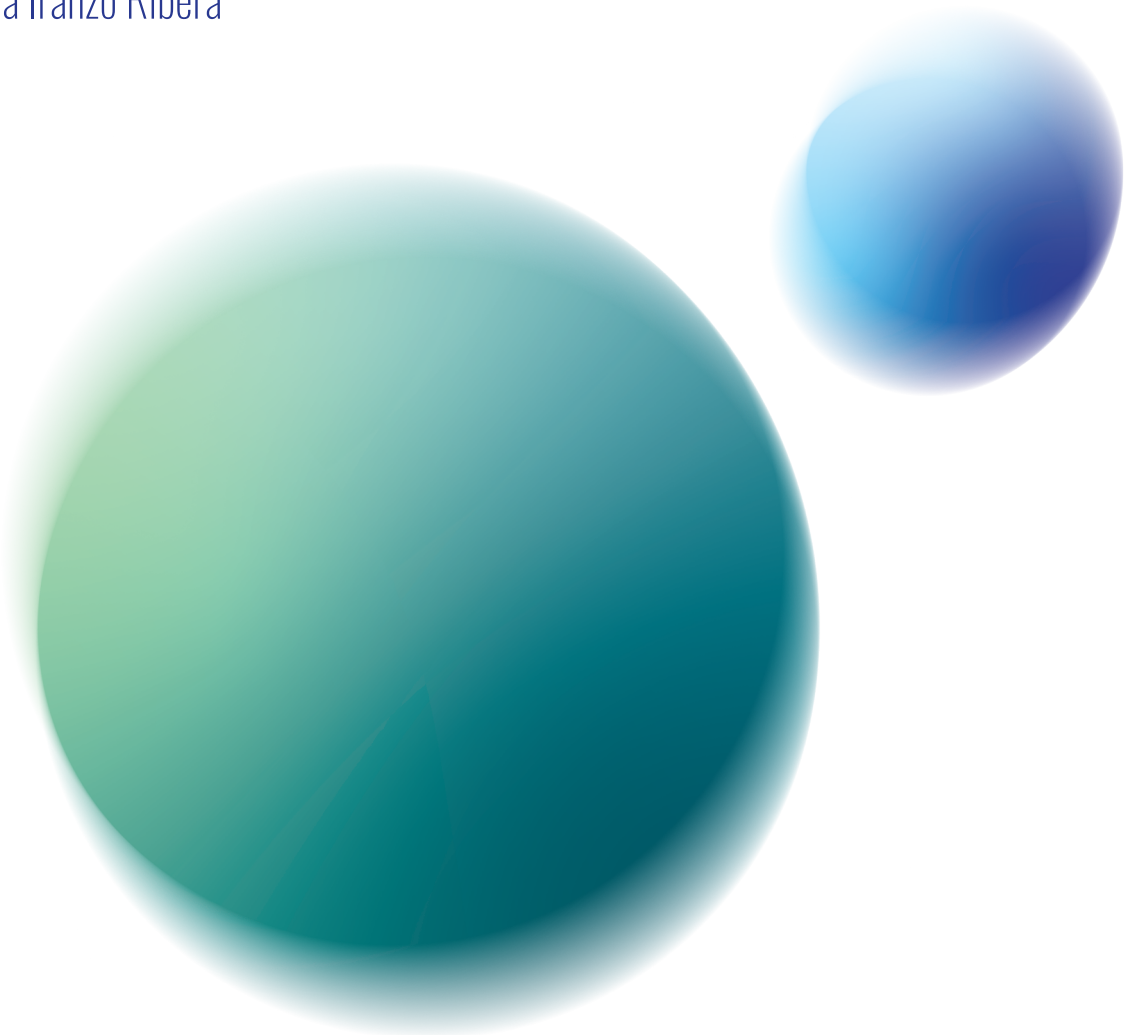
POSSIBILITIES FOR INTERVENTIONIST EXPLANATION

Conceptual, Physical & Fictional

By Noelia Iranzo Ribera



A thesis submitted to the University of Birmingham
for the degree of DOCTOR OF PHILOSOPHY

Department of Philosophy | School of Philosophy, Theology, and Religion
College of Arts and Law | University of Birmingham
February 2023

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

To my grandparents Carmeta and Josep,
my guardian angels.

A la iaia Carmeta i l'avi Josep,
els meus àngels de la guarda.

Abstract

This thesis examines a series of questions about *possibility*. The starting point is a neglected claim in James Woodward's interventionist theory of causation: interventions are non-trivially *possible*. Noting that interventionism is a realist theory of causation, in the first part of this thesis I explore what notion of possibility could possibly satisfy the interventionist's objectivity demands. In chapter 1, I challenge Woodward's argument that interventions need not be physically possible, an argument which sets the ground for what he takes to be the relevant notion of possibility: conceptual possibility. In chapter 2, I present a charitable interpretation of conceptual possibility as an objective kind of possibility arrived at via conceivability. I note that rationalist approaches to modal knowledge may fit the bill. To this end, I explore whether David Chalmers' modal rationalism is able to furnish interventionism with a suitable notion of possibility. In chapter 3, I turn to Woodward's Invariance(-based) View of laws to argue that, in fact, possibility ought to be physical, where physical possibility is construed as nomological possibility plus information about initial, boundary, and background conditions.

Chapter 3 puts the emphasis on scientific counterfactuals. In the second half of this thesis, I explore what sorts of counterfactuals feature in scientific explanation. More specifically, I investigate their explanatory functions in connection to scientific models and model-based reasoning. In chapter 3, I argue that scientists can sometimes make sense of counterfactuals which describe situations in which the laws had been different, and show how the Invariance View can be extended to accommodate these counterfactuals. In chapter 4, I introduce and address two challenges to the interventionist claim that counterfactuals are explanatory when they are 'same-object' counterfactuals, that is, counterfactuals which describe invariance within the same object. These are *the individuation challenge* and *the idealisation challenge*. Finally, in chapter 5 I propose two desiderata any naturalised account of counterfactuals and counterfactual reasoning in science ought to fulfil and, to this end, defend the *make-believe view of scientific counterfactuals*. This account, which uses a notion of fiction connected to the imagination, has a high unificatory potential: it accommodates regular counterfactuals, counternomics, and counterpossibles alike. Hence, the make-believe view of scientific counterfactuals outstrips the interventionist view of counterfactuals.

Acknowledgements

Many times had I been told that a PhD is like a marathon; little did I know how true that is. Circumstances have made it so that my PhD experience has not been precisely smooth: from some initial hardship adjusting to life in a new country alone, then covid-19 – which meant remote work for a big part of the process – to finally juggling writing up with work. If I have made it thus far is because of the following people, to whom I owe the completion of this thesis.

First and foremost, I owe Prof Alastair Wilson the opportunity to join the academic and non-academic family FraMEPhys has been during these years. He is a well of wisdom, but also – and even more remarkable – a caring and reachable supervisor and mentor who has had infinite patience with my ups and down, as well as my eventual questioning of the academic career path. Without his constant guidance and support, this thesis would not have reached completion, so I am truly indebted to him for making it *practically possible*. To this end, I am incredibly grateful to the European Research Council for the funding received under grant agreement no. 757295.

To Dr Katie Robertson, who has been more than a supervisor to me. She has invested infinite time and effort in me, going way beyond her contractual obligations, and has constantly encouraged me to build self-confidence when I struggled. She is a truly inspiring figure: not just an excellent scholar, but a kind, caring, and fun friend. I am also grateful to her for having helped start the Feminist Philosophy reading group at Birmingham, which, thanks to Kathleen Murphy-Hollies first and Jessica Sutherland now, is still up and running.

To Dr Henry Taylor, who has co-supervised part of this thesis. In particular, I am very grateful to him for patiently walking me through the material in chapter 2 from zero to ''hero" and for providing extremely detailed feedback on multiple drafts of this chapter.

To the other core members of FraMEPhys: Dr Michael Townsen Hicks, Dr Joaquim Giannotti, and Nicholas Emmerson. For their companionship and constant support (often in the form of endorsement in the academic Twitter jungle), reading groups, and feedback on different chapters of this thesis. You have all set the bar high; it will be difficult to find colleagues who are as bright, hard-working, and kind. Besides the FraMEPhys family, I am the luckiest to have encountered another gem: Kathleen Murphy-Hollies, whose invaluable support throughout this journey has made me feel understood, valued, and loved.

I am also indebted to Dr Niels van Miltenburg and Dr Janneke van Dis, who kindly offered me the opportunity to teach at Utrecht University. This job has provided the financial resources to make the writing-up period sustainable, significantly widened my teaching experience, and, even more importantly, kept me sane and motivated during this period. I am also grateful to my colleagues in Utrecht for the warm welcome, guidance, and trust they placed in me.

Finally, I would also like to thank my thesis examiners Dr Lina Jansson and Dr Michael T. Hicks for a very enjoyable viva discussion and their excellent feedback on how to further develop this research in future work. I would also like to thank Dr Merten Reglitz for his superb job at chairing the thesis examination.

Thanks should also go to many life-long and new friends. To Marta Conti Lorenzo, fellow philosopher and close friend, for her steadfast support since our Physics undergraduate days. I want to thank her for her readiness to provide last-minute feedback, read drafts, discuss career options, and be always available for a call whenever needed. To Nick Wiggerhaus and Sam Rijken, with whom I have had the pleasure to – first remotely, now in person – share this PhD journey. I am very happy we decided to organise the *Fictions in Science and Metaphysics* symposium for BSPS21, which led to chapter 5 of this thesis. Therefore, I would like to thank them for their feedback on this material.

Downstream the friendship road, I would like to thank Matilde Aliffi, whom I deeply admire. She introduced me to the department and life in Birmingham more broadly and was the gate to many other great friendships. Many thanks to Alessandro Carabia, Auro M. Perego, Flavia Vanni, Vicky Kaisidou, Greg Mathers, and Jodie Neville for bringing the fun (as well as the excellent food) to my Brummie life. I would also like to thank Luis Freijo and Francisco López-Santos – my *sanedrín de sabios* – for being incredible friends, confidants, and lunch companions, as well as for teaching me so much about the byzantine empire and non-American westerns, two topics I had not expected to learn so much about. Cheers to friends with niche interests who bring new insights into your life! I am also fortunate to have met many other amazing people along the way, especially Adrià Fontrodona Bach, Anabel Palacios Trujillo, Andrea Vecchi, Fanny Corlosquet, Silvia Fornara, Gabriele Humbert and Jenny Van.

A continuació, voldria també dedicar unes paraules a la meva família, per la formació rebuda i el seu suport incondicional. Després de més de set anys, encara resulta complicat ser tan lluny de vosaltres, així que gràcies per fer-me arribar l'escalf i donar-me forces per seguir sempre endavant. En especial, voldria agrair a la meva germana Eva haver-se passat anys interessant-se tant pel progrés d'aquesta tesi com per la meva salut mental. Evidentment, el màxim agraïment se

l'endú ma mare Carme, l'artífex de tot plegat. En primer lloc, pel suport econòmic rebut durant el màster, sense el qual no hagués pogut arribar fins aquí. Però sobretot, per haver-me criat des de la llibertat i la confiança plena en les meves eleccions.

Last but not least, I am genuinely indebted to Niels, who has most closely shared this crazy ride with me. First and foremost, I would like to thank him for his love and infinite patience: putting up with my incomprehensible explanations and rants for years must be no easy feat. He is also the one to credit for most of the artwork in this thesis: from the cover page to most of the figures (see `http://nielsdielen.nl`). But above anything else, I feel so blessed by his endless curiosity: from philosophy to art, sharing awe for the wonders of life is truly a gift. Dank je uit het diepst van mijn hart.

