulders.

KMRandom suggests that truth and linguistic denotation are operators entirely reducible to "the underlying commitments and entitlements, and to use these to show why it has the expressive function that it does" (2022, p. 277). Let us attempt to illustrate the strategy with the truth operator first. When we say that a proposition p is true, KMRandom suggests, we are not expressing a relation between p and the world: we are simply stating exactly the same propositional content originally expressed by p. The truth operator is just expressing an *endorsement* of that propositional content on the part of someone – for instance, the speaker, or the person that we are talking about (*ibid.*, p. 278). The operator "... is true", then, would just work as a prosentential operator that anaphorically refers to something that has already been said.

Further, KMR emphasise that endorsement "is an activity, not a further proposition" (*ibid.*). This should shed light on the general strategy: when we say that a proposition p is true, we are not adding any new propositional content than what is contained already in p. Given that the propositional content of p can be expressed via use and inferences, we seem to not need any appeal to correspondence to reality or problematic referential assumptions. So far so good.

Things get more complicated when we move from truth to linguistic denotation. Here, let me quote KMR at length:

> Not all semantic vocabulary can be treated as prosentential operators, since not all will have sentences as their anaphoric antecedents. "... refers ....," for example, has indefinite descriptions and deictic expressions as antecedents. In general, Brandom's strategy for semantic vocabulary is to treat such items as "proform" operators that anaphorically inherit content from antecedent expressions and endorse it in some way. (*ibid.*)

Let us assume, for the sake of KMR's argument, that KMRandom's expressivism proves itself to work well with linguistic denotation (that is, the relation from words to objects). Then, we have an explanation, in purely inferential-expressivist terms, of how some terms acquire their meanings. However, linguistic denotation is still there. It has just been explained in terms of endorsements of previous expressions. Instead of referring to "the thing", the word simply refers to previous uses of it.

It seems to me that DEKI is perfectly compatible with different theories of meaning and word-objects relations. Therefore, I do not see KMR's point about linguistic denotation as particularly troubling for my account. What is relevant as a critique of DEKI is what KMR say about the relation of scientific representation between a model and the target. So, let's move to that.

As we saw above, the authors say above that representation is *sui generis* with respect to linguistic reference. But then they add the following:

> Clearly, Brandom's account of semantic vocabulary will not apply directly to scientific representation. However, "M represents T" and "m denotes o" are very

> near cousins of the semantic vocabulary that expresses linguistic representation, and we will argue that their *epistemic* function in scientific discourse is analogous to that of proform operators. (KMR 2022, pp. 278-279)

So, now they say that the two relations are also analogous, being "cousins".[172] The challenge for KMR is then to show exactly how an account of model-target denotation, and not simply linguistic denotation, can be cashed out in expressivist terms, and how this would answer the smuggling objection.

To clarify where the problem lies, let us take again the Lotka-Volterra model of a prey-predator population system as an illustrative example, employed also by KMR as a case study. KMR tell us that, by their characterisation entitlement, we interpret, say, the variables as physical or biological quantities. For example, the letter V is interpreted as a prey population. According to KMRandom's expressivism, this characterisation is obtained on the basis of previous use and inferences. Again, so far, so good.

However, this model prey population, as we have seen above when discussing it in section 3.3.2, does not correspond to any actual prey population. For the prey in the Lotka-Volterra model is an abstract, idealised object. For example, is not constituted by discrete entities but it is continuous, and it grows exponentially unless a predator eats them. How, then, do we move from this abstract population to the actual one?

We can see how this problem is different from linguistic denotation. If I keep using a term, say, V, to stand for an ideal prey population, then V ends up denoting it. I have no quarrel with the expressivist on this. But this certainly does not exhaust representation, because one has still to say how this ideal population stands for an actual population of, say, fish in the Adriatic Sea.

Now, DEKI has a way to do that, namely by denotation. Furthermore, DEKI distinguishes representation from simple denotation, as we have already seen: we still need exemplification (which is a further form of reference), interpretation, the key, and imputation. The key is particularly important in this context, as it specifies in which way we climb down the ladder of abstraction, so to speak, and we impute properties to real target systems on the basis of the abstract Z-representation that we are dealing with.

Let us now see how KMR attempt to solve this problem importing the expressivist strategy. They do it for both representation and denotation, so I deal now with both of these attempts one after the other.

KMR try to argue for their expressivist treatment of representation analogously to KMRandom's expressivist treatment of truth as follows:

> All parties to the debate over scientific representation agree that when [a model] M represents [a target system] T, some surrogative inferences from what M says

[172] KMR's remark on denotation and representation being "cousins" is particularly concerning, given that the entire literature on representation can be seen as an effort to show that representation is not a mere matter of denotation – as, e.g., Callender and Cohen (2006) suggested.

> to conclusions about T are justified. It is *uncontroversial*, then, to take endorsing a set of surrogative inferences as a central function of asserting "M represents T" in scientific contexts. The surrogative inferences endorsed are exactly those justified by the inferential pedigree. "Represents" is thus analogous to "true" and the pro-form operators of semantics: it expresses and endorses an epistemic entitlement that is based on an independent body of epistemic entitlements. And in so doing, it introduces no new epistemic entitlements or semantic content. (KMR 2022, p. 279, my emphasis)

Here, KMR claim that the proposition "M represents T" expresses nothing more than the inferences we take as justified about T from M. However, this is far from being uncontroversial, and it is simply not true that all parties agree on this. The DEKI account, for one, does not agree with such a statement: to represent something is not reducible to the endorsement that a set of inferences are true about the target. DEKI is no exception: other accounts of representation, like the representation-as accounts, but also those based on similarity and isomorphism, do not take justification as a constitutive part of representation at all.

The reason for such resistance among many participants to the debate is that the (correct) interpretation of a model does not imply any relation to a specific target system in the world. Even when characterised/interpreted correctly, theoretical models like the Lotka-Volterra model often describe non-existing systems, and there is no reason to export their results to actual target systems. And this is true for material models as well. Even when the dynamic of a material model is understood correctly under a certain interpretation, there is no step to the target system yet, and no extrapolation to other systems is implied.[173]

In other words, it seems *highly* controversial, to say the least, to simply state that representation is an operator that works like truth, even admitting an inferentialist theory of language – which is a controversial choice in its own terms. Here, KMR seem just to presume that everybody agrees on a generally inferentialist approach to representation, but this should be the conclusion of their argument, not their starting assumption.

For now, I have assumed that KMR's take representation as an endorsement of the justification for our inferences, without any endorsement of their truth about the target system. Indeed, as we have seen, none of the five epistemic endorsements seem to imply such endorsement of truth. However, in a subsequent passage, KMR explicitly suggest that, when we take a model to represent a target, we also endorse the *truth* of our inferences from the former to the latter:

> To claim that M represents T, then, is to endorse just that set of surrogative inferences where propositions derived from M are inferred to be true of T (KMR

[173] Things are here analogous to my considerations about justifications in the previous chapters: what is true about the surrogative system is not necessarily true (let alone justified) about the target system.

2022, p. 280).

However, this last view seems simply incorrect. It is simply not true, semantically, that when we understand a model as a representation, then we are already endorsing that the inferences we are drawing within the model are also correct about the target. When I look at the Ptolemaic model of the Solar System, I understand it as a representation and a model, but I don't endorse what the model tells me about its target. KMR's point does not seem to be true even empirically: scientists use models to produce new testable hypotheses, but it seems at least a stretch to say that the very use of those models implies (by definition) an endorsement of the truth of those hypotheses on the part of the scientists.

In general what KMR suggest seems simply different from what actually happens in everyday scientific practice. There, there is no automatic endorsement, but rather just hypothetical and tentative inferring. So, KMR's analogy between truth and representation is very controversial, to say the least, both semantically and empirically.

KMR seem aware of this problem and try to specify how the two cases of truth and representation also differ:

> Brandom's project aims at understanding the semantic function of "...is true," while we are interested in understanding the capacity of models to represent their targets (and not the semantic function of the sentence "M represents T"). (*ibid.*, p. 279)

Then, KMR's points would not be about semantics of models and representation, but about the grounds for models' epistemic success. However, the exact goals of KMR's analysis now start to appear less clear. For, so far, they seemed fully committed to a semantic analysis of representation. They have given a definition of representation at the beginning, they have set as their main aim a purely inferentialist account of representation, and they tried to reduce the expression "representation" to endorsements of epistemic entitlements. In the second last quoted extract, they have even explicitly drawn an identity between the proposition "M represents T" and an endorsement of some epistemic entitlements about the inferences we draw from the model to the target. But now, KMR claim that they are *not* interested in the semantics of representation at all. As it is quite manifest, the two sets of claims are in overt conflict with each other.

Furthermore, they identify as their opponents those people who analyse the *semantics* of representation in terms of similarity or denotation. But if their account is not concerned with the semantics of representation, then their level of analysis is simply different with respect to all the other accounts of representation, Suárez's and DEKI included, both of which however KMR take as critical targets in their paper. It is not clear anymore, then, if KMR's attack to the accounts of representation that include denotation is meaningful.

In the rest of their paper, *pace* KMR, the authors seem to in fact deal with the semantics of representation:

> We conclude that "M represents T" functions in the scientific context in a way analogous to the semantic proform operators. Like semantic vocabulary, it depends on a number of entitlements and its function is to express endorsement of them. As distinct from the semantic operators, it inherits and *expresses* [my emphasis] epistemic entitlement. Specifically, "M represents T" inherits the entitlements that justify the inference of model derivations (M says that P) to a proposition true or false of a target (C). It thereby expresses entitlement to a set of surrogative inferences. (*ibid.*, p. 282)

Here, KMR conclude something very similar to their "uncontroversial" premise, namely that there is nothing more to representation than the endorsement of their set of entitlements. But, we have already seen that this is both semantically and empirically controversial. Also, this does not help them to overcome the smuggling objection, because models are idealised systems and we still miss a way to bridge them with real target systems.

We can see then that the last step of KMR's argument is fundamentally flawed. They try to avoid the smuggling objection by avoiding the problem of the semantic of "representation" altogether, focusing on the epistemological dimension alone. But this cannot be done. At least implicitly, some form of denotational relations has to be in place in order to even start thinking of the model as the source for our inferences about the target. Specifically, the very epistemic entitlements that KMR want to employ for their expressivist treatment of representation are still implicitly appealing to referential relations.

We have just seen KMR's attempt and failure to give an expressivist account of representation. Towards the end of the section where they defend their account from the smuggling objection, KMR further claim that their expressivist strategy could be applied also to model-target denotation specifically (KMR 2022, p. 282). They try to show that denotative statements like "In the model M the element m denotes the element t in the target system T" can be reduced to an inferential pattern of the sort: "Infer that o exists when m occurs in M" (*ibid.*, p. 283). The same would hold for inferences about relations in the models and other types of inferences.

However, how do we establish such inferential patterns? Here, KMR have to appeal again to their five epistemic entitlements: any "specific instance of these inferences will be justified (or unjustified) by their inferential pedigree" (*ibid.*). But this reduction once again does not help us, because we are still brought back to the epistemic entitlements (characterisation, relevance...), which, we have already seen, are still to be paired with some form of denotative assumptions about the model and the target, which cannot be reduced to simple linguistic denotation and its expressivist treatment.

In fact, this inevitability of some denotative relation between model and target beside mere linguistic meaning was exactly what KMR promised us, in order to accomplish their importation of expressivism to the case of models and thus overcome the smuggling objection. However, KMR do not eventually escape the objection, as their importation of the expressivist strategy remains seriously wanting. For representation is not analogous to truth, at least in the relevant respects of the discussion on models. So, a purely expressivist treatment of representation fails; and the same holds for model denotation, which remains, until proven otherwise, a quintessential aspect of representation.

5.4.3.3 Remarks on denotation and inferentialism

I want to end this paper with a final note concerning KMR's motivations for being suspicious about denotation, and try to show that these motivations are ill-founded. KMR say that while some inferentialist accounts (Hughes 1997; Contessa 2007) and the DEKI account take denotation as a requirement for representation, "such views put the cart before the horse: denotation is consequent on surrogative inference, not the other way around" (2022, p. 282). Furthermore, they say:

> [I]t is a mistake to try to explain surrogative inference in terms of a prior notion of denotation. Doing so misses the function of denotation. To say that the model denotes some object is to say that the model licenses inferences about that object. (2022, p. 284)

KMR's claims indicate that there may be some confusion on what KMR and I take dependency and explanation to be. In DEKI, inferences depend on denotation semantically, and representation is semantically explained by appeal to denotation (and the rest of the constituents of the DEKI account). However, there is no claim in DEKI that denotation does not depend, historically or causally, from the amount of success of our imputations. My point is simply that a representation can be incorrect or unjustified, and so missing the inferential pedigree altogether, and still count as representation. In contrast, it is not logically possible that there are successful representational inferences from a model *about* a target without already presuming some denotative relations.

The sort of dependence and explanation that KMR and DEKI are interested in, then, is simply different. DEKI's emphasis on denotation acknowledges the logical dependence of inferences on referential relations, while KMR's insistence on inferences concerns the practical, historical, causal process through which certain presumed or just hypothetical referential relations become more and more accepted.

I am happy to grant that how certain representations turn out to denote certain target *is* a result of the historical evolution of our imputation attempts (resulting from our surrogative inferences), and whether they worked well in the past or not. The notion of entrenchment employed by Goodman (1983) to solve the new riddle

on induction seems to follow the same strategy: certain predicates are used instead of others because they worked well, or simply have been used, until that moment. In a similar way, certain representations end up denoting certain targets because they effectively exemplify certain properties that we manage to successfully map onto our target systems of interest.

This should reassure KMR, as denotation and exemplification in DEKI are understood as very flexible and dynamic, and they will be shaped according to how well our tentative inferences go. My point is simply that reference, in the specific instances of denotation and exemplification, is a semantic requirement for eventual inferential success: there is no way to make sense of an inference from a representation to its target if we are not already interpreting one as referring to the other in some way.

While possibly reconcilable, I take it that DEKI's focus on the semantic priority more beneficial, as it does not rule out the historical dependence. In contrast, KMR's excessive focus on the historical explanation make them oblivious of the important referential relations implicitly at play in scientific representation.

5.5 Common sceptical themes, and summary of the chapter

From the piecemeal analysis offered above, I would like now to briefly emphasise the commonalities among the three groups of anti-representational sceptics that I have faced in this Chapter: Isaac's (2013) identification of representation with an assumption of realism; the artifactualist views as expressed by Knuuttila (2011,**Knuuttila:2021**) and Sanches de Oliveira (2021, 2022); and, a new inferentialist view proposed by Khalifa, Millson and Risjord (2022), who try to import Brandom's inferentialism in philosophy of language to the context of scientific representation.

First, all of these views offer a stark opposition to views of representation based on similarity or other substantial or intrinsic relations between surrogate systems and their targets. On this point, these views are in agreement with the area of the representationalist camp in which I place myself: representation is at least partially a mind-dependent, context-dependent relation, resulting from the interpretive activity of epistemic agents and the referential relations woven by these agents between representation systems and their targets. Still, representation is complex and does not often reduce to mere stipulation, nor to the individual mental states and attitudes of a single user. According to my view, representation is the result of the interpretive activities of scientists as an epistemic community, and the related referential relations connecting a surrogate system to its target.

However, the anti-representational sceptics whose views I have introduced in this Chapter generalise, and see all the representation talk as not worth the hassle. The artifactualist tries to bypass the problems by analysing surrogate systems as

epistemic tools or artefacts, and the inferentialist attempts the same bypass by taking inferences as the basis for a satisfactory philosophical enquiry of these systems. Furthermore, these two views seem to attain a very similar result in that both views are extremely slender in their general analysis of models and other surrogate systems. Artifactualism can say very little about these systems besides that they are artefacts and tools, and inferentialism has the problem that surrogative reasoning remains fundamentally a black-box, as we cannot further specify some of its general peculiarities and constitutive elements.

This pessimistic attitude towards representation as a dead end has become quite influential, not only among philosophers working on models specifically, but also among philosophers of science more generally. Also, philosophers of representation, and the DEKI authors in particular, have reasonably tended to focus on their direct competitors in the field (similarity views, structuralist accounts...) in order to ascertain the best way to characterise the concept of representation. As a consequence, the critical arguments against representation *tout court*, and the related alternative views of surrogative reasoning carried out by the sceptics, have remained to a large extent unchallenged.

My original contribution to the debate, then, is a systematic critique of anti-representationalism. My response overall shows that these forms of deflationism with respect to model practice are not the only alternative at our disposal. We can retain a representational framework and still say something informative about surrogative reasoning without problematic assumptions about similarity or realism. Namely, we can discern conceptually distinct elements of surrogative reasoning once we offer a representational account of it in the terms of the DEKI account.

In this sense, it is important to stress the philosophically substantial disagreement between my understanding of epistemic representation and all the sceptic views encountered so far. As regards the success-first view offered by Isaac (2013), the conflict is on the very concept of representation, and once one acknowledges that it does not problematically imply assumptions of realism, accuracy, or explanatory power, this subgroup of sceptics should be more willing to reconsider the potential of a representational account of models and surrogate systems.

The contrast of my position with both artifactualism and inferentialism, on the other hand, goes beyond the terminological disagreement and becomes a structural one, concerning not (only) the concepts that we have at our disposal, but a more fundamental view of the philosophy of models and surrogative reasoning *tout court*. Both subgroups take representation to be a non-starter and indicate different routes for an analysis of surrogative reasoning – tool- or artefact-talk, and inference-talk, respectively. Then, my strategy in this Chapter has been two-fold. First, I showed for both of these alternatives that their attacks to my view of representation are either ineffective or misplaced. Second, I emphasised how the positive features of these non-representational views do not match with the advantages provided by DEKI, or

that these positive aspects can be easily integrated because well compatible with a fair reading of the account. Finally, I highlighted the issues of the sceptics' own views. As regards artifactualism, I have argued that this view does not grant us much progress in the understanding of surrogative reasoning with models and the peculiarity of inferences made from a model to its target system. For what concerns expressivist-inferentialism does not manage to escape the smuggling objection, namely the fact that their inferentialist account still forces them to make appeal to representation in referential terms. Overall then, for both groups of proposals, their preferability in comparison to DEKI seems to vanish.

On a final, very general tone, I hope that this extensive and detailed analysis of anti-representational scepticism will help philosophers to finally dispel most of their fear of representation as basically unjustified and provide reasonable ground for optimism concerning a representational analysis of models, as well as thought experiments, pictures, and other instances of surrogative reasoning in the sciences. This should push philosophers of science to move forward, with an encompassing representational framework as a starting point for more local, detailed investigations on different instances of surrogative reasoning.

Chapter 6

Conclusion

Since our "point of departure" in Chapter 1, I have developed a fine-grained analysis of different types of epistemic representations employed in the sciences, moving from the more familiar terrain of models to thought experiments, model organisms and experimental specimens, and mechanically produced pictures. Drawing on this philosophical "field work", I then faced the challenges that arose from a variegated anti-representationalist camp in the philosophy of science and philosophy of modelling.

This thesis was created on the basis of reflections on representation and its application to more circumscribed enquiries. Besides Chapter 1, which introduces and develops the basic conceptual toolkit of my analysis, all the other chapters could be read mostly independently from each other: Chapters 2, 3 and 4 are the result of autonomous investigations, each focusing on more narrow groups of representations, and Chapter 5 is a general response to the multifaceted camp of anti-representational accounts of models and surrogative systems more generally.

This thesis is constituted by sometimes very different variations and motifs that intertwine and build on each other on a common theme, namely a systematic application of a general conceptual framework like DEKI to more specific instances of representation beyond models. The first aim of this Chapter is then to extrapolate the general harmonic features resulting from the different voices of my paper collection. Accordingly, I delineate below a more general, bird-eye-view illustration of the main results of my enquiry, with some principal take-home messages. The second aim is to look forward and sketch a few new promising lines of enquiry for future work.

Let us start with the take-home messages. To begin with, one of the most manifest results of my investigation is that representation is neatly separated from similarity, understood as sharing of properties. In all the cases studied in this work, similarity seems in itself unable to help us understanding the actual epistemic practices of representation put in place by scientists.

Chapter 2 insists on the difference between internal validity of a thought experiment and its external validity. In that context, I further characterised this difference as a difference between what is true in the thought experiment and what is the case

in the target system. Galileo's thought experiment instantiates the law of inertia, whereas very few real physical systems do. In order to infer something about the latter group of systems, we have to first of all acknowledge the differences from the thought experiment's scenario. More importantly, we need a key to apply the result that directly follows from our idealised assumptions to other contexts.

The attack to similarity is even more explicit in the chapters on model organisms and scientific pictures. In the former, I show that much of the disagreement about whether considering model organisms as models or not depend on a general misconception – or at least simplification – of the concept of representation, taken as synonym of analogical reasoning and reducible to similarity. Looking at real-life cases, I have shown that representation does not work this way. For what concerns pictures, I further cast doubts on an understanding of visual representation grounded in a sharing of properties, particularly if we focus on the visual ones.

The same thoughts apply when one moves from similarity to instances of mathematical morphism. For while isomorphisms or other types of mathematical mappings can be more precise and systematic in linking models with their target systems, they still do not provide an answer to the main issues of the similarity view. Particularly, given the vast generality of the concept of morphism and the flexibility of how a mathematical structure can be defined, it remains unclear how one identifies the relevant structure in both the model and the target, and how one chooses among many ways to associate the two structures. In general, I argue that it is not by focusing on similarity (or its special case of mathematical morphism) that one can shed light on the actual representational practices and strategies put in place by scientists, nor on the rational justification for such practices.