Contents

List of figures **11**

Introduction **13**

1 The Possibility of Interventionism **19**
1.1 Introduction . . . 19
1.2 Interventionism: The Formal Framework . . . 19
1.2.1 The Causal Modelling Framework . . . 20
1.2.2 What Are Interventions? . . . 22
1.2.3 Causal Representation & Explanation . . . 24
1.2.3.1 Causal Relata . . . 25
1.2.3.2 Causal Relations . . . 28
1.2.3.3 Levels of Explanation . . . 31
1.3 Interventionism & The Possibility Question . . . 32
1.3.1 Agency and Possibility . . . 33
1.3.2 The Moon-Tides Case Study . . . 35
1.3.2.1 Physical Possibility . . . 37
1.3.2.2 The Rejection of Physical Possibility Argument . . 39
1.3.2.3 Contesting the Rejection of Physical Possibility Argument . . . 44
1.3.2.4 Conceptual Possibility . . . 48
1.3.3 Causal Realism . . . 50
1.4 Conclusion . . . 52

2 Possibility via Conceivability: A Rationalist Account of Interventionism **55**
2.1 Introduction . . . 55
2.2 What is Weak Modal Rationalism? . . . 56
2.3 What is Conceivability? . . . 58
2.3.1 Prima facie vs ideal conceivability . . . 58
2.3.2 Positive vs negative conceivability . . . 59
2.3.3 Primary vs secondary conceivability and possibility . . . 60
2.3.4 Concerns with the CP thesis . . . 64
2.4 The Conceivability Argument (CA) . . . 66

2.5 The Moon-tides Conceivability Argument (MTCA) 68
2.6 Interventionism & Physical Properties 71
2.6.1 Functionalism about Physical Properties 71
2.6.2 Non-functionalism about Physical Properties 73
2.6.3 Functionalism or Non-Functionalism? 74
2.6.3.1 Russellian Monism: The Modified Conceivability Argument (CA*) 75
2.6.3.2 Russellian Physicalism: The Modified Moon-tides Conceivability Argument (MTCA*) 76
2.7 Does WMR Answer the Possibility Question? 79
2.8 Conclusion 80

3 Possibility via Invariance I: A Physical Account of Interventionism 83
3.1 Introduction 83
3.2 The Possibility Question Revisited 84
3.3 The Invariance View 86
3.3.1 What is Invariance? 86
3.3.2 What is the Invariance View? 89
3.3.3 The Methodology of Invariance 93
3.3.3.1 Independence Laws - ICs 93
3.3.3.2 Independence IC - IC 95
3.3.3.3 Independence Law - Law 96
3.3.4 The Epistemology of Invariance: The Epistemic Circularity Concern 96
3.3.5 The Metaphysics of Invariance: The Modal Priority Question 100
3.4 The Modal Priority Question Revisited 102
3.4.1 Marc Lange's View of Laws 103
3.4.2 Tim Maudlin's View of Laws 105
3.4.3 Woodward in Perspective: The Modal Priority Question Answered 107
3.5 The Extended Invariance View 110
3.5.1 What Would Have Happened Had the Laws Been Different 110
3.5.2 What is the Extended Invariance View? 111
3.5.3 The Possibility of Higher-Order Interventions 113
3.5.4 The Characteristics of Higher-order Invariance 116
3.5.4.1 The Explanatory Status of Higher-order Invariance 116
3.5.4.2 The Metaphysical Status of Higher-order Invariance 118

3.6 The Possibility Question Answered: Whence Causal Realism? . . . 121
3.7 Conclusion . . . 124

4 Possibility via Invariance II: The Object of Interventionist Counterfactuals 127
4.1 Introduction . . . 127
4.2 Interventionist Counterfactuals & Scientific Explanation . . . 128
4.2.1 The Standard View of Scientific Explanation: Other-Object Counterfactuals . . . 129
4.2.2 The Invariance View of Causal Explanation: Same-Object Counterfactuals . . . 130
4.2.3 Why Same-Object Counterfactuals? . . . 133
4.3 The Individuation Challenge . . . 135
4.3.1 What is the Individuation Challenge? Tokens vs Types . . . 136
4.3.2 A Case Study: Sea Water's Electrical Conductivity . . . 137
4.3.3 The Solution to the Individuation Challenge . . . 139
4.3.3.1 Science Considered: Models of Kinds . . . 140
4.3.3.2 Science Considered: Hypothetical Interventions . . 142
4.3.3.3 Types of Generality & Explanatoriness . . . 145
4.4 The Idealisation Challenge . . . 146
4.4.1 Idealisation in Science . . . 147
4.4.2 What is the Idealisation Challenge? Idealising Counternomics 149
4.4.3 The Solution to the Idealisation Challenge . . . 152
4.4.4 Finding SO-enabling-CFs Involving Idealisation . . . 154
4.4.4.1 Analysis . . . 159
4.4.4.2 Limitations . . . 159
4.5 No Abstraction Challenge . . . 162
4.5.1 Abstraction vs Idealisation . . . 163
4.5.2 An Objection: Abstraction as Distinctive of Non-Causal Explanation . . . 165
4.5.3 A Reply: Interventionism Beyond Causal Explanation . . . 167
4.5.4 An Own Goal? Two Types of RG Explanation . . . 168
4.5.5 Same-Object Counterfactuals and Abstract Models . . . 171
4.6 Conclusion . . . 174

5 Possibility via Fiction: Scientific Counterfactuals as Make-Believe 177
5.1 Introduction . . . 177

5.2 Counterfactuals in Science 178
5.2.1 Counterfactuals and Scientific Models 179
5.2.2 Counterfactuals and Non-vacuous Truth 181
5.3 The Toolbox for Counterfactual Analysis 182
5.3.1 A Philosopher's Take: The Similarity Analysis 182
5.3.2 A Cognitive Psychologist's Take: The Mental Models Theory 185
5.3.3 Assembling the Pieces: Scientific Models and the Imagination 187
5.4 Fictionalism About Scientific Counterfactuals 189
5.4.1 Contextualising Fiction 189
5.4.1.1 Fiction as Imagination: Make-believe 189
5.4.1.2 The Fiction View of Models 191
5.4.2 The Make-Believe View of Scientific Counterfactuals 194
5.4.2.1 Exposition 194
5.4.2.2 Extension on Previous Views 198
5.4.2.3 Case Studies 200
5.5 Conclusion 204

Conclusions **207**

Bibliography **213**

List of figures

1.1 Causal structures ruled out by **I3**. 23
1.2 Causal structure ruled out by **I4**. 23
1.3 Fine-tuned causal structure: the causal paths $X \rightarrow Z \rightarrow Y$ and $X \rightarrow Y$ cancel each other out. 30
1.4 Tidal bulges due to gravity (the right one, facing the Moon) and inertia (the left one, on the opposite side). 36
1.5 Tidal forces on the Earth's surface at baseline sea level created by the Moon's gravitational pull (OpenStax University Physics). 36
1.6 Simplest causal graph for an intervention which tests **Moon-tides**. 37
1.7 SPP and WPP delimit concatenated spaces of physical possibility. . 38
1.8 System comprised of Mars, the Moon, and the Earth. 40
1.9 Causal graph for MT-I_1. 41
1.10 Causal graph for MT-I_2. 41

2.1 Putnam's Twin Earth thought experiment. 61
2.2 Representation of the relative distances between the bodies in the three-body system Mars-Moon-Earth pre (r_{i-j_A}) and post intervention (r_{i-j_C}). The point between the Moon and the Earth represents r_{Ma-M}'s position pre-intervention. 69

3.1 Representation of the physical modal space. 88
3.2 Invariance View in terms of (N). 90
3.3 Lange's view of laws. 104
3.4 Breadth of invariance provides an objective ordering of generalisations. 112
3.5 Notions of physical possibility (strong, weak, and very weak) centered on actuality. 114

5.1 Simple pendulum representation 201
5.2 Relation of boiling points (ºC) versus molar mass (Kg/Kmol) for molecules in hydrides' groups 14, 15, 16, and 17. 203

Introduction

> "And shall we esteem it worthy the labour of a philosopher to give us a true system of the planets, and adjust the position and order of those remote bodies; [...]"
>
> — Hume, *Of the Different Species of Philosophy*

Modal reasoning is not only central to daily life but also science. The content of counterfactual conditionals, models, and laws informs us about what can (can't), could (couldn't), or would (wouldn't) be the case. No aircraft can travel faster than light, though they may reach supersonic speeds. This Airbus you are in could have been flying at 900km/h instead of 850km/h; had that been the case, you would be in New York by now. And if your parents were not affected by genetic anticoagulant deficiencies, you would now have a lower risk of suffering thrombotic events during long-haul flights (though thank them for your blue eyes; hadn't they had blue eyes, you wouldn't have them either!). You are about to embark on a modal journey into what is possible, which via the lens of interventionism about causation, will take you from causation and counterfactual conditionals to laws and scientific models. The second part of the journey will concern how we reason about and understand the nature of these possibilities.

Since antiquity, the regularities we experience have been described in terms of causes and effects. Causation has given rise to many challenges and questions, but importantly, philosophers have ever since wondered about its nature: is it a feature of the world, or rather a useful feature of our psychological make-up which endows us with the certainty that, e.g. the sun will rise tomorrow? Those who advocate the former view are causal realists. Versions of causal realism have come in many and diverse flavours; among the most prominent views, we find those which understand causation as nomological subsumption (Davidson 1980), statistical correlation (Hitchcock 2001; Pearl 2000; Spirtes et al. 2000; Suppes 1970), counterfactual dependence (D. K. Lewis 1986a, 2000; Menzies 1989), or physical processes (Dowe 1992, 2000; Salmon 1984, 1998).[1]

However, not everyone subscribes to realism about causation. In line with Humean scepticism about causation, Russell famously advocated eliminativism about causation when he called it "a relic of a bygone age, surviving, like the

[1] See Schaffer (2016b) for a complete taxonomy of all the extant realist analyses of causation.

monarchy, only because it is erroneously supposed to do no harm" (Russell 1912). Echoing this notorious analogy, Price (2004, 2007) has suggested causal republicanism as an alternative which views causation as both useful – perhaps even ineliminable – and anthropocentric. Causal republicanism is the view that notions of agency and manipulation are central to understanding causation (see Menzies & Price 1993; von Wright 1971). Agential manipulability theories are pragmatic in aspiration: their goal is to account for the role of causation in agents' practical and cognitive lives. We can also view these agential theories of causation as functional, for according to Nancy Cartwright (1979) causal laws are 'effective' strategies for obtaining particular goals. Her understanding of causal relations as those which can be exploited to reach desired outcomes was pivotal in carving the path for manipulationist approaches to causation. Generally speaking, in manipulability theories of causation *A* causes *B* if bringing about *A* is an effective way of bringing about *B* (Menzies & Price 1993, p.187). This manipulationist line of thought, conceived as anthropocentric and reductive in aspiration by its early proponents, culminated in the state-of-the-art interventionist theories of causation, which are non-anthropocentric and non-reductive.

Interventionist theories of causation originated in Judea Pearl's *Causality* (2000) and Peter Spirtes, Clark Glymour, and Richard Scheines' *Causation, Prediction and Search* (2000). These theories use causal models, characterised as collections of causal mechanisms and formally represented and described by systems of equations. Causal mechanisms are disrupted via manipulations of the values of the variables in these causal models, which we call 'interventions.' In the following thesis, I focus on James Woodward's more qualitative version of the interventionist theory of causation. Woodward gave a philosophical twist and extensively elaborated on the framework introduced by Pearl and Spirtes et al., for which his work has received much attention and praise. Only in chapter 4 do I explore his joint work with Christopher Hitchcock (Hitchcock & Woodward 2003; Woodward & Hitchcock 2003) on causal explanation.

One feature that makes Woodward's interventionist theory of causation – to which I shall hereafter refer by the name of 'interventionism' – as attractive as it is puzzling is its unique combination of causal realism and causal republicanism. On the one hand, it presupposes a minimal metaphysical realism, for it views causes and effects as constituents of the world. In order to accommodate this causal realism interventionism is severed from human agency; since interventions need not be carried out by human beings the notion of an intervention appeals to a more abstract notion of manipulation. On the other hand, it acknowledges that human beings are interested in causal relationships insofar as they can be

manipulated and controlled in line with their interests. The title of Woodward's magnus opus, *Making Things Happen*, is suggestive of this functional motivation of interventionism: we, agents, intervene in the world to make things happen according to our aims and abilities. However, as said, interventions need not be practically possible. In fact, they need not even be actual – causes and effects must be manipulable at least *in principle*. After all, interventionism is a counterfactual theory of causation: the meaning of causal claims is clarified via counterfactuals associated with actual or hypothetical experiments. It is via these we evaluate *what-if-things-had-been-different* claims. The use of counterfactuals to understand causal claims does not rely on any form of modal realism and/or appeal to similarity relations among possible worlds – recall that within philosophy, Stalnaker's (1968) and Lewis' (1973) are the best-known examples of theories which attempt to explain causal claims in terms of counterfactual claims – but instead on the observation that both people in their daily life and scientists have found it helpful to think of causal claims in terms of counterfactuals. This connection between causal and counterfactual thinking has also been explored in the psychological literature, though not much dialogue has been established between the two disciplines, namely, philosophy and psychology – one such exception is, again, Woodward's most recent book *Causation with a Human face* (Woodward 2021). The present thesis starts by examining what sorts of possibilities interventions are, to then explore how we reason with and understand the limits of these possibilities. In the following, I provide an outline of the chapters.

The sense in which hypothetical interventions are possible, what I call the 'possibility question,' constitutes the backbone of the first half of the present thesis. As highlighted, one element of interventionism that has created confusion is the fact that the notion of manipulability embedded in interventions goes beyond actual experimentation. This has led Woodward himself to deny that interventions are *physically possible* and defend instead *conceptual possibility* as the kind of modal notion which underwrites interventions. In **chapter 1**, I introduce the possibility question via an example that will accompany us throughout this thesis: the causal claim that changes in the position of the Moon cause changes in the magnitude of the Earth's tides, and the associated counterfactual that describes a hypothetical intervention that would double the distance between the Moon and the Earth. I then reconstruct and contest Woodward's argumentation that such intervention need not be physically possible but only conceptually so, and bring up a challenge: conceptual possibility does not mesh well with the presupposed causal realism of interventionism.

In **chapter 2**, I suggest that one way to interpret Woodward's conceptual

possibility – and the associated conceivable ideal interventions – which preserves the realism interventionism presupposes is via a rationalist strategy such as David Chalmers' weak modal rationalism (WMR). WMR is based on a core thesis which establishes an entailment relation between ideal conceivability and possibility. I argue that to know whether the interventionist can make use of said conceivability-possibility link, a metaphysics for the individuation of physical properties needs to be spelt out. I then explore how this metaphysics affects the prospects of finding an answer to the possibility question.

Of course, the rationalist strategy just described may sound rather wild to the empirically-minded philosopher. Indeed, there are more straightforwardly empirically-friendly modal epistemologies. In **chapter 3**, I return to the view that interventions need to be physically possible, where physical possibility is construed in terms of laws and initial and background conditions. After dispelling a twofold epistemic and metaphysical circularity concern regarding the view that interventions are physical possibilities, I conclude that the physical possibility approach is the best answer to the PQ. Chapter 3 thus marks the end of the first half of this thesis.

In the second part of this thesis, I explore what sorts of counterfactuals feature in scientific explanation and what their connection to laws and models is. Because we are interested in counterfactuals insofar as they help us, e.g. make predictions, investigate the scope of theories, and explain the world, the focus is more broadly on how we reason with these. Quite often, counterfactuals in science describe situations that can't be realised, either because the laws of nature don't permit it or because they contradict a posteriori necessities concerning constitutive facts. Nonetheless, we generally seem to be able to assign truth-values to counterfactuals that describe impossibilities; this is the case of counternomics which describe situations in which other-than-actual first-order laws apply. In chapter 3, I develop an extension of Woodward's invariance-based view of laws which accommodates this type of counterfactuals – I even go further and argue that they describe 'very weak physical possibilities.'

In **chapter 4**, I turn to another type of counternomics: those that appeal to idealised models, which I accordingly call 'idealising counternomics.' I explore their role in explanation in the context of a distinction drawn by Woodward and Hitchcock (2003a, 2003b) between explanatory and non-explanatory generalisations, as well as between the counterfactuals these support. Given that generalisations are explanatory insofar as they encode modal information about *one and the same object*, interventionist counterfactuals are also known as 'same-object' counterfactuals (SO-CFs). Counterfactuals which provide information about the relations

between *different objects*, and are thereby not causally explanatory, are 'other-object' counterfactuals (OO-CFs). In chapter 4, I first contest the idea that SO-CFs do not provide information about scope on the basis that in scientific explanations SO-CFs typically provide information about classes of systems, and not particular systems. This is the *individuation challenge*. The SO vs OO distinction also poses the *idealisation challenge*: if all there is to scientific explanation is causal explanation, Woodward and Hitchcock can't account for the function of idealising counternomics in explanation, which is to *enable* causal explanations. To remedy this, I explore the prospects of an interventionist strategy that justifies their explanation-enabling role. I also discuss why abstraction does not raise any analogous challenge – though, via the local context of abstraction in RG explanation, we will see how it provides valuable insight into how the SO vs OO distinction regarding counterfactuals might be extended beyond causal counterfactuals.

Finally, **chapter 5** aims to bring together the philosophical literature on counterfactuals and scientific models, on the one hand, and the psychological literature on counterfactual reasoning, on the other. The goal is to provide an account of the meaning of counterfactuals in science. In order to do so, I set interventionism aside and develop a fiction-based account of counterfactuals in science which, just like interventionism, is functional in aspiration and takes scientific practice at face value. This account is the 'make-believe view of scientific counterfactuals', which characterises counterfactuals via the propositional attitude of pretense or 'make-believe' introduced by Kendall Walton in his book *Mimesis as Make-Believe* (Walton 1990), and later developed into the 'fiction view of models' (see e.g. Frigg 2010a,b; Levy 2015; Salis 2020; Toon 2012 or Frigg & Nguyen 2020). Given the close connection between scientific models and counterfactuals, I propose to apply it to counterfactuals. This application which gives us the expected result that one may reason successfully about impossible states of affairs. Since make-believe is a notion of fiction qua *imaginability* this chapter thus complements chapter 2, which revolved around the related notion of conceivability (though a technical version thereof).

Chapter 1

The Possibility of Interventionism

1.1 Introduction

James Woodward's interventionist account of causation rests on the idea that "X is a cause of Y, iff there is a possible intervention on X that changes [the value or probability distribution over values of] Y" (Woodward 2003, p.45).[2] Interventions have to satisfy further formal requirements (e.g. be causes themselves, be localised, etc.), but most crucially they have to be *possible*. While actual interventions are always possible (since actuality entails possibility in all major approaches to modal logic), these interventions might also be 'hypothetical.' After providing a concise summary of interventionism, in this chapter I introduce one of the main although overlooked controversies surrounding interventionism: the sense in which interventions are possible. This is what I call the 'possibility question' (PQ hereafter). Insofar as possible interventions cannot be identified, causal claims are left vague, indeterminate, or contentless, hence the importance of providing a satisfactory answer to the PQ.

1.2 Interventionism: The Formal Framework

Interventionism is a theory of causation and causal explanation. While in the current chapter I focus on the modal component of interventions, it is paramount I first provide an introduction of the formal framework within which interventionism is embedded. This formal framework is necessary background to contextualise both the account of causation and causal explanation interventionism provides. Therefore, in §1.2.1 and §1.2.2 I introduce causal models, interventions, and interventionist causal models formally. Then, in §1.2.3 I focus on how the interventionist uses these causal models to explain.

[2]Square brackets mine. If the system is deterministic an intervention on X changes the value of Y; if the system is indeterministic it changes the probability distribution over values of Y.

1.2.1 The Causal Modelling Framework

Interventionism uses the causal modelling framework, an umbrella term that applies to a wide range of formal methods for inferring and representing both deterministic and probabilistic causal relationships. Contributions to the causal modelling literature have come from fields as diverse as computer science – where the works *Causality* (2000) by Judea Pearl, and *Causation, Prediction, and Search* (2000) by Peter Spirtes, Clark Glymour, and Richard Scheines are usually highlighted – but also statistics, epidemiology, or econometrics.[3]

Together with prediction, the other main function of causal models is representation, key for explanatory purposes. Causal models represent causal structure via directed acyclic graphs (DAGs), that is, graphs where the nodes or vertices are variables connected via directed edges in a way that no cycles or loops are formed. The variables in DAGs are classical variables which are either binary and represent the occurrence of an event or the having of a property by a physical system, or take a value from a discrete or continuous set of values. For instance, my choice of which online series I am going to watch tonight is the variable 'series', and the values it may take come from a domain restricted by e.g. Netflix's series repository. Causal models may be deterministic or probabilistic; provided models are probabilistic, these variables are random, i.e. the values these variables may attain have probabilities attached.

On a terminological note, variables can be divided into 'exogenous' or 'endogenous' depending on whether their values are fixed by factors external to the model under study (exogenous) or they are dependent on the value of variables internal to the model (endogenous). Intervention variables, to be defined shortly, are exogenous variables. Variables may also be described via the terminology of 'parent-child' relationships. In this fashion, exogenous variables correspond to parentless nodes (also known as 'roots'). Similarly, endogenous variables which represent final nodes in a causal path – that is, variables which are not causes – are also known as childless nodes or 'sinks.'

If variables are the relata of causal models, causal relationships are graphically represented via edges, and formally described via structural equations, in which each endogenous variable may be written as a function of all and only the variables which are its direct causes Woodward (2003, p.42). This is why this framework is also known as the 'structural equations' framework.

Consider a deterministic causal model where an endogenous variable Y may

[3]See, e.g. the work done by Haavelmo (1944), R. Frisch ([1938] 1995), or Strotz & Wold (1960) in econometrics.

be written as a function of variables $X_1, ..., X_N$, which are its direct causes, in the following way:

$$Y = F_Y(X_1, ..., X_N) \tag{1.1}$$

Equation (1.1) is the model's structural equation. As a matter of example, imagine a scenario where one aims to model the causal behaviour of a toy spring; in this case, the dependency at play is captured by Hooke's law $F = kX$, which tells us that the elongation of the spring ΔX is a function of the force applied to the spring ΔF with a factor k.[4] The respective structural equation is thus $\Delta X = k^{-1}\Delta F$. In the interventionist framework, values (e.g. $F_1 = 3$ N) stand to variables (e.g. F) in a relationship of *determinates* to *determinables*, and causal relationships are always instantiated by particular individuals; e.g. the toy spring in front of me being stretched by a certain amount of force ΔF. Structural equations should thus not be understood as representing any average or aggregate response of the value of Y to various values of X, but as representing how the value of Y possessed by *some* individuals would change when the value of X possessed by these individuals also changes.

Finally, an observation: in the metaphysics of causation it is standard to think of causal relata as properties or events. Within interventionism, however, given that the causal structure of systems is described via causal models, it is commonplace to speak of causal relationships between variables or, to be more precise, values of variables. Woodward is never really clear about how to relate explanations between variables to explanations between events and properties, but since causal relationships are worldly dependencies one way to view variables would be à la Hausman, who defines these as features of the world denoted by symbols (Hausman 2005, p.34).[5] The motivation for talk of variables is that it is more in line with how scientists describe causal relationships: for instance, one may speak of the pressure exerted by a gas P, as causally depending on its temperature T and volume V. As we will see, this practice-based focus of interventionism will be a recurring theme throughout this thesis (see chapter 3 and chapter 5).[6]

[4]Note that the equation gets a negative sign if F is taken to represent the restoring force of the spring instead of the force applied to it. See Woodward (2003, p.39; p.112) for some more details on this example.