The alternative, however, is not necessarily a complete surrender to some forms of anti-representationalism as depicted in Chapter 5. We can give a valuable philosophical account of representation, which is also able to increase our descriptive and normative understanding of actual scientific practices, without appeals to similarity or isomorphism as fundamental concepts. The account of epistemic representation proposed in this work, the DEKI account, focuses on both interpretation and reference: the interpretive dimension is primarily captured by the I-function and the key, while reference enters the scene in two ways, namely as denotation (reference of a symbol towards an object) and exemplification (reference of an object to a property). Interpretive functions and referential relations are here distinguished conceptually, though in practice they participate to the realisation and development of each other. Without interpretation, we would not often be able to even start understanding an object as a symbol, or stand-in, for something else, and without denotation and exemplification it would be arduous to define the interpretive functions connecting the material carrier with the Z-representation, and the latter with the target. This mutual dependence of interpretive functions and referential relations notwithstanding, the account helpfully clarifies their interplay by also identifying them from each other.

Focusing on interpretation and reference, the account is also able to capture a few additional crucial aspects of epistemic representation in science. First, representation clearly depends on epistemic agents, situated in epistemic communities of reference, who in virtue of this situatedness will interpret and referentially relate systems according to specific programmes of study and lines of enquiry. The relevance of epistemic agents as the responsible for the interpretative activities and the creation of referential links, though, does not imply a too narrow view of representation as simply a user-centred phenomenon. Interpretation practices, as well as referential relations, are the products of a collective endeavour of epistemic communities operating within broad conceptual frameworks and research programmes. The nature of the surrounding social structures of representational practices are of course relevant for any reflection on the normative dimension of our surrogative reasoning, particularly when it comes to possible strategies to improve the system of knowledge acquisition from an institutional point of view. Still, the analysis provided so far constitutes an important contribution as it clarifies the internal structure of our inferences from representations and provides philosophers with a useful starting point for broader reflections on the social aspects of scientific representation.

Second, as already implicitly suggested, representation is theorised here as a local and highly context-dependent phenomenon: depending on the context, different properties will be exemplified, thus made epistemically accessible. Depending on the target system of interest, different keys will have to be employed and different interpretations of the carrier required. Of course I acknowledge that the interpretive functions adopted will depend on the specific question we want to answer, and this in turn will depend on the purposes of our enquiry. However, in contrast with much focus on purposes and goals that characterises more pragmatically oriented approaches, I would insist that the interesting aspect highlighted by my investigation is how, once the question has been set, different interpretations, sets of exemplified properties, and keys do fare better or worse. The emphasis, therefore, is not on the set of purposes, but in the strategies to best address those purposes.

Third, it is important to stress that the account is not to be read as implying that representation is a static relation: given its interpretive and referential nature, a representation system and its relation to a target is dynamic, with the different elements of the account (interpretation, denotation, exemplification, keying-up) constantly affecting each other internally, and, extrinsic to the single representation system, the resulting imputations are also affected by further empirical and theoretical advancements. This is a crucial aspect of representation, which otherwise would result in a hypostatic, reified, and ultimately inaccurate concept. Representation is an activity, of which my analysis simply individuates the conceptually distinct elements, which are though in turn themselves dynamic in nature (denotation, exemplification, keys, and imputation). This internal dynamic, however, should be understood in logical, or conceptual terms, and not as a temporal succession: there is no fixed temporal

order of appearance of the four constitutive elements of representation, nor causal hierarchy among them (with perhaps the exception of imputation, which seems to naturally depend on the other elements).

Fourth, the development of the account that I offered in this work, particularly with reference to thought experiments, enlightens the deep connections between interpretation and representation-as, on the one hand, and scientific imagination and idealisation, on the other. By appealing to representation-as in terms of interpretation, we make sense of the critical level of freedom characterising scientific reasoning, and we appreciate the value of idealisations as a path to epistemic selectivity and epistemic accessibility. At the same time, the account is well compatible with the idea that scientific imagination and scientific interpretation are normative in essence: they follow rules, and they can thus be licit or illicit. In this sense, they allow for intersubjective communication and collaboration. Also, one can make this account of imagination and idealisation still compatible with the general factive nature of scientific representation via the concept of a key, namely a function that translates idealised properties into the ones we eventually impute to the target system. Therefore, the apparent paradox of idealisation fades away: distortions are acknowledged as useful tools to make certain properties salient, but they should not be taken as literal descriptions of their target systems; rather, they can be reconciled with a factive understanding of scientific representational inferences by the fact that we use de-idealising keys before imputing properties to the target system.

Another important aspect of representation developed in the present work concerns the nature of justification in the case of inferences from a representation about its target system. This justification, I have argued, is always at least in part extrinsic to the single representational system under investigation. As I have suggested already above, it is important to distinguish two levels of correctness (derivational and factual), which can interplay but are conceptually distinct. When it comes to factual correctness, one has to acknowledge that representational inferences are crucially ampliative in nature, therefore they always remain fundamentally tentative and conjectural. This fact also pushes us to appreciate once again the deeply holistic character of knowledge in general, and science in particular. Even if philosophers can provide useful "models" of knowledge and focus on single cases and practices (for example, in this context, a single thought experiment, a single picture, or a single model organism), one has also to accept that this is just a way to "model", to represent, the actual, very messy scientific practice of scientific representation. A philosophical "model" that, nevertheless, gives us an interesting perspective to better understand both the actual scientific practice and its rational foundations.

One further common theme across the chapters has been the external validity of inferences drawn from representations, as well as the justificatory strategies put in place when external validity was not accessible directly. In Chapter 4, I have further identified a peculiarity of inferences made on the basis of measurement outputs, of

which mechanically produced pictures are a subclass. This type of representational systems, which I will call here for brevity *measurement representations*, allow us to make inferences about their targets in the way illustrated by DEKI. However, the way in which these inferences are justified substantially differs from the way in which one justifies inferences from a model. Inferences drawn from measurement representations are usually justified by appealing to the causal mechanisms that constitute the mode of production of the measurement output, which have to counterfactually ground the relation between the target system and the representation. This constitutes a major difference from the inferences drawn from models (even material ones) and thought experiments. There, the justification comes through the theoretical and empirical knowledge used to support specific assumptions in the model or thought experiment.

If one asked for the reason of this epistemological difference, the best way to answer would be to say that the mode of production is basically built into measurement representations – even in their very interpretation as representations. Asking why they differ from models, then, seems just an ill-posed question: it is not the case that the mode of production is relevant because a surrogate system is a measurement representation, rather a system is a measurement representation when the justification for its inferences depends on the mode of its production. This does not stop us to classify measurement representations, like the picture of the black hole, as epistemic representations, just like models, independently of the way in which we justify their inferences. At the same time, this distinction is useful in order to shed light on a relevant epistemological difference concerning the justificatory strategies put in place for different types of representational systems.

As a consequence, my analysis also allows us to see more clearly the potential and limits of comparing models and other types of representations, like measurement representations. DEKI allows to see the common structure to all these types of representations, while at the same time leaving the necessary space to appreciate the relevant epistemological differences. For example, given the results of my investigation, it would be too hasty to assume that because the mode of production is important for the justification for the inferences from a picture, or a map, then it will be important or relevant for the justification for any representation inference.

A final consequence of the work that I want to emphasise concerns what Frigg and Nguyen (2020, pp. 9-10) called the *question of style* as concerns scientific, and more broadly, epistemic representation. Scientific representation comes in many ways, and different representational systems and strategies are employed even to represent the same phenomena. As Frigg and Nguyen suggest (*ibid.*, p. 180), DEKI seems to imply a multidimensional notion of style, resulting from the interplay of the features of the carrier, the consequent way in which those material properties are interpreted and thus mapped onto the properties of the resulting Z-representation, the way in which properties are exemplified (and thus highlighted), and the keys. To this list, I add the justificatory strategies associated with different types of representations. For

example, as we have seen before, measurement representations exhibit justification strategies crucially characterised by an appeal to the mode of production of the measurement. While I have argued that the way in which we justify the inferences from a representation transcends that single representation, justification still counts as a relevant aspect of that representation once understood as an element of a more complex network of assumptions, beliefs, and inferences. In practice, all the stylistic levels just mentioned interact, but there seems to be no reason to think that they determine each other entirely, nor that one can be simply reduced to another. What seems to be the case is only that there are combinations of stylistic features that seem to work better, depending on the specific purposes at stake.

Now that I listed some of the most worthwhile features and consequences of my work, I want to conclude with a few suggestions for further exploration and development.

First, my work does not indicate that specific types of keys correlate with specific disciplines or traditional categories of surrogate systems (thought experiments, experimental organisms, pictures...). A noteworthy exception is the use in biology of functional identity keys that I conceptualised in Chapter 3. A functional identity key associates elements of a mechanism in the model with elements of a functionally equivalent mechanism in the target. This notion relies on an understanding of function that seems peculiar to biology (and perhaps, engineering), and uncommon to other sciences like physics. Similar thoughts seem to arise with what I called phylogenetic keys, which associate genes of the model organism with homologous genes of the designated target organism. From this work, I can anticipate at least two lines of further investigation. First, my conceptualisation of functional identity keys and phylogenetic keys remained quite general, and further work is required to refine the details of actual mapping between representations and targets in the biological sciences. Second, it may be the case that not all notions of mechanism and function will work well once applied to the question of model keying-up, once all the details are cashed out.

More generally, this specific work on biological representations and their peculiarities may motivate and constitute a template for similar research on other keys that tend to be associated with other disciplines, for example in chemistry, economics, and psychology.

Second, moving beyond the traditional boundaries of science, a question arises on whether the results obtained in this work can further improve our understanding of artistic representation as well. While admitting that artworks can function as epistemic representations like models and thought experiments do in the sciences, one may still wonder whether there is a principled, if gradual, way to distinguish artistic representation from scientific representation. Are there dimensions along which, the more we move along the axis, the more it is likely to define a representation as artistic or, in contrast, scientific?

In his *Ways of Worldmaking*, Goodman (1978) lists five symptoms of the aesthetic: syntactic density, semantic density, relative repleteness, exemplification and complex and indirect reference.

> (1) syntactic density, where the finest differences in certain respects constitute a difference between symbols—for example, an ungraduated mercury thermometer [...]; (2) semantic density, where symbols are provided for things distinguished by the finest differences in certain respects—for example, [...] ordinary English [...]; (3) relative repleteness, where comparatively many aspects of a symbol are significant [...]; (4) exemplification [...]; and finally (5) multiple and complex reference, where a symbol performs several integrated and interacting referential functions, some direct and some mediated by other symbols. (Goodman 1978, pp. 67-68)

From the work carried out in this thesis, it seems that only the fifth feature, that is, multiple and complex reference, actually counts as a genuine symptom of aesthetic representation, in the sense that none of the others seem to track our intuitions on the distinction between science and art. Science makes use of syntactically dense symbol system (think of the autoradiograph in Figure 4.6), and of course scientific representations can be semantically dense as well, given that the phenomena we want to represent are often dense. Repleteness again does not seem to effectively signal the artistic in contrast with the scientific: a model organism or an MRI scan are paradigm examples of scientific representations that are also comparatively more replete than many ordinary aesthetic representations. Finally, exemplification is in this work taken as a crucial aspect of epistemic representation *tout court*.