[5]In Woodward (2003, p.14) he says "I will often speak of causal relationships as obtaining between variables or their values, trusting that it is obvious enough how to sort out what is meant."

[6]However, this focus on variables will also create problems for the individuation of the systems interventionist causal models represent (see chapter 4).

1.2.2 What Are Interventions?

So far, I have briefly introduced the formal machinery interventionism uses, and only informally introduced interventions as surgical changes of the values of variables representing causes. The formal definition of interventions obeys a two-stage fashion. First, Woodward introduces the type-level definition of an intervention variable, after which he gives a token-level definition of interventions.[7]

Briefly, the difference between token-level and type-level causal claims is the following. Token-level (i.e. actual) causal relations link particular events: the alarm going off at 8 am today, 15 May 2020, caused me to get out of my bed. The alarm going off on a particular day at a particular time is a *token*. Type-level causal relations link kinds of events or properties.[8] The alarm going off at 8am every day (generally) causes me to get out of bed. In this example, the alarm going off at 8 am is a *type*. In a nutshell, the type vs token distinction can be interpreted in terms of the determinable vs determinate distinction. Let us now focus on defining intervention variables and interventions.

I is an intervention variable I for X with respect to Y iff the following criteria are satisfied (Woodward 2003, p.98):[9][10]

> **(IV)**
>
> **I1 – CAUSATION.** I causes X. If I is an intervention variable, there is a direct causal path from I to X.
>
> **I2 – ARROW-BREAKINGNESS.** I switches off all other interferences with X. When I attains certain values, X ceases to depend on the values of any other variables that were previously causes of X.
>
> **I3 – SURGICALITY.**[11] I has no other influences on Y other than through X. That is, with the exception of those causes of Y, if any, that are built into the $I - X - Y$ connection itself, I does not directly cause Y and it is not a cause of any causes of Y that are distinct from X.[12]

[7] Woodward attributes the term 'intervention' to Meek & Glymour (1994) and Pearl (2000).

[8] Since properties are not particulars located in space and time and can have many instances, kinds of events or properties can be used interchangeably (Hausman 2005, p.33).

[9] With the hope that the reader shall keep in mind that the notion of an intervention is a three-place relation, I shall hereafter omit "for X with respect to Y" when referring to intervention variables.

[10] 'IV' is short for 'intervention variable.' I1, . . . , I4 is Woodward's labelling of the conditions, the conditions' names are mine.

[11] Notion indebted to Pearl (2000, p.41).

[12] **I3** assumes 'modularity,' the property that any equation in a system of structural equations may be disrupted or replaced without any disruptions on the other equations which form the set (see Hausman & Woodward 1999 and Woodward 2003, Ch.7). Modularity was inspired by the notion

In other words, **I3** rules out causal structures like these ones:

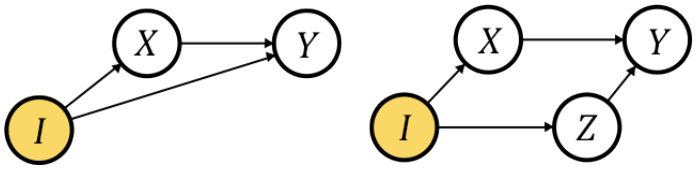


Figure 1.1: Causal structures ruled out by **I3**.

I4 – STATISTICAL INDEPENDENCE. I is statistically independent of any other variables Z that also cause Y.

In other words, **I4** prohibits the following causal structure:

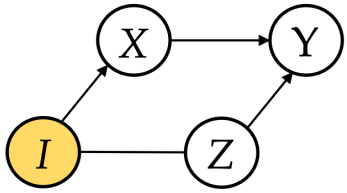


Figure 1.2: Causal structure ruled out by **I4**.

Regarding (IV), note that **I1** does not entail that the notion of an intervention is viciously circular, for it does not presuppose that there exists a causal relation between X and Y; the only causal information required to characterise the notion of an intervention on X is that there is a causal relation between I and X. Furthermore, **I1** is key to understanding why, since the notion of cause is ineliminable from Woodward's account of causation, it is a non-reductive account of causation.

Note as well that in the above characterisation 'cause' refers to 'contributing' rather than 'total' cause. The reason why the interesting notion of cause here is that of contributing cause is because the aim of interventionism is to capture whether a variable X influences another variable Y along a specific route even if (because of cancellation among different causal paths between X and Y) X has no total effect on Y. See Woodward (2003, pp.50-51) for a more detailed discussion of this topic.

Having defined intervention variables, an intervention on X with respect to Y will be defined as follows (Woodward 2003, p.98):

> **(IN)** I's assuming some value $I = z_i$ is an intervention on X with respect to Y iff I is an intervention variable for X with respect to Y and $I = z_i$ is an actual cause of the value taken by X.

of 'autonomy' in the econometrics literature, developed by e.g. Haavelmo (1944) or R. Frisch ([1938] 1995), which captures this idea that causal mechanisms can be modified independently from each other. A similar concept, the "wiping out of equations," can be found in Strotz & Wold (1960).

Note that (*IN*) defines a token-level causal notion, but the notion of contributing cause is a type-level notion. In chapter 4 I examine this type-token characterisation in some more detail. Note as well that what defines interventions is not that they be carried out by humans, but that the requirements (*IV*) are satisfied. As I discuss in §1.3.1 criteria (*IV*) are non-anthropocentric, since no reference to free will or any kind of agency is required in order to make sense of interventions.

In more qualitative terms, an intervention can be defined as a *wiggle* of the value of a variable representing a system's property, a process which cuts off said variable from other variables in the model which are putative causes. It is by tracing how said wiggle "percolates through the system" that we acquire information about the system's causal structure (Woodward 2014, p.78).

1.2.3 Causal Representation & Explanation

In §1.2.1 we have seen what the components of causal models are. In this section I focus on their representational function, which is key for the interventionist account of causal explanation. In this context, a question which arises is whether causal models are 'apt,' namely, whether they are adequate representations of the causal structure of the systems they model.

In Woodward's interventionist account of causal explanation, the question of aptness has epistemic and pragmatic overtones (Jansson 2018, p.256): judgments of aptness are constrained by prior experience/a posteriori knowledge of causal dependencies, and are relative to goals of inquiry and context.[13]

In this section, I summarise the interventionist's position with respect to aptness, which requires a compromise between 'what's out there,' what one knows, and pragmatic constraints. To do so, I split the analysis into the components of causal models, namely, their relata and their relationships. I then illustrate aptness via the archetypal example of inter-theoretic relations: the reduction of thermodynamics to statistical mechanics.

[13] Grounding interventionists like Schaffer (2016a) or Wilson (2018) would of course deny this; for them, aptness is just a matter of accurate representation of worldly causal structure.

1.2.3.1 Causal Relata

In interventionism, the causal relata *are* variables.[14] The question of what constitutes a good causal model is then largely dependent on the question of which variables ought to be incorporated into a causal model for some system; in other words, what makes a given set of variables good or better than others.

Woodward approaches the question of variable choice as both a descriptive and normative project, by offering some heuristic principles that do and ought to guide variable choice.[15] The framework within which the problem of variable choice is approached is one of 'means-ends': the goodness of a certain set of variables is assessed with respect to whether the resulting representations capture the causal explananda of interest (Woodward 2016, p.1048).[16] However, this normative aspect of interventionism does not undermine the objectivity of the resulting explanations. While goals of inquiry and context influence the choice of explananda (these determine which possibilities we shall take seriously), once explananda are fixed it is empirical considerations about causal structure that ultimately determine what explains what (Woodward 2010b, p.297). In Potochnik's words, "[i]f the process under investigation does not conform to some causal pattern, then that pattern cannot explain the outcome of the process" (Potochnik 2010, pp.224-225). This interplay between the pragmatic aspects of interventionist explanation – it is interest-relative and context-dependent – and its ontic aspect – it is mind-independent – will be illustrated in more detail in §1.2.3.3. For now, let us look at the basic ingredients of causal aptness.

Among the principles for variable choice in causal representation and explanation, we find the distinctness of relata, sparsity, specificity, proportionality, and stability.[17] Woodward thinks of these principles as 'heuristic' in the sense that they guide variable choice, though they might not necessarily yield a unique best

[14] See e.g. Woodward (2003, p.39): "[...] on a manipulability account of causation, it is most perspicuous to think of causal relationships as relating *variables* or, to speak more precisely, as describing how changes in the value of one or more variables will change the value of other variables." These variables, are "properties or magnitudes that, as the name indicates, are capable of taking more than one value" (ibid.). As explained in §1.2.1, variables are symbols by which we denote features of the world. According to this definition, one may prefer to say that the causal relata are the features of the world modelled as variables.

[15] See Woodward (2016, p.1053) for a defence of the descriptive claim.

[16] The term 'means-end' is taken from Hitchcock (2012), who coins it to characterise the tradition, going back to Kant, of "recasting metaphysical questions as questions about the utility of a conceptual scheme, heuristic framework, or methodological rule for achieving some particular end" (p.79)
'Interests' should not be understood narrowly as the subjective preferences of investigators for certain variables, but rather more broadly as the cognitive aims such as the provision of (objective) causal explanations.

[17] See Woodward (2016) for a non-exhaustive list of criteria for variable choice and their justification.

choice of variables. Let us briefly discuss the aforementioned criteria.

The criterion of 'distinctness' of the causal relata tells us that variables which are logically, conceptually, or constitutively related do not stand in causal relations. For instance, 'being a woman' and 'being a sister' cannot stand in a causal relation because these properties are conceptually related, as are 'redness' and 'scarletness' (the property of being scarlet).[18] Besides the criterion that causes are distinct, they need to be 'independently fixable,' which as the name indicates tells us to choose variables whose values we can in principle manipulate independently of the values taken by other variables (see e.g. Woodward 2015, p.316; Woodward 2016, p.1054).[19] Independent fixability is covered by the criteria (*IV*) specified in §1.2.2, in particular, by *ARROW-BREAKINGNESS* and *SURGICALITY*.

Next we have 'stability,' whose first cousin 'invariance' will be the focus of chapter 3. Stability captures the robustness of causal relationships under changes in background conditions; the idea behind it is that the more stable a relationship is, the more causal information it may provide, because it is more generalisable.

Another criterion that guides the construction of good causal models is 'sparsity,' the property of there being fewer variables, and consequently fewer causal relations. Sparsity is motivated by its connection to 'specificity': specific causal variables can be more easily exploited for manipulation and control. For instance, the connection between oxygen transportation in the blood from the lungs to the tissues and hemoglobin is highly specific. The aforementioned connection between sparsity and specificity may be illustrated via another biological example, the dependence between enzymes and their substrates: the fewer substrates an enzyme can bind, the greater its specificity (Woodward 2010b, p.309).

While sparsity is desirable, one must be wary that any given causal representation does not include too few variables either, because an overly coarse-grained representation might lead to the identification of non-existent causal relations; vice versa, a too fine-grained representation might omit existent causal dependency relations. This tension is resolved by the inclusion of the criterion of 'proportionality,' which Woodward borrows from Yablo (1992). Along general lines, proportionality is the criterion that causes are *just enough* for or *line up with* their effects, in the sense that they neither omit relevant details (they are not overtly coarse-grained) nor are they so specific that they include irrelevant details (they are overtly fine-grained)

[18] The distinctness of the causal relata is a criterion which is not specific to interventionism. For instance, D. K. Lewis (1986b), for whom events are spatiotemporal regions, thinks that overlapping events are not sufficiently distinct to be causal relata.