When he talks about multiple and complex reference, Goodman seems to have in mind two aspects. First, there is what he calls "expression", which he defines in *Languages of Art* as metaphorical exemplification: an object O expresses a property P *iff* (i) O metaphorically possesses P and (ii) O refers to P (1976, p. 95).[174] On this, it has to be noticed that this does not seem peculiar to art either. If we accept the lesson from DEKI that often models are interpreted so that some of their material properties are translated in theoretical ones, there seems to be no principled distinction between metaphorical possession and possession via theoretical interpretation.

Second, this possibility of not only possessing or literally exemplify properties, but also of expressing them, paired with syntactic and semantic density, often allows a symbol to refer to different types of properties simultaneously, and this in turn allow a complex, intertwined net of referential relations to arise.

I take it that this complexity of aesthetic reference is also the fundamental basis of some important developments made by Elgin about disagreement in the arts vs. disagreement in the sciences (2017, pp. 171-182). Elgin suggests that discussion of artworks does not usually aim at finding a consensus among experts: the same

[174]Both Goodman and primarily Elgin (1983, pp. 59-71) have worked on the concept of metaphor and metaphorical reference. I do not delve into these issues for reason of space.

painting by Cézanne, for example, exemplifies certain properties according to a critic, and other properties for another. And these differences about the interpretation of the same painting are simply "irreconcilable": they provide completely conflictual readings of the same artwork. In this respect, aesthetic disagreement seems interestingly different from scientific one, which is usually considered as highly problematic. This, as Elgin herself suggests, is probably due to the fact that scientific research usually aims to provide a basis for coordinated action, both in theoretical and interventionist terms.

Certainly, there may be forms of irreconcilable disagreement among scientists about the correct interpretation of a scientific representation: extreme cases can be found in medical diagnostics, but in general scientists often discuss the correct implications not only of what I have called measurement representations, but also about experimental specimens and theoretical models. The difference is likely to lie not in the fact that there is a disagreement, but how this disagreement is evaluated. In the sciences, the disagreement is seldom seen as an acceptable point of arrival, that is, as almost a goal in itself. A reasonable explanation for this is that science most of the time aims at building knowledge systems useful for cooperation, and in general for collective, coordinated action and intervention. In other words, scientists have to agree to a large extent in order to proceed in their work. In contrast, artistic representations are rarely created in order to provide the basis for collaboration and coordinated action – except perhaps cases where the aim is mere propaganda or the flat outline of political agenda, about which one could though even doubt whether these cases genuinely count as artistic works. In most cases of artistic representation, the fact that we can reasonably interpret the same symbol in different ways is in fact regarded as a positive achievement.

I want to suggest that it would be fruitful to further explore this comparison between epistemic representation as presented in this work and the aesthetic nature of artistic representation.[175] Specifically, I think a way to characterise the distinction we usually draw between artistic and scientific representation concerns the keys involved: the more the urgency to find an unequivocal, precise way to impute properties to a target system, the more systematic and rigid the keys become.

A further set of questions arise from the rising of application of AI technology in the sciences. Scientists already largely employ complex computer algorithms to help them with images (one case is exactly the picture of a black hole, where AI is employed not only to interpolate data but also to conduct robustness analyses on the processes of picture production). Many in the scientific and philosophical literature already suggest that AI could help us interpret pictures more broadly, for example to produce medical diagnoses from MRI and CT imaging. Given the potential of the use of AI for epistemic purposes, a philosophically informed account of how machines

[175]Some work in this direction has already been done in the collection of papers edited by Ivanova and Murphy (2023), devoted to the aesthetics of experiments.

process and "read" pictures is crucial to understand the extent to which these AI outputs can count as evidence in scientific and clinical contexts.

Besides addressing a still largely unexplored territory in the philosophical literature on AI (except for a few exploratory works, e.g., Sullivan 2023), a philosophical investigation on AI analysis of pictures may have important implications as concerns the concept of representation overall. In the philosophical framework that I have adopted in this thesis, interpretation is what really makes something a symbol and, *a fortiori*, a representation of a target system. At the same time, the hope for AI technology is that it will not really need interpretation: machines will be able to detect novel patterns, objective similarities, and eventually make inferences without any higher order interpretation of images. In fact, they would be better than us, allegedly, exactly because they are freed of a pre-imposed interpretation of what they are "looking at". This opens a set of epistemological questions about the use of AI to interpret pictures and other forms of representations. Namely, whether we are going towards a use of representation that is interpretation-free, or at least where the concept of interpretation is not in fact appropriate in the case of AI readings of images. Or, alternatively (and in my view more plausibly), whether this alleged neutrality of AI analysis of representations is just a new form of interpretation.

If I am right, this reveals important epistemological issues on the justification required to rely on AI-based readings of pictures, how the human interpretation will be a necessary part of this activity of justification, and the potentially problematic biases that may have serious downstream effects in the way we use the results of AI-based analyses.

Related to this last point, it becomes more and more evident the crucial role played by experts in the interpretation of our representational systems, particularly in the case of epistemic uncertainty. This uncertainty often concerns how to interpret representations of complex phenomena in context where the stakes are particularly high: models of the climate, diagnostic scans in medicine,[176] and animal models for sentience and consciousness are classic examples of the sort. It would be useful to understand how such instances of representational practice under uncertainty pair up with expert judgement, and what are the best ways to elicit and use such contribution from the experts in the most effective way.[177]

Besides the descriptive intent of improving our philosophical understanding of scientific representational practices, this work will allow to draw out normative advice concerning how experts can properly contribute to the controversies under discussion. Most importantly, it could help us understand the complex structure of evidence

[176] Particularly fascinating in this respect is the interpretive activity of medical scans on the part of expert radiologists (see e.g. Fanti and Lalumera 2023).

[177] The literature on expert judgement and its role in scientific modelling has been increasing in the last years – as an entry point, see Ericsson *et al.* (2018) and Ward *et al.* (2019) – but it remained quite isolated from more traditional approaches to the philosophy of representation. For interesting points of departure, see Hanea *et al.* (2021) and Hanea, Hemming and Nane (2022).

used in context of uncertainty, and where many inferential and interpretive steps are involved. More generally, if interpretation is such a fundamental dimension of representation, who sets the norms and boundaries that distinguish a correct interpretation from an incorrect one? Here again, the role of experts prominently emerges as concerns the standards of correctness for the interpretation of representations and for the justification for the inferences we draw from them.

Bibliography

Abuter, Roberto *et al.* (2018). "Detection of the Gravitational Redshift in the Orbit of the Star S2 near the Galactic Centre Massive Black Hole". *Astronomy and Astrophysics* 615.L15, pp. 1–10.

Alexandrova, Anna (2010). "Adequacy for Purpose: The Best Deal a Model Can Get". *The Modern Schoolman* 87.3/4, pp. 295–301.

Ankeny, Rachel A. (2000). "Fashioning Descriptive Models in Biology: Of Worms and Wiring Diagrams". *Philosophy of Science* 67, S260–S272.

Ankeny, Rachel A. and Sabina Leonelli (2011). "What's So Special about Model Organisms?" *Studies in History and Philosophy of Science Part A* 42.2, pp. 313–323.

— (2020). *Model Organisms*. Cambridge University Press.

Arcangeli, Margherita (2010). "Imagination in Thought Experimentation: Sketching a Cognitive Approach to Thought Experiments". In: *Model-Based Reasoning in Science and Technology*. Ed. by Lorenzo Magnani, Walter Carnielli, and Claudio Pizzi. Berlin-Heidelberg: Springer, pp. 571–587.

— (2018). "The Hidden Links between Real, Thought and Numerical Experiments". *Croatian Journal of Philosophy* 18.1, pp. 3–22.

Barr, Nicholas (2000). "The history of the Phillips Machine". In: *A. W. H. Phillips: Collected works in contemporary perspective*. Ed. by Robert Leeson. Cambridge University Press, pp. 89–114.

Bartha, Paul F.A. (2010). *By Parallel Reasoning*. Oxford: Oxford University Press.

Batterman, Robert W. (2009). "Idealization and Modeling". *Synthese* 169, pp. 427–446.

Batterman, Robert W. and Collin C. Rice (2014). "Minimal model Explanations". *Philosophy of Science* 81.3, pp. 349–376.

Beck, Lukas and Marcel Jahn (2021). "Normative Models and Their Success". *Philosophy of the Social Sciences* 51.2, pp. 123–150.

Bee, Emma *et al.* (2012). "#HazardMap - Real time hazard mapping using social media". URL: https://www.researchgate.net/publication/260336051_HazardMap_-_Real_time_hazard_mapping_using_social_media.

Begg, D. et al. (2014). *Economics* (11th ed.) New York: McGraw-Hill Education.

Beisbart, Claus (2021). "Opacity Thought Through: On the Intransparency of Computer Simulations". *Synthese* 199.3, pp. 11643–11666.

Berkovitz, L. and E. Donnerstein (1982). "External Validity Is More than Skin Deep". *American Psychologist* 37.3, pp. 245–57.

Black, Max (1962). *Models and Metaphors: Studies in Language and Philosophy*. Ithaca (NY): Cornell University Press.

Bogen, James and James Woodward (1988). "Saving the Phenomena". *The Philosophical Review* 97.3, pp. 303–352.

Bohr, Niels (1949). "Discussion with Einstein on Epistemological Problems in Atomic Physics". In: *Albert Einstein: Philosopher-Scientist*. Ed. by Paul A. Schilpp. Evanston (IL): The Library of Living Philosophers, pp. 199–242.

Bokulich, Alisa (2017). "Models and Explanation". In: *Springer Handbook of Model-Based Science*. Ed. by Lorenzo Magnani and Tommaso Bertolotti. Cham: Springer, pp. 103–118.

Bolker, Jessica A. (1995). "Model Systems in Developmental Biology". *BioEssays* 17, pp. 451–455.

Bouman, Katherine L. (2020). "The Inside Story of the First Picture of a Black Hole". *IEEE Spectrum*. URL: https://spectrum.ieee.org/the-inside-story-of-the-first-picture-of-a-black-hole.

Bradley, Seamus and Karim Thébault (2017). "Models on the Move: Migration and Imperialism". *Studies in History and Philosophy of Science Part A* 77, pp. 81–92.

Brandom, Robert (1994). *Making it explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.

Brendel, Elke (2018). "The Argument View: Are Thought Experiments Mere Picturesque Arguments?" In: *The Routledge Companion to Thought Experiments*. Ed. by Michael T. Stuart, Yiftach Fehige, and James R. Brown. London: Routledge, pp. 23–43.

Brentano, Franz (1874). *Psychologie vom empirischen Standpunkt*. Berlin: Duncker and Humblot.

Brown, James R. (1992). "Why Empiricism Won't Work". In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 1992. 2, pp. 271–279.

— (2004). "Why Thought Experiments Transcend Empiricism". In: *Contemporary Debates in Philosophy of Science*. Ed. by Christopher Hitchcock. Hoboken (NJ): Blackwell, pp. 23–43.

— (2011). *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. New York: Routledge.

Brown, James R. and Yiftach Fehige (2022). "Thought Experiments". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/archives/sum2017/entries/thought-experiment/.

Bueno, Otávio, Steven French, and James Ladyman (2012). "Models and Structures: Phenomenological and Partial". *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 43.1, pp. 43–46.

Buxton, Richard B. (2009). *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*. Cambridge University Press.

Buzzoni, Marco (2008). *Thought Experiment in the Natural Sciences. An Operational and Reflective-Transcendental Conception*. Würzburg: Königshausen and Neumann.

— (2018). "Kantian Accounts of Thought Experiments". In: *The Routledge Companion to Thought Experiments*. Ed. by Michael T. Stuart, Yiftach Fehige, and James R. Brown. London: Routledge, pp. 327–341.

Callender, Craig and Jonathan Cohen (2006). "There Is No Special Problem About Scientific Representation". *Synthese* 21.55, pp. 67–85.

Camp, Elisabeth (2007). "Thinking with Maps". *Philosophy of Mind* 21, pp. 145–182.

Campbell, Donald T. (1957). "Factors Relevant to the Validity of Experiments in Social Settings". *Psychological Bulletin* 54, pp. 297–312.

Cartwright, Nancy (1980). "The Truth Doesn't Explain Much". *American Philosophical Quarterly* 17.2, pp. 159–163.
— (1983). *How the Laws of Physics Lie*. Oxford University Press.
— (1999). *The Dappled World*. Cambridge University Press.
— (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press.
— (2010a). "Models: Parables v Fables". In: *Beyond Mimesis and Convention: Representation in Art and Science*. Ed. by Roman Frigg and Matthew Hunter. Dordrecht: Springer, pp. 19–31.
— (2010b). "What Are Randomised Controlled Trials Good For?" *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 147.1, pp. 59–70.
— (2012). "Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps". *Philosophy of Science* 79.5, pp. 973–989.
Cartwright, Nancy and Jeremy Hardie (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press.
Christensen, L. B. and Meg A. Waraczynski (1988). *Experimental Methodology*. Boston: Allyn & Bacon.
Cohen, I. Bernard and Anne (assisted by Julia Budenz) Whitman (1999). *Isaac Newton.* Principia*: Mathematical Principles of Natural Philosophy*. Oakland (CA): University of California Press.
Contessa, Gabriele (2007). "Scientific Representation, Interpretation, and Surrogative Reasoning". *Philosophy of Science* 74.1, pp. 48–68.
Crane, Tim (2013). *The Objects of Thought*. Oxford University Press.
Curiel, Erik (2019). "The Many Definitions of a Black Hole". *Nature Astronomy* 3.1, pp. 27–34.
Currie, Adrian (2017). "From Models-as-Fictions to Models-as-Tools". *Ergo* 4.27, pp. 759–781.
Currie, Adrian and Arnon Levy (2019). "Why Experiments Matter". *Inquiry* 62.9-10, pp. 1066–1090.
Da Costa, Newton C. and Steven French (1990). "The Model-Theoretic Approach to the Philosophy of Science". *Philosophy of Science* 57.2, pp. 248–265.
— (2000). "Models, Theories, and Structures: Thirty Years on". *Philosophy of Science* 67.S3, S116–S127.
Daston, Lorraine and Peter Galison (2021). *Objectivity*. Princeton University Press.
Davidson, Donald (1973). "On the Very Idea of a Conceptual Scheme". *Proceedings and Addresses of the American Philosophical Association* 47, pp. 5–20.
Davies, Stephen (2007). *Philosophical Perspectives on Art*. Oxford: Clarendon Press.
de Chadarevian, Soraya (2004). "Models and the Making of Molecular Biology". In: *Models: The Third Dimension of Science*. Ed. by Soraya de Chadarevian and Nick Hopwood. Stanford University Press, pp. 339–68.
De Regt, Henk W. (2017). *Understanding Scientific Understanding*. Oxford University Press.
Dennett, Daniel C. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Penguin Books.
— (1996). "Intuition Pumps". In: *Third Culture: Beyond the Scientific Revolution*. Ed. by John Brockman. New York: Simon and Schuster, pp. 181–197.
Dewey, John (1929). *The Quest for Certainty*. New York: Minton, Balch & Company.

Díez, Jose A. (2020). "An Ensemble-Plus-Standing-For Account of Scientific Representation: No Need for (Unnecessary) Abstract Objects". In: *Abstract Objects*. Ed. by J. L. Falguera and C. Martínez-Vidal. Cham: Springer, pp. 133–149.

DiFrisco, James, Alan C. Love, and Günter P. Wagner (2020). "Character Identity Mechanisms: A Conceptual Model for Comparative-Mechanistic Biology". *Biology & Philosophy* 35.4, pp. 1–32.

Doboszewski, Juliusz and Jamee Elder (2024). "Robustness and the Event Horizon Telescope: The Case of the First Image of M87". *arXiv preprint arXiv:2401.16323*.

Doyle, Yannick *et al.* (2019). "Non-Factive Understanding: A Statement and Defense". *Journal for General Philosophy of Science* 50, pp. 345–365.

Eckart, Andreas and Reinhard Genzel (1997). "Stellar Proper Motions in the Central 0.1 pc of the Galaxy". *Monthly Notices of the Royal Astronomical Society* 284.3, pp. 576–598.

Einstein, Albert (1905). "Zur Elektrodynamik bewegter Körper". *Annalen der Physik* 7.891. Eng. trans. *On the Electrodynamics of Moving Bodies* in *The Principle of Relativity. A Collection of Original Memoirs on the Special and General Theory of Relativity*, ed. by H.A. Lorentz, A. Einstein, H. Minkowski and H. Lorentz (1952), Mineola, Dover Publications, pp. 35-65.

— (1915). *Sitzungsberichte der königlich preußischen Akademie der Wissenschaften*. Berlin: Deutsche Akademie der Wissenschaften zu Berlin.

— (2002). *Relativity: The Special and the General Theory* (1916), Eng. transl. by Robert W. Lawson. London-New York: Routledge.

Einstein, Albert and Leopold Infeld (1938). *The Evolution of Physics*. Cambridge University Press.

El Skaf, Rawad (2018). "The Function and Limit of Galileo's Falling Bodies Thought Experiment: Absolute Weight, Specific Weight and the Medium's Resistance". *Croatian Journal of Philosophy* 18.52, pp. 37–58.

— (2021). "Probing Theoretical Statements with Thought Experiments". *Synthese*, pp. 1–29.

El Skaf, Rawad and Cyrille Imbert (2013). "Unfolding in the Empirical Sciences: Experiments, Thought Experiments and Computer Simulations". *Synthese* 190.16, pp. 3451–3474.

Elgin, Catherine Z. (1983). *With Reference to Reference*. Indianapolis: Hackett Publishing.

— (1996). *Considered Judgement*. Princeton University Press.

— (2010). "Telling Instances". In: *Beyond Mimesis and Convention*. Ed. by Roman Frigg and Matthew C. Hunter. Dordrecht: Springer, pp. 1–18.

— (2017). *True Enough*. Cambridge (MA): MIT Press.

Elkins, James (1999). *The Domain of Images*. Cornell University Press.

Elliott-Graves, Alkistis and Michael Weisberg (2014). "Idealization". *Philosophy Compass* 9.3, pp. 176–185.

Elowitz, Michael B. and Stanislas Leibler (2000). "A Synthetic Oscillatory Network of Transcriptional Regulators". *Nature* 403, pp. 335–338.

Ericsson, K. Anders *et al.* (2018). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press.

Event Horizon Telescope Collaboration, *et al.* (2019a). "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole". *The Astrophysical Journal Letters* 875.1, p. L1.

Event Horizon Telescope Collaboration, *et al.* (2019b). "First M87 Event Horizon Telescope Results. II. Array and Instrumentation". *The Astrophysical Journal Letters* 875.1, p. L2.
— (2019c). "First M87 Event Horizon Telescope Results. III. Data Processing and Calibration". *The Astrophysical Journal Letters* 875.1, p. L3.
— (2019d). "First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole". *The Astrophysical Journal Letters* 875.1, p. L4.
— (2019e). "First M87 Event Horizon Telescope Results. V. Physical Origin of the Asymmetric Ring". *The Astrophysical Journal Letters* 875.1, p. L5.
— (2019f). "First M87 Event Horizon Telescope Results. VI. The Shadow and Mass of the Central Black Hole". *The Astrophysical Journal Letters* 875.1, p. L6.
Fagan, Melinda B. (2016). "Generative Models: Human Embryonic Stem Cells and Multiple Modeling Relations". *Studies in History and Philosophy of Science Part A* 56, pp. 122–134.
Fanti, Stefano and Elisabetta Lalumera (2023). "The Epistemology of Imaging Procedures and Reporting". *European Journal of Nuclear Medicine and Molecular Imaging* 50.5, pp. 1275–1277.
Feyerabend, Paul (2001). *Conquest of abundance: A tale of abstraction versus the richness of being*. University of Chicago Press.
Fodor, Jerry A. (1975). *The Language of Thought*. New York: Thomas Crowell.
Franklin, A. (1986). *The neglect of experiment*. Cambridge University Press.
Frappier, Mélanie, Letitia Meynell, and James Robert Brown (2013). *Thought Experiments in Philosophy, Science, and the Arts*. New York: Routledge.
French, Steven (2020). "Imagination in Scientific Practice". *European Journal for Philosophy of Science* 10.27.
French, Steven and James Ladyman (1999). "Reinflating the Semantic Approach". *International Studies in the Philosophy of Science* 13, pp. 103–121.
French, Steven and Alice Murphy (2021). "The Value of Surprise in Science". *Erkenntnis* 2021, pp. 1–20.
Friedman, Milton (1953). "The Methodology of Positive Economics". In: *Essays in Positive Economics*. Ed. by Milton Friedman. University of Chicago Press, pp. 3–46.
— (1961). "The Lag in the Effect of Monetary Policy". *Journal of Political Economy* LXIX, pp. 447–466.
— (1970). "Counter-Revolution in Monetary Theory". *Wincott Memorial Lecture*, p. 33.
Friend, Stacie (2020). "The Fictional Character of Scientific Models". In: *The Scientific Imagination*. Ed. by Arnon Levy and Peter Godfrey-Smith. Oxford University Press, pp. 102–127.
Frigg, Roman (2002). "Models and Representation: Why Structures Are Not Enough". *Physics and Economics Project Discussion Paper Series, DP MEAS 25/02, London School of Economics*.
— (2010). "Models and Fiction". *Synthese* 172.2, pp. 251–269.
— (2022). *Models and Theories*. Oxon-New York: Routledge.
Frigg, Roman and Stephan Hartmann (2018). "Models in Science". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/archives/sum2018/entries/models-science/.
Frigg, Roman and James Nguyen (2016). "The Fiction View of Models Reloaded". *The Monist* 99.3, pp. 251–269.