[19] Whether DAGs and the interventionist framework are applicable to non-causal dependencies and the criterion of independent fixability applies in those contexts is something I don't consider here. See Woodward (2015) and Jansson (2018), as well as Schaffer (2016a) and Wilson (2018), for a discussion of this question.

(Woodward 2021, p.243). Therefore, proportionality helps us select the right *level of grain*.

As an example, consider Yablo's (1992) famous example of the pigeon trained to peck at targets of any shade of red. When the pigeon is presented with a scarlet target, it pecks. Here we can construct two possible causal models of the situation. Model 1 has two variables, RED: {red, not red} and PECK: {peck, not peck}. Model 2 is a fine-grained version of the first, where the variable RED is replaced by SCARLET: {scarlet, not scarlet}. Note that model 2 omits certain dependencies, in particular those causal dependencies that exist between red but not scarlet targets and pigeons pecking. As a result, model 2 is overtly specific, and so SCARLET fails to be proportional to PECK. Model 1 is thus better that model 2. Last, consider a model (model 3) where the variable SCARLET is replaced by the variable COLOUR: {scarlet, cyan}. COLOUR is not a good variable because it is non-exhaustive, i.e. there are other colours it could take. Proportionality requires that the variables included in a given causal model *exhaust* the possibility space (ibid., p.249). [20] From these three models, model 1 is the most proportional and, therefore, the best causal model of the situation.

The relativity of variable choice in interventionism has been one of its most contested aspects. Strevens, for instance, raised this issue in his review of *Making Things Happen* (Strevens 2008), where he argues that this is a damaging sort of relativity, for it may lead to spurious explanations. Consider that drinking bottled water, BW: {0, 1}, and dying of heart disease, HD: {0, 1}, are correlated because they share the consumption of salty food, SF: {0, 1}, as a common cause. Whether one includes SF as a variable in their causal model of the system affects whether BW and HD are causally related; they are if SF is not included, but they aren't if it is (ibid., pp.242-244). Woodward defends variable relativity by arguing that, first, it is not a feature that's specific to interventionism; it equally affects other apparatuses of causal representation. Second, it is explanatory relativity of an innocuous kind (Woodward 2003, p.56; p.89): relativity with respect to the level of grain with which we describe physical systems.

Statham has offered an extensive analysis of the charge that interventionism is affected by a harmful variable relativity in her paper 'Woodward and variable relativity' (Statham 2018). There, she claims that while interventionism doesn't suffer from 'relativity to a set of system variables,' thus countering Strevens' charge, it is affected by 'relativity to background conditions.' She argues that, if strengthened – she proposes that the values of the variables which are not part of the explanation

[20] Model 3 is an example taken from Franklin-Hall (2016).

be fixed at their actual values – this relativity to background conditions is good, for it avoids the over-permissiveness of the relativity to sets of system variables. Furtermore, it aligns with our intuitions about these cases: for instance, we want to say that Covid-19 will only be a cause of death in unvaccinated individuals and/or those with underlying conditions, not a cause of death simpliciter.

One common way to handle variable relativity is to appeal to notions of naturalness. D. K. Lewis (1983) or Sider (2011) have notoriously adopted such a strategy.[21] However, Woodward has rejected this metaphysical criterion on the basis of the following concerns. First, and most importantly, judgements of naturalness obliterate the pragmatic aspect of explanation: which properties count as natural is dictated by metaphysical considerations, and so it is independent of the goals of inquiry and the context of the explanation. Second, there is the epistemic constraint imposed by fundamental physics: if it is fundamental physical theories which are meant to tell us which properties count as perfectly natural, we wouldn't be able to identify the perfectly natural properties until a unified theory of fundamental physics were proposed. Last, there is the related concern that even in the event that fundamental natural properties were identified, without an account of the emergence/reduction of higher-level properties from/to fundamental properties naturalness would not be of much help to identify and describe higher-level properties. See Woodward (2016, pp.1056-1057) for a detailed response to the naturalness solution to the problem of variable relativity. Perhaps there are ways to defend that higher-level naturalness may be possible even in the absence of knowledge of lower-level naturalness and hence respond to Woodward's second concern about naturalness. However, I think the first concern, namely, that naturalness does not sit well with the pragmatic aspect of interventionist explanation, is a forceful reason to abandon naturalness as a criterion for variable choice.

1.2.3.2 Causal Relations

In the previous section, I briefly discussed the complicated task of variable choice. Recall that one such guiding criterion was sparsity: if finding sparse representations is a constraint on the construction of causal models it is because it is difficult to find variable aggregations which preserve screening-off conditions.

Screening-off conditions are important because they enable the isolation of causal from merely correlational dependencies. Consider the following philosophical stock example: we have a causal structure in which a drop in atmospheric

[21] One might also check Knox (2016) for a physics-based example of how to reduce thermodynamical properties to statistical mechanical properties in terms of naturalness.

pressure P causes a dip of the mercury in the barometer's reading B and the onset of a storm S. While B and S are correlated, they do not cause each other – rather, they are effects of a common cause. P screens off S from B (namely, renders B conditionally independent from S) because $P(S|P \cup B) = P(S|P)$. This screening off can e.g. be observed when the barometer's needle is kept fixed (there is no dip in the mercury level). Provided the atmospheric pressure drops, there is a storm; the probability of there being a storm depends only on the probability of a drop in atmospheric pressure.

In the probabilistic version of the causal modelling framework, the screening-off is provided by the *CAUSAL-MARKOV* (*CM*) condition, which was first described by Kiiveri & Speed (1982), and which Hausman & Woodward (1999, p.523) define as follows:

> **CM:** For all distinct variables X and Y in the variable set V, if X does not cause Y, then $Pr(X/Y\&Parents(X)) = Pr(X/Parents(X))$.

Here, $Parents(X)$ is the set of all the direct causes of X in the set V. In technical terms, CM entails that children in a causal graph are conditionally independent of their non-descendants (non-effects) given their parents. In less technical terms, CM establishes that the values of variables depend only on the values of their direct causes or parents: once the values of X's parents have been specified, all other variables in the set become irrelevant for predicting X except for X's descendants. Note that this is a generalisation of Reichenbach's notion of common cause (Reichenbach 1956), which can be summarised under the slogan 'no correlation without causation.'

The converse condition for CM is the *CAUSAL-FAITHFULNESS* (*CF*) condition, also called 'stability' (Pearl 1988) or 'no fine-tuning' (P. W. Evans 2018). This condition roughly says that the joint probability distribution for a DAG satisfies only the (conditional and unconditional) independencies implied by the causal structure the model represents. Hence, it has the role to restrict the values the variables in a DAG may take in order to prevent accidental independencies; i.e. independencies not due to the causal structure.

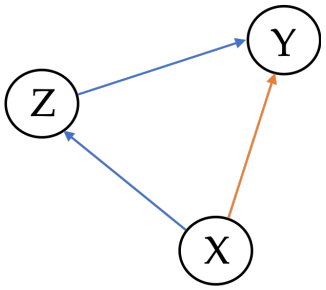


Figure 1.3: Fine-tuned causal structure: the causal paths $X \to Z \to Y$ and $X \to Y$ cancel each other out.

CF forbids situations such as the one represented in fig. 1.3, where one causal path (the blue) exactly cancels the influence on Y along the other path (the orange). In other words, *CF* prevents that certain parametrisations of a given graph lead to independencies which other parametrisations don't yield.

The *CM* and the *CF* are the central methodological principles of the causal modelling framework, and thereby interventionism, because they respectively constrain the dependencies and independencies in causal models. The *CM* is the most fundamental compatibility principle, for it ensures that a given probability distribution represents a causal structure. In this respect, the *CF* is somewhat less central than the *CM* because its failure doesn't jeopardise the whole causal modelling enterprise as a failure of the *CM* would. A failure of the *CF* might only indicate an untraceable causal relation, and thus foreground our epistemic access to causal relations, but does not necessarily mean that there is no causal structure.

The *CM* and *CF* do not always uniquely identify causal structures, so other assumptions such as causal minimality (the graph is as sparse as possible) or causal sufficiency (all common causes have been included) are required to eliminate the underdetermination which may affect DAGs (Hitchcock 2021). Furthermore, failures *CM* and *CF* have been the focus of an intense debate as to whether these failures should be interpreted as a failure of causal explanation altogether.[22] Since all these general conditions for the aptness of causal models have been extensively discussed in the literature, in §1.3 I focus on an overlooked constraint on aptness specific to interventionism, the *possibility* of performing interventions on the mod-

[22] For instance, Cartwright (1999, 2002) has argued that *CM* is not satisfied for probabilistic causes, and Andersen (2013) has argued that many equilibrium-maintaining mechanisms in the special sciences won't satisfy *CF*. One interesting case study in this respect is quantum entanglement: there is no consensus as to whether there is no causal explanation for it (Woodward 2003), or it is a causal phenomenon which does not satisfy *CM* (Glymour 2006), or *CF* (Näger 2016). There have also been proposals to quantise the classical causal modelling formalism (Costa & Shrapnel 2016; Wood & Spekkens 2015), or to even abandon any causal explanation of quantum entanglement in favour of a non-causal common ground explanation (Ismael & Schaffer 2020). For more details on this case study, see Iranzo-Ribera (2020).

els' variables. But before then, let me illustrate how these aptness conditions for variables and causal relations may come together.

1.2.3.3 Levels of Explanation

We cannot end this section without discussing a topic at the core of scientific explanations: the so-called 'levels of explanation,' also known as 'levels of reality' if seen under a realist lens. Considerations of level play a key role in selecting the explananda and explanantia variables. For instance, if the explanandum is how *individual* neurons generate spike trains with particular temporal features, then it is the causal details of the anatomy and molecular constitution of the individual neurons which are relevant for the explanandum, but if instead one is interested in an explanation of the *collective* dynamics of neuronal cultures, then it is the dynamics and connectivity of the whole culture instead of details about individual neurons that will be relevant to this new explanandum (Iranzo-Ribera 2015).

Proportionality was highlighted as the principle to select the right variable sets (see §1.2.3.1). While introduced with upper-level explanations in mind (Woodward 2021, p.261, fn.30), proportionality may help us select the right level of explanation, together with 'conditional irrelevance,' a screening off conditional similar to CM (see §1.2.3.2), which may be described as follows:[23]

> **UNCONDITIONAL (IR)RELEVANCE:** a set of variables X_i is *unconditionally (i)relevant* to [explanandum] E if there are (no) some changes in the values of each X_i when produced by interventions that are associated with changes in E.
>
> **CONDITIONAL IRRELEVANCE:** a set of variables Y_k is irrelevant to variable E *conditional* on additional variables X_i if both the X_i and the Y_k are respectively unconditionally relevant to E *and* conditional on the values of X_i, changes in the value of Y_k are unconditionally irrelevant to E.

As an illustration of how these notions apply to real scientific cases, consider the archetypal reduction example: the reduction of thermodynamics to statistical mechanics, respectively macro-level and micro-level descriptions of large physical systems. Take increases in temperature T as a cause of the increase in the pressure P of the gas. In this case, we can say that T is *unconditionally relevant* for P, because P depends on T (the converse also holds).

[23] Definitions adapted from Woodward (2021, pp.253-254).