— (2018). "Scientific Representation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/archives/win2018/entries/scientific-representation/.
— (2019). "Of Barrels and Pipes: Representation-as in Art and Science". In: *Thinking about Science and Reflecting on Art: Bringing Aesthetics and the Philosophy of Science Together*. Ed. by Otávio Bueno et al. New York: Routledge, pp. 41–61.
— (2020). *Modelling Nature: An Opinionated Introduction to Scientific Representation*. Dordrecht: Springer.
— (2021). "Mirrors without Warnings". *Synthese* 198.3, pp. 2427–2447.
— (2022). "DEKI and the Mislocation of Justification: A Reply to Millson and Risjord". In: *Scientific Understanding and Representation*. Ed. by Insa Lawler, Kareem Khalifa, and Elay Shech. New York: Routledge, pp. 296–300.
Frigg, Roman and Julian Reiss (2009). "The Philosophy of Simulation: Hot New Issues or Same Old Stew?" *Synthese* 169.3, pp. 593–613.
Frigg, Roman and Fiora Salis (2017). "Of Rabbits and Men". In: *Fictionalism in Philosophy*. Ed. by Bradley Armour-Garb and Frederick Kroon. Oxford University Press, pp. 187–206.
Gadamer, Hans-Georg (1960/2013). *Truth and Method*. London - New York: Bloomsbury.
Galilei, Galileo (1638). *Dialogues Concerning Two New Sciences* (1954), Eng. transl. by Henry Crew and Alfonso de Salvio. New York: Dover Publications.
Gao, Xiaojing and Michael Elowitz (2016). "Precision Timing in a Cell". *Nature* 538, pp. 462–463.
Gebhardt, Karl *et al.* (2011). "The Black Hole Mass in M87 from Gemini/NIFS Adaptive Optics Observations". *The Astrophysical Journal* 729.2, pp. 1–13.
Gendler, Tamar S. (1998). "Galileo and the Indispensability of Scientific Thought Experiment". *The British Journal for the Philosophy of Science* 49.3, pp. 397–424.
— (2004). "Thought Experiments Rethought – and Reperceived". *Philosophy of Science* 71.5, pp. 1154–1163.
Ghez, Andrea M. *et al.* (1998). "High Proper-Motion Stars in the Vicinity of Sagittarius A*: Evidence for a Supermassive Black Hole at the Center of Our Galaxy". *The Astrophysical Journal* 509.2, pp. 678–686.
Giddings, Steven (20017). "Astronomical Tests for Quantum Black Hole Structure". *Nature Astronomy* 1.0067, pp. 1215–1220.
Giere, Ronald N. (2004). "How Models are Used to Represent Reality". *Philosophy of Science* 71.5, pp. 742–752.
— (2010). "An Agent-Based Conception of Models and Scientific Representation". *Synthese* 172.269, pp. 269–281.
Gilbert, Scott F. (2009). "The Adequacy of Model Systems for Evo-Devo: Modeling the Formation of Organisms/Modeling the Formation of Society". In: *Mapping the Future of Biology*. Ed. by Annouk Barberousse, Michel Morange, and Thomas Pradeu. Dordrecht: Springer, pp. 57–68.
Giovannelli, Alessandro (2017). "Nelson Goodman's Aesthetics, in The Stanford Encyclopedia of Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/archives/sum2022/entries/goodman/.
Godfrey-Smith, Peter (2006). "The Strategy of Model-Based Science". *Biology and Philosophy* 21.5, pp. 725–740.

Godfrey-Smith, Peter (2017). "Dewey and Anti-Representationalism". In: *The Oxford Handbook of Dewey*. Ed. by Steven Fesmire. Oxford University Press, pp. 151–172.

— (2020). "Models, Fictions, and Conditionals". In: *The Scientific Imagination*. Ed. by Arnon Levy and Peter Godfrey-Smith. Oxford University Press, pp. 154–177.

Goodman, Nelson (1976). *Languages of Art*. Indianapolis and Cambridge: Hackett.

— (1978). *Ways of Worldmaking*. Hackett: Hackett.

— (1983). *Fact, Fiction, and Forecast*. Harvard University Press.

Grimm, Stephen R., Christoph Baumberger, and Sabine Ammon (2017). *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. New York and Abingdon: Routledge.

Guala, Francesco (2005). *The Methodology of Experimental Economics*. Cambridge University Press.

Hacking, Ian (1993). "Do Thought Experiments Have a Life of Their Own? Comments on James Brown, Nancy Nersessian and David Gooding". In: *Proceedings of the Philosophy of Science Association Conference 1992*. Ed. by David Hull, Mickey Forbes, and Kathleen Okruhlik. Vol. 2. University of Chicago Press, pp. 291–301.

Häggqvist, Sören (2009). "A Model for Thought Experiments". *Canadian Journal of Philosophy* 39.1, pp. 55–76.

— (2013). "Modal Knowledge and the Form of Thought Experiments". In: *The A Priori in Philosophy*. Ed. by Albert Casullo and Joshua C. Thurow. Oxford University Press, pp. 53–68.

Hanea, Anca M. *et al.* (2021). *Expert Judgement in Risk and Decision Analysis*. Cham: Springer.

Hanea, Anca M., Victoria Hemming, and Gabriela F. Nane (2022). "Uncertainty Quantification with Experts: Present Status and Research Needs". *Risk Analysis* 42.2, pp. 254–263.

Harms, Richard J. *et al.* (1994). "HST FOS Spectroscopy of M87: Evidence for a Disk of Ionized Gas around a Massive Black Hole". *Astrophysical Journal, Part 2-Letter* 435.1, pp. L35–L38.

Harris, Margherita (2021). "Conceptualizing Uncertainty: The IPCC, Model Robustness and the Weight of Evidence". Ph.D. Thesis. London School of Economics and Political Science.

Hartmann, Stephan (1995). "Models as a Tool for Theory Construction: Some Strategies of Preliminary Physics". In: *Theories and Models in Scientific Processes (Poznan Studies in the Philosophy of Science and the Humanities 44)*. Ed. by I. Niiniluoto W. E. Herfel W. Krajewski and R. Wojcicki. Amsterdam and Atlanta: Rodopi, pp. 49–67.

Hawking, Stephen W. (1976). "Breakdown of Predictability in Gravitational Collapse". *Physical Review D* 14 (10), pp. 2460–2473.

Heidegger, Martin (2001). *Being and Time* (1927). Oxford: Blackwell Publishing Ltd.

Held, Carsten *et al.* (2013). *Mental Models and the Mind: Current Developments in Cognitive Psychology, Neuroscience and Philosophy of Mind*. Amsterdam: Elsevier.

Herfeld, Catherine (2024). "Model Transfer in Science". In: *The Routledge Handbook of Philosophy of Scientific Modeling*. Ed. by Tarja Knuuttila, Natalia Carrillo, and Rami Koskinen. New York: Routledge, pp. 105–121.

Hesse, Mary (1963). *Models and Analogies in Science*. London: Sheed and Ward.

Hoefer, Carl and Alexander Krauss (2021). "Measures of Effectiveness in Medical Research: Reporting both Absolute and Relative Measures". *Studies in history and philosophy of science* 88, pp. 280–283.

Hudson, Robert (2014). *Seeing Things: The Philosophy of Reliable Observation.* Oxford University Press.

Huettel Scott A. and, Allen W. and Gregory McCarthy (2004). *Functional Magnetic Resonance Imaging.* Sunderland: Sinauer Associates.

Hughes, R.I.G. (1997). "Models and Representation". *Philosophy of Science* 64, S325–336.

Humphreys, Paul (2009). "The Philosophical Novelty of Computer Simulation Methods". *Synthese* 169.3, pp. 615–626.

Huneman, Philippe (Ed.) (2013). *Functions: Selection and Mechanisms.* Dordrecht: Springer.

Hyman, John (2006). *The Objective Eye: Color, Form, and Reality in the Theory of Art.* University of Chicago Press.

— (2012). "Depiction". *Royal Institute of Philosophy Supplement* 71, pp. 129–150.

Illari, Phyllis (2019). "Mechanisms, Models and Laws in Understanding Supernovae". *Journal for General Philosophy of Science* 50.1, pp. 63–84.

Intemann, Kristen (2010). "Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties". *Intergovernmental Panel on Climate Change.*

Isaac, Alistair (2013). "Modeling without Representation". *Synthese* 190.16, pp. 3611–3623.

Ivanova, Milena and Alice Murphy (2023). *The Aesthetics of Scientific Experiments.* New York - Abingdon: Routledge.

Jacquette, Dale (2004). "Kacquette, Dale". In: *Brentano's Concept of Intentionality.* Ed. by Id. Vol. 10. Cambridge University Press, pp. 98–130.

Jebeile, Julie and Ashley Kennedy (2015). "Explaining with Models: the Role of Idealization". *International Studies in the Philosophy of Science* 29.4, pp. 383–392.

Kendrew, John C. *et al.* (1958). "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis". *Nature* 181, pp. 662–66.

Khalifa, Kareem (2017). *Understanding, Explanation, and Scientific Knowledge.* Cambridge University Press.

Khalifa, Kareem, Jared Millson, and Mark Risjord (2022). "Scientific Representation: An Inferentialist-Expressivist Manifesto". *Philosophical Topics* 50.1, pp. 263–292.

Kirkham, Richard L. (1992). *Theories of Truth: A Critical Introduction.* Cambridge (MA): MIT Press.

Klee, Paul (1920). "Schoepferische Konfession". In: *Tribüne der Kunst und der Zeit. Eine Schriftensammlung.* Ed. by Kasimir Edschmid. Vol. XIII. Berlin: Erich ReißVerlag, pp. 28–40.

Klein, Colin (2010). "Images Are Not the Evidence in Neuroimaging". *The British Journal for the Philosophy of Science* 61.2, pp. 265–278.

Knott, Cargill G. (1911). *Life and Scientific Work of Peter Guthrie Tait.* Vol. 1. Cambridge University Press.

Knuuttila, Tarja (2005). "Models as Epistemic Artefacts: Toward a Non-Representationalist Account of Scientific Representation". Ph.D. Thesis. University of Helsinki.

Knuuttila, Tarja (2011). "Modelling and Representing: An Artefactual Approach to Model-Based Representation". *Studies in History and Philosophy of Science Part A* 42.2, pp. 262–271.

Kohler, Robert E. (1991). "Systems of Production: *Drosophila*, Neurospora, and Biochemical Genetics". *Historical Studies in the Physical and Biological Sciences* 22.1, pp. 87–130.

— (1993). "*Drosophila*: A Life in the Laboratory". *Journal of the History of Biology* 26.2, pp. 281–310.

Kormendy, John and Douglas Richstone (1995). "Inward Bound - The Search for Supermassive Black Holes in Galactic Nuclei". *Annual Review of Astronomy and Astrophysics* 33, pp. 581–624.

Kostić, Daniel (2019). "Minimal Structure Explanations, Scientific Understanding and Explanatory Depth". *Perspectives on Science* 27.1, pp. 48–67.

Krämer, Sybille (2013). "The Mind's Eye': Visualizing the Non-visual and the 'Epistemology of the Line". *Ontos Verlag: Publications of the Austrian Ludwig Wittgenstein Society-New Series* 17, pp. 275–293.

Kuhn, Thomas S. (1977). "A Function for Thought Experiments". In: *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Ed. by Id. University of Chicago Press, pp. 240–265.

Le Bihan, Soazig (2021). "Partial Truth versus Felicitous Falsehoods". *Synthese* 198.6, pp. 5415–5436.

Leonelli, Sabina (2007). "Growing Weed, Producing Knowledge: An Epistemic History of *Arabidopsis thaliana*". *History and Philosophy of the Life Sciences* 29, pp. 193–223.

Levins, Richard (1966). "The Strategy of Model Building in Population Biology". *American Scientist* 54.4, pp. 421–431.

— (1968). *Evolution in Changing Environments: Some Theoretical Explorations*. Princeton University Press.

Levy, Arnon (2012). "Models, Fictions, and Realism: Two Packages". *Philosophy of Science* 79.5, pp. 738–748.

— (2015). "Modeling without Models". *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 172.3, pp. 781–798.

Levy, Arnon and Adrian Currie (2015). "Model Organisms are Not (Theoretical) Models". *The British Journal for the Philosophy of Science* 66.2, pp. 327–348.

Levy, Arnon and Peter Godfrey-Smith (2020). *The Scientific Imagination*. Oxford University Press.

Lynden-Bell, Donald (1969). "Galactic Nuclei as Collapsed Old Quasars". *Nature* 223, pp. 690–694.

Macarthur, David and Huw Price (2007). "Pragmatism, Quasi-Realism and the Global Challenge". In: *The New Pragmatists*. Ed. by Cheryl Misak. Oxford: Clarendon Press, pp. 91–121.

Mach, Ernst (1896). "Über Gedankenexperimente". *Zeitschrift für Physikalische Chemie Unterrichten* 10. pp.1-5. Eng. transl. by W.O. Price and S. Krimsky, *On Thought Experiments* (1973), *Philosophical Forum* 4, 3, pp. 446–457.

— (1919). *The Science of Mechanics*. Eng. transl. by Thomas J. MacCormack. London-Chicago: The Open Court Publishing.

Markie, Peter (2021). "Rationalism vs. Empiricism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/archives/fall2017/entries/rationalism-empiricism/.

Massimi, Michela (2022). *Perspectival Realism*. Oxford University Press.
Maxwell, James C. (1871). *The Theory of Heat*. London: Longmans Green and Co..
Mey, Tim De (2003). "The Dual Nature View of Thought Experiments". *Philosophica* 72, pp. 61–78.
Meynell, Letitia (2013). "Parsing Pictures: on Analyzing the Content of Images in Science". *The Knowledge Engineering Review* 28.3, pp. 327–345.
— (2014). "Imagination and Insight: A New Account of the Content of Thought Experiments". *Synthese* 191.17, pp. 4149–4168.
Miščević, Nenad (1992). "Mental Models and Thought Experiments". *International Studies in the Philosophy of Science* 6.3, pp. 215–226.
Miyoshi, Makoto *et al.* (1995). "Evidence for a Black Hole from High Rotation Velocities in a Sub-Parsec Region of NGC4258". *Nature* 373.6510, pp. 127–29.
Moretti, Charlotte, Isabelle Stévant, and Yad Ghavi-Helm (2020). "3D Genome Organisation in *Drosophila*". *Briefings in Functional Genomics* 2, pp. 92–100.
Morrison, Margaret and Mary S. Morgan (1999). "Models as Mediating Instruments". In: *Models as Mediators: Perspectives on Natural and Social Science*. Ed. by Mary S. Morgan and Margaret Morrison. Cambridge University Press, pp. 10–37.
Mößner, Nicola (2018). *Visual representations in science: concept and epistemology*. London: Routledgeo.
Muhr, Paula (2023). "The "Cartographic Impulse" and Its Epistemic Gains in the Process of Iteratively Mapping M87's Black Hole". *Media & Environment* 5.1, pp. 341–357.
Murphy, Alice M. L. (2020). "Thought Experiments and the Scientific Imagination". Ph.D. Dissertation. University of Leeds.
— (2022). "Imagination in Science". *Philosophy Compass* 17.6, e12836.
Murzi, Julien and Florian Steinberger (2017). "Inferentialism". In: *A Companion to the Philosophy of Language (2nd ed., Vol. 1)*. Ed. by Bob Hale, Wright Crispin, and Alexander Miller. Chichester: Blackwell-Wiley, pp. 197–326.
Nambyiah, Pratheeban and Andre E.X. Brown (2021). "Quantitative Behavioural Phenotyping to Investigate Anaesthesia Induced Neurobehavioural Impairment". *Scientific Reports* 11.19398, pp. 1–10.
Neill, Alex and Aaron Ridley (1995). *The Philosophy of Art: Readings Ancient and Modern*. McGraw-Hill Education.
Nersessian, Nancy J. (1992). "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling". *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992.2, pp. 291–301.
— (2007). "Thought Experimenting as Mental Modeling: Empiricism Without Logic". *Croatian Journal of Philosophy* 7.20, pp. 125–161.
— (2018). "Cognitive Science, Mental Modeling, and Thought Experiments". In: *The Routledge Companion to Thought Experiments*. Ed. by Michael T. Stuart, Yiftach Fehige, and James R. Brown. London: Routledge, pp. 309–326.
Nguyen, James (2020). "It's Not a Game: Accurate Representation with Toy Models". *The British Journal for the Philosophy of Science* 71.3, pp. 1013–1041.
Nguyen, James and Roman Frigg (2017). "Mathematics Is Not the Only Language in the Book of Nature". *Synthese* 34, pp. 1–22.
— (2020). "Unlocking Limits". *Argumenta* 6.1, pp. 31–45.
— (2022a). "Maps, Models, and Representation". In: *Scientific Understanding and Representation*. Ed. by Insa Lawler, Kareem Khalifa, and Elay Shech. New York: Routledge, pp. 261–279.

Nguyen, James and Roman Frigg (2022b). *Scientific Representation*. Cambridge University Press.

Nicholson, Daniel J. (2012). "The Concept of Mechanism in Biology". *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43.1, pp. 152–163.

Norton, John D. (1985). "What Was Einstein's Principle of Equivalence?" *Studies in History and Philosophy of Science* 16.3, pp. 203–246.

— (1991). "Thought Experiments in Einstein's Work". In: *Thought Experiments in Science and Philosophy*. Ed. by Tamara Horowitz and Gerald J. Massey. Savage (MD): Rowman & Littlefield, pp. 129–148.

— (1996). "Are Thought Experiments Just What You Thought?" *Canadian Journal of Philosophy* 26.3, pp. 333–366.

— (2004a). "On Thought Experiments: Is There More to the Argument?" *Philosophy of Science* 71.5, pp. 1139–1151.

— (2004b). "Why Thought Experiments Do Not Transcend Empiricism". In: *Contemporary Debates in the Philosophy of Science*. Ed. by Christopher Hitchcock. Blackwell, pp. 44–66.

— (2008). "The Dome: An Unexpectedly Simple Failure of Determinism". *Philosophy of Science* 75.5, pp. 786–798.

— (2012). "Approximation and Idealization: Why the Difference Matters". *Philosophy of Science* 79.2, pp. 207–232.

— (2013). "Chasing the Light: Einstein's Most Famous Thought Experiment". In: *Thought Experiments in Philosophy, Science, and the Arts*. Ed. by Melanie Frappier, Letitia Meynell, and James R. Brown. London: Routledge, pp. 123–140.

— (2018). "Maxwell's Demon Does Not Compute". In: *Physical Perspectives on Computation, Computational Perspectives on Physics*. Ed. by Michael E. Cuffaro and Samuel C. Fletcher. Cambridge University Press, pp. 240–256.

— (2021). *The Material Theory of Induction*. University of Calgary Press.

Oriel, Christine and Paul Lasko (2018). "Recent Developments in Using *Drosophila* as a Model for Human Genetic Disease". *International Journal of Molecular Sciences 19* 19.2041, pp. 1–10.

Parke, Emily C. (2014). "Experiments, Simulations, and Epistemic Privilege". *Philosophy of Science* 81.4, pp. 516–536.

Parker, Matt W. (1998). "Did Poincaré really Discover Chaos?" *Studies in History and Philosophy of Modern Physics* 29.4, pp. 575–588.

Parker, Wendy S. (2020). "Model Evaluation: An Adequacy-for-Purpose View". *Philosophy of Science* 87.3, pp. 457–477.

Peacocke, Christopher (1987). "Depiction". *The Philosophical Review* 96.3, pp. 383–410.

Penrose, Roger (1965). "Gravitational Collapse and Space-Time Singularities". *Physical Review Letters* 14 (3), pp. 57–59.

Perini, Laura (2005). "The Truth in Pictures". *Philosophy of Science* 72.1, pp. 262–285.

— (2010). "Scientific Representation and the Semiotics of Pictures". In: *New Waves in Philosophy of Science*. Ed. by P.D. Magnus and Joseph Busch. London: Palgrave McMillan, pp. 131–154.

— (2013). "Diagrams in Biology". *Knowledge Eng. Review* 28.3, pp. 273–286.

Pitt, David (2022). "Mental Representation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2022. URL: https://plato.stanford.edu/archives/fall2022/entries/mental-representation/.

Portides, Demetris (2007). "The Relation between Idealisation and Approximation in Scientific Model Construction". *Science and Education* 16, pp. 699–724.

Potochnik, Angela (2015). "The Diverse Aims of Science". *Studies in History and Philosophy of Science Part A* 53, pp. 71–80.

— (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Price, Huw (2008). *René Descartes Lectures*. Tillburg. URL: http://philsci-archive.pitt.edu/archive/00004430/.

— (2010). "One Cheer for Representationalism?" In: *The Philosophy of Richard Rorty*. Ed. by Randall E. Auxier and Lewis Ewin Hahn. Chicago: Open Court, pp. 269–289.

— (2011). "Expressivism for Two Voices". In: *Pragmatism, Science and Naturalism*. Ed. by Jonathan Knowles and Henrik Rydenfelt. Zurich: Peter Lang, pp. 87–114.

— (2013). *Expressivism, Pragmatism, and Representationalism*. Cambridge University Press.

Psillos, Stathis (1999). *Scientific Realism: How Science Tracks Truth*. Routledge.

Radder, Hans (1996). *In and about the World: Philosophical Studies of Science and Technology*. State University of New York Press.

Rescorla, Michael (2009). "Cognitive Maps and the Language of Thought". *The British Journal for the Philosophy of Science* 2.60, pp. 377–407.

Reutlinger, Alexander, Dominik Hangleiter, and Stephan Hartmann (2018). "Understanding (with) Toy Models". *The British Journal for the Philosophy of Science* 4.69, pp. 1069–1099.