Consider now $\rho_i = (\vec{p_i}, \vec{q_i}) = \sum_{j=1}^{j=3} (q_{ij}, p_{ij})$, a variable which describes the position and momentum of each individual particle, a point in phase space (where in a gas with N particles, $i = 1, ...N$). This variable ρ_i is also *unconditionally relevant* for P, because the motion of the particles affects P. However, and crucially, conditional on T – that is, once we know the temperature of the gas – ρ_i becomes *conditionally irrelevant* for P, because changes in the values of ρ_i which don't alter T (e.g. swapping the descriptions for two particles) make no difference to P. For distinguishable particles, each point in phase space will be a microstate of the system, and as Robertson (forthcoming, p.10) argues, they can be seen as equivalence classes for a certain macrostate (P, V, T).[24]

For thermodynamic systems, the conditional irrelevance of the micro-level details for macro-level explananda explains why, e.g., T is a much better choice of variable than ρ_i to explain changes in P. Besides illustrating an application of the criteria for aptness, hopefully this example also shows how the pragmatic and ontic aspects of interventionist explanation come together: while causation is 'variable-relative,' it is not 'description-relative,' since whether a variable is causally relevant to another does depend on which variable set is selected, but *not* on how these variables are described (Woodward 2007, p.90). Furthermore, it provides some context for chapter 3, which will discuss fundamentality relations between laws.

Having provided a comprehensive summary of interventionism, qua theory of causation and causal explanation, we are now in the position to introduce an overlooked question in discussions of the aptness of interventionist causal models and interventionist explanation, a question which is the backbone of this thesis: what is meant by interventions being *possible*. Interventions being possible is key to elucidate the meaning of counterfactuals, hence the importance to tackle this question.

1.3 Interventionism & The Possibility Question

Interventionism is formulated in term of counterfactuals about what would happen under *possible* interventions. While the formal criteria for intervention variables in §1.2.2 have been extensively scrutinised since the publication of Woodward's *Making Things Happen*, I believe the question of what sort of possibility is required

[24] For *indistinguishable* particles, each microstate will have $N!$ points, corresponding to all the permutations of the particles.

of possible interventions has been insufficiently examined.[25] This is the 'possibility question' (PQ), to which I now turn. In §1.3.1 I discuss how agency features (or rather, does not feature) in the PQ. Then, in §1.3.2 I examine the PQ by looking at the infamous case study of how the Moon's position affects the Earth's tides. There, I contest Woodward's argument that the relevant notion of possibility for interventions is not physical possibility but conceptual possibility. In §1.3.3, I show how conceptual possibility is in tension with the avowed causal realism of interventionism.

1.3.1 Agency and Possibility

Consider a study which aims to settle whether there is any causal connection between treatment with a newly developed drug and the alleviation of a disease's symptoms. Treating a group of research subjects with such a drug constitutes an intervention. Similarly, in the study of how the amount of water available to tomato plants affects their growth, sudden rain is another example of an intervention. Interventions so considered are none other than manipulations of systems – be it by human beings or nature – which satisfy criteria (*IV*). Woodward refers to these as *experimental* interventions (Woodward 2003, p.15; Woodward 2015, p.310). In this section, I discuss how the notion of 'intervention' is (or rather, is not) connected to agency.

The drug-disease example is a good starting point, as it allows us to draw on useful distinctions which are already in place in fields such as epidemiology, clinical medicine, psychology, statistics, etc. Research in these fields relies on two kinds of studies: *observational studies* (OSs), and *experimental studies* (ESs). Broadly speaking, the main difference between them is the degree of agency or involvement of the researcher: while in OSs the researcher merely observes the dependency between variables, e.g. a treatment, a risk factor, etc. and their effect (that is, the researcher is passively involved in the experiment), in ESs the researcher *alters* the dependency between these (the researcher is actively involved in the experiment). In short: in OSs there is no manipulation, just observation of the natural variation in a population. Contrarily, in ESs researchers introduce interventions and study their effects.

Another distinction between OSs and ESs is that the latter are usually randomised (e.g. randomised control trials, or 'RCTs'), which means that subjects are randomly assigned to two or more groups, one of which is the control group,

[25] With few exceptions. To the best of my knowledge, these are Reutlinger (2013), Gebharter (2017), and Andreas & Casini (2019).

which e.g. receives the placebo instead of the treatment whose efficacy is under study. OSs are useful when randomisation is not possible for ethical reasons (e.g. giving cigarettes to a group of people over decades, since this would be harmful for their health) or technological reasons (e.g. a brain transplant). However, OSs don't allow to control for potential confounders, and so they are generally not a good tool for causal inference. The exception are *natural experiments*, a specific type of OSs where an event or situation allows for a (seemingly) random assignment of study subjects to different groups. This is the case of e.g. Mendelian randomisation, where genetic makeup (presence or absence of a certain gene) is used to identify factors that influence the risk of certain diseases. Policy changes, weather events, natural disasters, etc. are other situations that create appropriate circumstances for natural experiments (see 'Natural Experiment'). Surely, causal notions play a role in these cases even if they are not controlled studies and humans are not involved. This is precisely a feature Woodward (2003, p.103) wants to capture with (*IV*) and (*IN*): that interventions need not be caused by humans, but could be caused by natural phenomena of any kind. By acknowledging that interventions are not limited to human capabilities interventionism distances itself from the anthropocentrism of the previous manipulability theories of von Wright (1971) and Menzies & Price (1993), which assign a central role to human agency – the reason why they are labelled 'agential theories of causation.'

Given intervention variables need not represent human agents, *practical* possibility is not a suitable marker for the type of possibility behind interventions. Correspondingly, 'impossibility' does not refer to a lack of capabilities or technological means. In §1.3.2 I investigate what notion of possibility is at play in interventionism. While human beings will naturally be interested in those causal relations they *are able to* discover (or for which they can at least have some evidence), manipulate, and control, interventionism – which given the fact that it is theory which aims to put causal information and reasoning at the service of our goals and purposes (Woodward 2014, p.692) has been called a 'functional' approach to causation – is a metaphysically realist theory. As I explain in §1.3.3, for the interventionist about causation causal relations are out there in the world regardless of our discovery enterprises.

A final point to take into account is that interventions *can* but *need not* be actual. A crucial aspect of Woodward's framework is that, as a counterfactual theory, it only requires that there are hypothetical interventions, namely, non-actual but possible manipulations of a variable's value. Thus, the meaning of causal claims may be spelt out via actual or hypothetical experiments (Woodward 2003, p.114). The sense in which interventions on systems in RCTs or natural experiments

are possible is obvious: they take place, so they are possible. As mentioned above, Woodward sometimes calls this type of interventions 'experimental,' but talk of experiments has many connotations: that intentions be involved, that conditions be controlled, that observers are present, etc. Since interventions are not anthropomorphically characterised, to avoid confusions in the remainder of this chapter I will call the interventions which take place 'actual.' The contrasting class is, as mentioned, 'hypothetical' interventions, i.e. possible but non-actual interventions.

To sum up, the sense in which interventions are possible can't be reduced to practical possibility because interventions need not be carried out by humans. The possibilities we are after are objective possibilities, yet from a functional perspective we focus on those objective possibilities which are epistemically accessible and epistemically interesting to us. Finally, interventions need not be actual. Since it is an obvious one the sense in which actual interventions are possible, the PQ is essentially a question about the possibility of hypothetical interventions.

1.3.2 The Moon-Tides Case Study

In the following, I focus on an often neglected example by Woodward himself in *Making Things Happen* (Woodward 2003), which I will hereafter refer to as the **Moon-tides** example, in order to discuss the ways in which one may address the PQ. Consider the following causal claim (adapted from Woodward 2003, p.129):

> **Moon-tides:** Changes in the *distance* of the Moon with respect to the Earth, through corresponding changes in the gravitational attraction exerted by the Moon on various points on the Earth's surface, cause changes in the *magnitude* of the tides.[26]

The causal claim **Moon-tides** describes the phenomenon of tidal forces exerted by the Moon on the Earth, which deform the Earth's surface and cause the rising and falling of the water in the oceans, as shown in fig. 1.4.[27]

[26] The italicised words point out the changes with respect to Woodward's original wording of the example. 'Distance' has replaced 'position,' as he clearly considers modifications of the relative distance between the two bodies and not of the way the Moon moving around in its orbits changes the tides in a particular place. Similarly, in the second part of the generalisation I have replaced 'motion' with 'magnitude' to make the claim more precise. In the original text 'Earth' is not capitalised.

[27] Notably, the effects of ocean tides are way larger than the effects of land tides, so this is probably the reason why Woodward focuses on the former.
As an interesting datum regarding the tides, note that at the surface of the Earth, the gravitational force of the Moon is about 2.2 times greater than that of the Sun (see 'Tide').

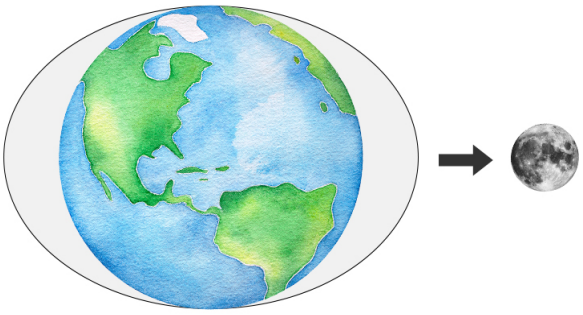


Figure 1.4: Tidal bulges due to gravity (the right one, facing the Moon) and inertia (the left one, on the opposite side).

These tidal forces are a result of the force imbalance between the gravitational force exerted by the Moon on the Earth and the inertial or centrifugal force produced by the Earth's 'rotation' around the common center of mass or barycenter of the Moon-Earth system (which can be considered an inertial frame of reference). While the inertial field is uniform, the gravitational field is not, and this is what causes this force imbalance. This force imbalance is manifested at the Earth's surface as follows: there is a net tide-producing force on both the hemisphere of the Earth turned towards the Moon (or in other words, the nearest point to the Moon) and the hemisphere turned opposite the Moon (the furthest point), acting respectively in the direction of and away from the Moon, as seen in fig. 1.5:



Figure 1.5: Tidal forces on the Earth's surface at baseline sea level created by the Moon's gravitational pull (OpenStax University Physics).

Figure 1.5 represents the Earth seen from the North Pole. You should imagine the Moon to be all the way to the right or left, where the tidal forces at the ocean surface are the strongest. These forces would diminish to zero as you approach the Earth's centre. Gravitational forces follow an inverse-square law, but tidal forces, since they are related to their gradient, are inversely proportional to the cube of the distance (Ferrari & Joyce 2004). Note that centrifugal forces are inertial

or pseudo forces which feature only in non-inertial frames of reference: in this case, the frame of reference that rotates with the Moon-Earth center of mass. Since we are interested in tidal forces on the Earth, the geocentric frame of reference is the most natural choice for this explanation.[28]

Moon-tides is a true causal claim because changes in the relative distance between the Moon and the Earth *do* cause changes in the tidal force exerted by the Moon on the Earth. But how do we know so? If upon intervening on the relative distance between the Earth and the Moon the magnitude of the tides changes accordingly, these two properties of the Earth-Moon system are causally related. The simplest and most general possible graph of the situation is:[29]



Figure 1.6: Simplest causal graph for an intervention which tests **Moon-tides**.

Here, r_{M-E} is the relative distance between the Moon and the Earth, and $TIDES_E$ the magnitude of Earth's tides. The intervention illustrated in fig. 1.6 would have to satisfy the requirements for intervention variables (*IV*) listed in §1.2.2 and *be possible in a non-trivial way*. The PQ is the question of what the relevant notion of possibility which characterises interventions is, and thus aims to characterise the *modal* aspect of interventions. In the following, I explore the PQ in the context of the scenario described by **Moon-tides**.

1.3.2.1 Physical Possibility

Applied to the Moon-tides case study, the PQ reads as follows: in which sense are changes in r_{M-E} which change $TIDES_E$ possible? Among the most fitting candidate notions of possibility, we find physical possibility, metaphysical possibility, conceptual possibility, or logical possibility. Given that we are dealing with an example of a physical phenomenon, there is reason to expect interventions to be physically possible. Woodward (2003, p.128) himself opens up this line of inquiry by entertaining two kinds of physical possibility, which I respectively call 'Strong Physical Possibility' (SPP) and 'Weak Physical Possibility' (WPP). According to

[28] Nonetheless, should one dislike the *fictitious* nature of inertial forces, an explanation of the Earth's tides from an inertial frame of reference is also available.