Rorty, Richard (1980). *Philosophy and the Mirror of Nature*. Princeton University Press.

— (1982). *Consequences of Pragmatism: Essays, 1972-1980*. University of Minnesota Press.

Rosen, Gideon (2010). "Metaphysical Dependence: Grounding and Reduction". In: *Modality: Metaphysics, Logic, and Epistemology*. Ed. by Bob Hale and Aviv Hoffmann. Oxford University Press, pp. 109–136.

Roskies, Adina L. (2008). "Neuroimaging and Inferential Distance". *Neuroethics* 1, pp. 19–33.

Roussos, Joe (2022). "Modelling in Normative Ethics". *Ethical Theory and Moral Practice* 25, pp. 865–889.

Ruyant, Quentin (2021). "True Griceanism: Filling the Gaps in Callender and Cohen's Account of Scientific Representation". *Philosophy of Science* 88.3, pp. 533–553.

Saatsi, Juha (2013). "Idealized Models as Inferentially Veridical Representations: A Conceptual Framework". In: *Models, Simulations, and Representations*. Ed. by Paul Humphreys and Cyrille Imbert. New York: Routledge, pp. 234–259.

Salis, Fiora (2016). "The Nature of Model-World Comparison". *The Monist* 99.3, pp. 243–259.

Salis, Fiora and Roman Frigg (2020). "Capturing the Scientific Imagination". In: *The Scientific Imagination*. Ed. by Arnon Levy and Peter Godfrey-Smith. Oxford University Press, pp. 17–50.

Salis, Fiora, Roman Frigg, and James Nguyen (2020). "Models and Denotation". In: *Abstract Objects*. Ed. by J.L. Falguera and Concha Martínez-Vidal. Synthese Library, vol. 422. Cham: Springer, pp. 197–219.

Sanches de Oliveira, Guilherme (2021). "Representationalism Is a Dead End". *Synthese* 198.1, pp. 209–235.

— (2022). "Radical Artifactualism". *European Journal for Philosophy of Science* 12.36, pp. 1–33.

Sartori, Lorenzo (2023). "Putting the 'Experiment' back into the 'Thought Experiment'". *Synthese* 201.34, pp. 1–36.

— (in press). "Model Organisms as Scientific Representations". *The British Journal for the Philosophy of Science*. URL: https://doi.org/10.1086/728259.

Schabas, Margaret (2018). "Thought Experiments in Economics". In: *The Routledge Companion to Thought Experiments*. Ed. by Michael T. Stuart, Yiftach Fehige, and James R. Brown. London: Routledge, pp. 171–182.

Schlaepfer, Guillaume and Marcel Weber (2018). "Thought Experiments in Biology". In: *The Routledge Companion to Thought Experiments*. Ed. by Michael T. Stuart, Yiftach Fehige, and James R. Brown. London: Routledge, pp. 243–254.

Schwarzschild, Karl (1916). *Sitzungsberichte der königlich preußischen Akademie der Wissenschaften*. Berlin: Deutsche Akademie der Wissenschaften zu Berlin.

Seim, Gretchen L. *et al.* (2019). "Two-Stage Metabolic Remodelling in Macrophages in Response to Lipopolysaccharide and Interferon-γ Stimulation". *Nature Metabolism* 1, pp. 731–742.

Shinod, N.K. (2017). "Why Thought Experiments Do Have a Life of Their Own: Defending the Autonomy of Thought Experimentation Method". *Journal of Indian Council of Philosophical Research* 34.1, pp. 75–98.

Sorensen, Roy A. (1998). *Thought Experiments*. Oxford University Press on Demand.

Starikova, Irina and Marcus Giaquinto (2018). "Thought Experiments in Mathematics". In: *The Routledge Companion to Thought Experiments*. Ed. by Michael T. Stuart, Yiftach Fehige, and James R. Brown. London: Routledge, pp. 257–278.

Stegenga, Jacob (2015). "Measuring Effectiveness". *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 54, pp. 62–71.

Strand, R, R. Fjelland, and T. Flatmark (1996). "In Vivo Interpretation of In Vitro Effect Studies". *Acta Biotheoretica* 44.1, pp. 1–21.

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

Stuart, Michael T. (2016). "Norton and the Logic of Thought Experiments". *Axiomathes* 26.4, pp. 451–466.

— (2020). "The Productive Anarchy of Scientific Imagination". *Philosophy of Science* 87.5, pp. 968–978.

Stuart, Michael T., Yiftach Fehige, and James R. Brown (2018). *The Routledge Companion to Thought Experiments*. Abingdon-on-Thames: Routledge.

Suárez, Mauricio (2003). "Scientific Representation: Against Similarity and Isomorphism". *International Studies in the Philosophy of Science* 17.3, pp. 225–44.

— (2004). "An Inferential Conception of Scientific Representation". *Philosophy of Science* 71.5, pp. 767–779.

— (2010). "Scientific Representation". *Philosophy Compass* 5.1, pp. 91–101.

— (2015). "Deflationary Representation, Inference, and Practice". *Studies in History and Philosophy of Science Part A* 49, pp. 36–47.

— (2024). *Inference and Representation: A Study in Modeling Science*. University of Chicago Press.
Suárez, Mauricio and Albert Solé (2006). "On the Analogy between Cognitive Representation and Truth". *Theoria* 55.1, pp. 39–48.
Suppes, Patrick (1960). "A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences". In: *Studies in the Methodology and Foundations of Science. Selected Papers from 1951 to 1969*. Ed. by Patrick Suppes. Dordrecht: Springer-Science+Business Media, pp. 10–23.
— (1967). "What Is a Scientitic Theory?" In: *Philosophy of Science Today*. Ed. by Sidney Morgenbesser. New York: Basic Books, pp. 55–67.
— (1970). *Set-Theoretical Structures in Science*. Stanford: Institute for Mathematical Studies in the Social Sciences, Stanford University.
Swoyer, Chris (1991). "Structural Representation and Surrogative Reasoning". *Synthese* 87, pp. 449–508.
Tan, Peter (2021). "Inconsistent Idealizations and Inferentialism about Scientific Representation". *Studies in History and Philosophy of Science Part A* 89A, pp. 11–18.
Thoma, Johanna (2016). "On the Hidden Thought Experiments of Economic Theory". *Philosophy of the Social Sciences* 46.2, pp. 129–146.
Thomasson, Amie (2020). "If Models Were Fictions, Then What Would They Be?" In: *The Scientific Imagination*. Ed. by Arnon Levy and Peter Godfrey-Smith. Oxford University Press, pp. 51–74.
Thompson, Richard A., James M. Moran, and George W. Swenson (2017). *Interferometry and Synthesis in Radio Astronomy*. Springer Nature.
Thomson-Jones, Martin (2010). "Missing Systems and Face Value Practise". *Synthese* 172.2, pp. 283–299.
— (2011). "Structuralism about Scientific Representation". In: *Scientific structuralism*. Ed. by Alisa Bokulich and Peter Bokulich. Dordrecht: Springer, pp. 119–141.
Tobin, James (1970). "Money and Income: Post Hoc Ergo Propter Hoc?" *The Quarterly Journal of Economics* 84.2, pp. 301–317.
Todd, Cain (2020). "Imagination, Aesthetic Feelings, and Scientific Reasoning". In: *The Aesthetics of Science: Beauty, Imagination and Understanding*. Ed. by Milena Ivanova and Steven French. London: Routledge, pp. 63–85.
Toon, Adam (2010). "Models as Make-Believe". In: *Beyond Mimesis and Convention*. Ed. by Roman Frigg and Matthew C. Hunter. Dordrecht: Springer, pp. 6–29.
— (2012). *Models as Make-Believe: Imagination, Fiction and Scientific Representation*. London: Palgrave Macmillan.
Tufte, Edward R. (1997). *Visual and Statistical Thinking: Displays of Evidence for Decision Making*. Chelshire: CT: Graphic Press.
Tversky, Amos (1977). "Features of Similarity". *Psychological Review* 84.4, pp. 327–352.
Tversky, Amos and Itamar Gati (1978). "Studies of Similarity". In: *Cognition and Categorization*. Ed. by Eleanor Rosch and Barbara B. Lloyd. Hillside (NJ): Lawrence Elbaum Associates, pp. 79–98.
Vaihinger, Hans (1924). *The Philosophy of "as if": A System of the Theoretical, Practical and Religious Fictions of Mankind* (English translation). London: Kegan Paul.
van Fraassen, Bas (1980). *The Scientific Image*. Oxford University Press.

van Fraassen, Bas (2008). *Scientific Representation: Paradoxes of Perspective*. Oxford University Press.

Volterra, Vito (1926). "Fluctuations in the Abundance of a Species Considered Mathematically". *Nature* 118.2972, pp. 558–560.

— (1928). "Variations and Fluctuations of the Number of Individuals in Animal Species Living Together". *Journal du Conseil* 3, pp. 3–51.

Walsh, Jonelle L. *et al.* (2013). "The M87 Black Hole Mass from Gas-Dynamical Models of Space Telescope Imaging Spectrograph Observations". *The Astrophysical Journal* 770.2, pp. 1–11.

Walton, Kendall L. (1990). Mimesis *As Make-Believe: On the Foundations of the Representational Arts*. Cambridge (MA): Harvard University Press.

Ward, Paul *et al.* (2019). *The Oxford Handbook of Expertise*. Oxford University Press.

Weber, Marcel (2004). *Philosophy of Experimental Biology*. Cambridge University Press.

— (2014). "Experimental Modeling in Biology: In Vivo Representation and Stand-Ins as Modeling Strategies". *Philosophy of Science* 81.5, pp. 756–769.

Weisberg, Michael (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Weisberg, Michael and Kenneth Reisman (2008). "The Robust Volterra Principle". *Philosophy of Science* 75.1, pp. 106–131.

Willats, John (1997). *Art and Representation: New Principles in the Analysis of Pictures*. Princeton University Press.

Wilson, James (2016). "VII—Internal and External Validity in Thought Experiments". *Proceedings of the Aristotelian Society* 116.2, pp. 127–152.

Wimsatt, William C. (1981). "Robustness, Reliability, and Overdetermination". In: *Reengineering Philosophy for Limited Beings*. Ed. by Id. Cambridge (MA): Harvard University Press, pp. 43–74.

— (1987). "False Models as a Means to Truer Theories". In: *Neutral Models in Biology*. Ed. by Matthew H. Nitecki and Antoni Hoffman. Oxford University Press, pp. 23–55.

Winsberg, Eric (2001). "Simulations, Models, and Theories: Complex Physical Systems and their Representations". *Philosophy of Science* 68 (Proceedings), S442–S454.

Wollheim, Richard (1987). *Painting as an Art*. London: Thames and Hudson.

Worrall, John (2007). "Evidence in Medicine and Evidence-Based Medicine". *Philosophy Compass* 2.6, pp. 981–1022.

Wouters, Arno G. (2003). "Four Notions of Biological Function". *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 34.4, pp. 633–668.

Young, James (2003). *Art and Knowledge*. London: Routledge.