[29] For a more realistic pictures other variables would have to be introduced; e.g. the strength of the gravitational force, the mass of the oceans, etc.

each of these types of physical possibility, an intervention is physically possible iff it is consistent with:

> **Strong Physical Possibility (SPP):** the actual laws of nature and the actual initial conditions (ICs).
>
> **Weak Physical Possibility (WPP):** the actual laws of nature and some range of ICs consistent with the actual laws of nature including the actual ICs.

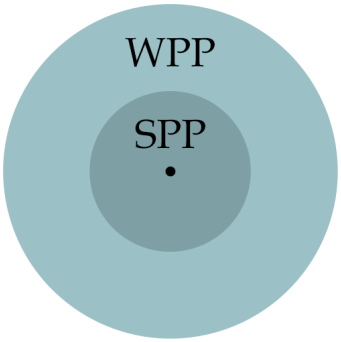


Figure 1.7: SPP and WPP delimit concatenated spaces of physical possibility.

For the sake of clarity, in the definitions above *actual* means *actually obtaining*. Note that both SPP and WPP construe physical possibility as a combination of nomological possibility and ICs, where the ICs are the "specifications of the values taken by variables figuring in a law when it is applied to particular systems" (Woodward 2013, p.61).[30] This categorical distinction between laws and ICs goes back to at least Wigner (1967), who roughly suggests the following characterisation: "the [world's] complications are called initial conditions; the domains of regularities, laws of nature." (ibid., p.3) As we will see in chapter 3, this distinction between laws and ICs serves as the basis of Woodward's invariance-based account of laws. Note as well that, since both types of physical possibility are construed as versions of nomological possibility, they both rule out interventions which are not allowed by the laws, such as moving faster than the speed of light or transforming all the heat produced by a system into work.

Woodward discards SPP as the relevant notion of possibility which characterises interventions on the basis that, if conjoined with determinism, it is too strong because it implies that interventions would not be possible unless they

[30] SPP and WPP also suggest that Woodward implicitly thinks of causation as bound up with time's arrow. In this respect, the types of laws that can explain change-relating generalisations are dynamical laws.

actually occur. As a result, counterfactuals associated with non-actual interventions would come out vacuously true. One option here would of course be to reject determinism, but in order to avoid getting tangled up in a debate about the truth of determinism (I presume!) Woodward opts for weakening SPP to WPP. Unlike SPP, WPP allows to label counterfactual scenarios such as the withholding of treatment with a newly developed drug from subjects (when in fact we gave it to them), or the lack of rain to provide hydration for our tomato plants (when in fact it did rain and they did not dry out) as physical possibilities regardless of the truth of determinism. These counterfactual scenarios require that "... initial conditions at some earlier time had been different, but given these different initial conditions no further violation of physical law would be required" (Woodward 2003, p.128).

Despite this plausible weakening of physical possibility, Woodward eventually rejects WPP due to its inability to capture the relevant kind of possibility concerning interventions, a rejection which he justifies on the basis of a generalisation from the Moon-tides case study introduced and discussed in this section (ibid., pp.130-131). In the following section, I dispute Woodward's argument for the rejection of WPP as a satisfactory answer to the PQ, an argument which motivates his turn to conceptual possibility as the right modal notion underlying his interventionist theory of causation.

1.3.2.2 The Rejection of Physical Possibility Argument

Woodward's **Rejection of Physical Possibility (RPP)** argument draws on an analysis of putative interventions on the Moon-tides scenario.[31] Imagine a hypothetical intervention that would double the relative distance between the Moon and the Earth as follows:

$$\mathbf{MT-I}: r_{M-E} = r_{M-E_A} \rightarrow r_{M-E} = r_{M-E_H} = 2r_{M-E_A} \tag{1.2}$$

In expression (1.2), subscript 'A' stands for 'actual' value and subscript 'H' stands for 'hypothetical' value. They respectively correspond to a variable's pre-intervention and post-intervention values in hypothetical interventions.

The conclusion of RPP is that MT-I is not physically possible, provided physical possibility is exhausted by WPP. Recall that a *weakly physically possible* intervention is one that is consistent with the actual laws of nature and some set of possible ICs, including the actual ICs. To see why Woodward argues that MT-I

[31] My naming.

is not physically possible, let us start by examining two realistic ways in which, according to him said intervention could occur (Woodward 2003, p.129).

One possible way in which such intervention could take place would be through the change in the position of another body, say Mars, which by coming closer to the Moon would thereby alter its orbit. Let us call this version of the intervention MT-I 'MT-I_1.'

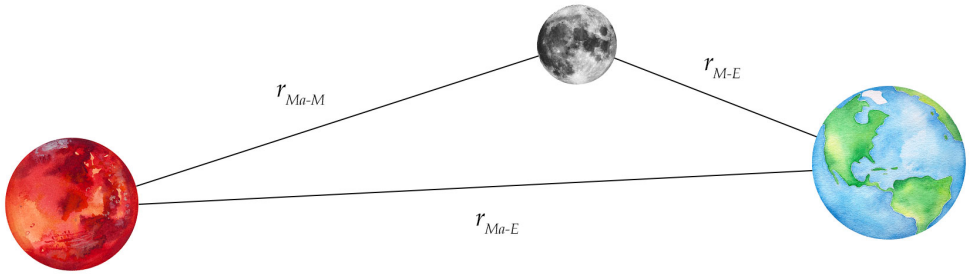


Figure 1.8: System comprised of Mars, the Moon, and the Earth.

Recall that the gravitational interaction between two bodies in classical settings is described by Newton's law of universal gravitation:

$$F_{ij} = -G\frac{M_i M_j}{r_{ij}^2} \tag{1.3}$$

In eq. (1.3), M_i and M_j are the masses of the two bodies, r_{ij} the distance between them, and $G = 6.67x10^{-8}cm^2/g\ s^2$ the gravitational constant.

MT-I_1 would change Mars' position (r_{Ma-M}) in such a way that it would affect the gravitational interaction between Mars and the Moon, and thereby the Moon's position. In turn, this would affect the value of the magnitude of the Earth's tides ($TIDES_E$). Furthermore, it would also directly affect the tides via a change in the position of Mars with respect to the Earth (r_{Ma-E}), which would affect the gravitational interaction between Mars and the Earth. Figure 1.9 below represents said intervention, where r_{Ma-M} is the intervention variable, r_{M-E} the variable intervened upon, and the magnitude of the tides, $TIDES_E$, the variable that changes as an effect of the intervention.

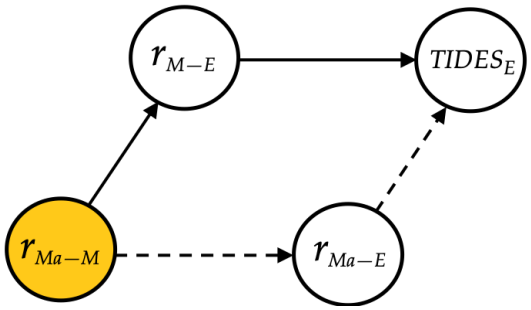


Figure 1.9: Causal graph for MT-I_1.

Another possible way the intervention MT-I could come about would be through the impact of a massive body, say a planetoid, on the Moon. Let us call it MT-I_2. This situation is described by fig. 1.10 below: the relative distance between the planetoid and the Moon (r_{p-M}) is the intervention variable, which like in the previous scenario causes the relative distance between the Moon and the Earth (r_{M-E}) to vary, which then causes the magnitude of the tides $TIDES_E$ to vary accordingly.

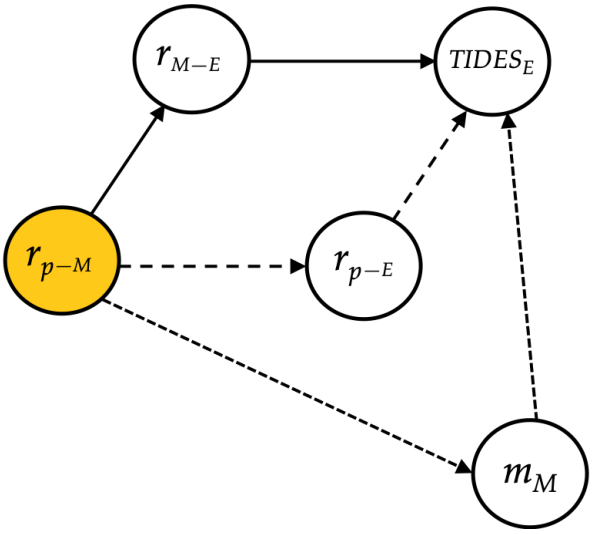


Figure 1.10: Causal graph for MT-I_2.

So far, so good. Nonetheless, there is a problem with these interventions: they are "too ham-fisted."[32] Both a change in Mars' or the planetoid's position with respect to the Earth would result in a direct change in the gravitational field at the Earth's surface not directly caused by a change in the relative distance between the Moon and the Earth, in violation of $SURGICALITY$. For instance, consider MT-I_1: an intervention to change the value of r_{M-E} via a change on r_{Ma-M} would also change

[32]This term is due to Elliott Sober, in correspondence with Woodward (as acknowledged in Woodward 2003, p.129).

r_{Ma-E}.

Additionally, Woodward argues, it is likely that the impact of a massive body like a planetoid on the Moon would not only change its position, but also its mass distribution; maybe a part would come off, or the center of mass would be displaced. In this scenario $TIDES_E$ would not only change because of a variation of the values of r_{M-E} or r_{p-E}, but also because of the change in the Moon's mass distribution m_M, which constitutes a violation of $SURGICALITY$.

Similarly, these physically possible interventions on the Moon's position would neither fulfill $ARROW\text{-}BREAKINGNESS$, as they would not erase the gravitational influence of other astronomical bodies on the Moon unless fine-tuned (see M. Frisch 2014, pp.81-82). It could be the case that the additional force modified the gravitational interactions between other planetary objects and the Moon in such a way that their effects would be effectively masked; i.e., that the variation of the net force of the other gravitational interactions with the Moon would be null. Or in terms of $SURGICALITY$, it could be the case that an intervention I on X influenced Y through paths other than $I - X - Y$ which exactly cancelled each other out, giving the appearance of a perfect fulfilment of $SURGICALITY$.

Although possible, this fine-tuned scenario seems highly improbable, as it would require exact compensation of the different force contributions (namely, module and direction of the different force vectors). In short, there is no guarantee that all interventions will respect the intervention conditions (IV). In light of this, Woodward concludes that since it is not guaranteed that there will always be physically possible interventions, the relevant notion of possibility the interventionist is after can't be physical possibility (Woodward 2003, pp.128-131).

As a first attempt, we might formally reconstruct the RPP argument as follows:

RPP:

P1. All genuine cases of causation can be characterised via interventionism.

P2. If a genuine case of causation can be characterised via interventionism, then the associated interventions are possible.

P3. **Moon-tides** describes a genuine case of causation.

P4. MT-I_1 and MT-I_2 are amongst the interventions associated with **Moon-tides**.

P5. MT-I_1 and MT-I_2 are not possible in the sense expressed by WPP.

C. Physical possibility is unsuited to characterise genuine cases of causation.

Premise P1 describes causal monism. While Woodward has softened his position

regarding this point, in *Making Things Happen* (Woodward 2003) he clearly advocates causal monism. Note that P1 is, however, not strictly necessary: were one a causal pluralist, the following version of C would obtain: 'Physical possibility is unsuited to characterise genuine cases of *interventionist* causation.'

P2 is built into the theory: X is a cause of Y iff there is a possible intervention on X that changes Y (Woodward 2003, p.45).

Following Woodward, I take P3 to be true.[33] We have seen that Woodward only considers two types of physical possibility, SPP and WPP, and granted the claim that the intervention MT-I associated with such case, implemented as either MT-I_1 or MT-I_2, is not even possible in its weakest form, namely WPP, P4 is also true. As it stands, its conclusion that physical possibility is unsuited to characterise genuine cases of causation is, however, logically invalid.

WPP is a particular type of physical possibility, and so without having exhausted all the types of physical possibility there can be, from WPP's unsuitability to characterise interventionist causation it cannot be concluded that the more general notion of physical possibility simpliciter is wrongheaded. Fortunately, this can be easily remedied by adding an extra premise which states that there is no form of physical possibility weaker than WPP and thus warrants the validity of RPP. The result is a valid argument, RPP*, which takes the following form:

RPP*:

P1. All genuine cases of causation can be characterised via interventionism.

P2. If a genuine case of causation can be characterised via interventionism, then the associated interventions are possible.

P3. **Moon-tides** describes a genuine case of causation.

P4. MT-I_1 and MT-I_2 are amongst the interventions associated with **Moon-tides**.

P5. MT-I_1 and MT-I_2 are not possible in the sense expressed by WPP.

P6. If WPP does not permit interventions, no other type of physical possibility does.

C. Physical possibility is unsuited to characterise genuine cases of causation.

Even if valid, RPP* conceals some unstated assumptions under which premises P3, P5, and P6 may turn out false. In §1.3.2.3 I lay out these assumptions, which threaten the soundness of RPP*.

[33] See Jansson (2014) for a different opinion: Jansson argues that the resulting explanation is not causal but law-based.

1.3.2.3 Contesting the Rejection of Physical Possibility Argument

Consider RPP* anew (see §1.3.2.2). First, note that the implementations considered for the MT-I are defined with respect to classical mechanics: we are only considering which interventions are possible according to a Newtonian framework. Woodward states it very clearly in the following passage:

> **W1.** "Although it may be true that any *actual physical process* that changes the position of the Moon will also directly influence the tides, *Newtonian theory* and *familiar rules about the composition of forces* tell us how to subtract out any direct influence from such process so that we can calculate just what the effect of, say, doubling the moon's orbit (and no other changes) would be on the tides, even though it also may be that there is no way of actually realizing this effect alone" (Woodward 2003, p.131).[34]

Unless specified, nothing in the argument precludes the existence of interventions with respect to more fundamental physical theories that do satisfy the intervention requirements (*IV*). Our world is not Newtonian and it is not obvious what we are talking about when we consider 'causation in a Newtonian world.' Correspondingly, it is not obvious which laws we should be basing the relevant kind of physical modality on. There could be interventions which are not consistent with the laws of Newtonian mechanics but are consistent with the laws of quantum mechanics or by some more fundamental and yet unknown theory of quantum gravity that do fulfill (*IV*). Imagine e.g. an intervention on r_{M-E} via quantum teleportation, which would cause the Moon to move from its initial to its final position in small jumps.[35] Given that physical possibility is essentially nomological possibility, in order to assess whether WPP permits interventions on a given system it ought to always be specified with respect to which laws the relevant causal claims have to be evaluated; it is considerations like those discussed in §1.2.3 that help determine the appropriate level of grain. Furthermore, P6 stipulates that WPP is the weakest form of physical possibility, but it need not be so. In chapter 3 I explore this avenue by proposing an even weaker type, Very Weak Physical Possibility (VWPP), under which interventions need only be consistent with the actual higher-order laws of nature. Under this proposal, P6 turns out false.

[34] Italics mine.

[35] The 'small jumps' is a continuity approximation. This sort of approximation is common in physics: e.g. the quasi-static approximation in thermodynamics tells us that if a thermodynamic process is sufficiently slow the process can be considered a succession of equilibrium states.

Second, recall that interventions ought to satisfy criteria (*IV*). Note that these criteria are presented as regulative ideals; in cases of imperfect experiments where there is no intervention variable that satisfies all of the criteria, "we should think of ourselves as trying to determine what would happen in an ideal hypothetical experiment in which X is manipulated in such a way that the conditions in [IV] are satisfied" (Woodward 2003, p.114).[36] The hypothetical scenario in question is one where the Moon's orbit is twice its actual value. The *ideal* hypothetical scenario in question is one where the effect of the Moon on the Earth is surgical because we ignore the gravitational pull of the other bodies on the Earth and vice versa. This formal aspect of interventionism is independent of the modal aspect of interventionism: that is, one thing is the nature of these possibilities, the other whether these are causal according to the criteria (*IV*) set by interventionism.

In fact, given that "it seems wildly optimistic to assume that appropriately surgical intervention processes must be available for all true causal claims" (Woodward 2003, p.130), one could argue that it is not WPP which is too stringent but (*IV*); in particular, *SURGICALITY* and *ARROW-BREAKINGNESS*. This is precisely the assumption that seems to underlie Woodward's claim that criteria for interventions (*IV*) are regulative ideals: it does not matter whether they are strictly met, because they barely ever are. Frisch agrees that "[...] in the case of at least one fundamental physical force, the gravitational force, such a shielding is impossible" (2014, p.82). Hence, since there are physical phenomena which are non-ideal, it is hopefully clear how a process *being physically possible* and it *being ideal* are conceptually independent conditions. Furthermore, provided the fulfilment of *SURGICALITY* and *ARROW-BREAKINGNESS* proves challenging, why uphold these criteria and sacrifice physical possibility?

Understandably, relaxing the criteria for interventions would imply a modification of the theory's core; the main motivation for introducing the notion of an intervention characterised via (IV) was to exclude cases of confounding, that is, situations in which changes in variable X would cause indirect changes in variable Y via an intermediate variable Z. That said, there might be some room for relaxing these criteria and retaining WPP, which I believe is a modal notion which aligns with our intuitions here: after all, it is information about physical dependence what helps us evaluate **Moon-tides**. Couldn't one consider approximate surgicality instead of ideal or perfect hypothetical surgicality? Couldn't one fix the non-surgical character of MT-I_1 and MT-I_2 by considering some of these interactions negligible? Certainly, the tidal effects of Mars on the Earth are

[36] In other words, 'ideal' rules out spurious correlations.
The original quote says 'IN' instead of 'IV,' which I think is a mistake. Italics mine.

negligible compared to those of the Moon, as are those of the planetoid, whose mass is negligible compared to the masses of both the Moon and the Earth. In fact, these are the sorts of approximations which are routinely made in order to find analytical solutions to gravitational three-body-problems. Since one will rarely (if ever) find any perfectly surgical process in nature, perfect surgicality seems too strong a criterion to impose.[37]

Should one retain perfect *SURGICALITY* and *ARROW-BREAKINGNESS*, it is the truth of P3 which is compromised. If one cannot find ideal physically possible interventions in this context, it may well be because there are none (provided one is not a pluralist about causation and believes interventionism is the correct theory of causation), or because interventionism is a defective theory of causation (provided one is a monist about causation). As a response of the second kind, Pearl has famously objected that interventionism can't apply to global models of the universe, for "if you wish to include the whole universe in the model, causality disappears because interventions disappear – the manipulator and the manipulated lose their distinction" (2000, p.350). As a response of the first kind, the worry that causal notions play no role in fundamental physics has extensively been discussed in Woodward (2007), and contested by M. Frisch (2014). I believe this question concerning the scope of application of interventionism to subject matters – the 'applicability question' (AQ) – is, even if related, distinct from the PQ. As indicated, it is the PQ that is the focus of this thesis, but should one be interested in the AQ, see the discussion of these matters in Price (2007) and M. Frisch (2014, 2020).

Last, note that Woodward has defined MT-I_1 and MT-I_2 as interventions that *could be realised*. This *forward-looking* conception of possibility places too stringent demands on possibility: under this characterisation, possible interventions are those that have the potential to be *actualised* at a given time. This is precisely the notion of SPP which Woodward himself has rejected, for it presupposes consistency with the actual ICs. However, elsewhere he clearly seems to think that counterfac-

[37] The possibility of relaxing the criteria for interventions has been the topic of intense debate in the context of constitutive mechanistic explanations in the life sciences. The most prominent account of constitutive explanation is Carl Craver's mutual manipulability approach to constitutive relevance, spelt out in terms of interventions (Craver 2007a, 2007b). Critics have argued that the account is inconsistent because constitutive relations are non-causal, and interventions exclusively demarcate causal relations (Krickel 2018, p.58). Given that interventions into phenomena with respect to its mechanistic components violate *SURGICALITY*, some of these critics (see, e.g. Baumgartner & Gebharter 2016; Romero 2015) have endorsed a *fat-handedness* criterion for interventions. However, for these critics such fat-handed definition of intervention demarcates non-causal relations. It would be interesting to explore whether, outside of inter-level explanations, a fat-handed definition of interventions might make it into an account of causation and causal explanation.

tual possibilities might be seen as interventions. In the study of whether there is any causal connection between receiving treatment with a newly-developed drug and the alleviation of a disease's symptoms, the counterfactual "If those in the control group had received the drug, the incidence (or expected incidence) of recovery in that group would have been much higher" (Woodward 2003, p.122) describes a well-defined intervention, though not an actual one: there is no actual process that assigns the drug to the control group once it has already been withheld from them. Nonetheless, Woodward thinks this is a perfectly well-defined hypothetical intervention, possible in the sense of WPP.

The example above shows how hypothetical interventions amount to considering that the system in question is disrupted in such a way that it adopts possible but non-actual ICs. I believe this is equivalent to characterising interventions as possibilities that *could have been realised*. To this end, nothing precludes the characterisation of a hypothetical intervention that doubles the orbit of the Moon as the possibility that ICs at some earlier time had been different in such a way that evolving the system forward according to the actual laws would have made the distance between the Earth and the Moon to be $2r_{M-EA}$ instead of r_{M-EA}. Against P5, given that such ICs are compatible with actual laws, the possibility of MT-I can be accommodated by WPP.[38] In relation to this point, an important question to ask is whether the notion of a hypothetical intervention makes sense at all, for it is not clear conditions (IV) can be satisfied *unless* we already know X and Y – in this case, r_{M-E} and $TIDES_E$ – are causally related.

In short, I have contested the soundness of RPP* on the basis of three main objections, which can be summarised as follows:

- *Physical possibility is relative to the relevant laws.* Since physical possibility is essentially nomological possibility, the possibility of interventions will depend on what the relevant laws are. From the fact that MT-I is not physically possible according to Newtonian mechanics, it does not follow that it is not physically possible simpliciter.

- *Physical possibility is not ideal possibility.* Being a physical process and being an *ideal* process, i.e. satisfying criteria (IV), are two independent conditions. From the fact that the criteria (IV) are not satisfied it does not follow that there is no possible physical process; it only follows that there is no possible *ideal* physical process.

[38] As the discussion in the next section will highlight, there is a clear tension between hypothetical possibilities and the notion of an intervention: we can successfully evaluate whether changes to the position of the Moon cause changes to the magnitude of the Earth's tides because we already have knowledge of the relevant causal model.

- *Physical possibility does not reduce to actualisability.* There might be no actual intervention which sets the Moon to r_{M-EH} given the actual ICs (i.e. r_{M-EA}), but this is fine, because interventions need only be hypothetically possible. MT-I is consistent with some possible but non-actual ICs at some earlier time which would not have required any further violation of the actual physical laws. Therefore, MT-I may be possible in the sense of WPP.

I shall return to the idea that interventions are physically possible processes in chapter 3. In the remainder of this chapter, I look at how Woodward fills the void left by physical possibility with conceptual possibility after having rejected physical possibility as the possibility notion characetrising interventions (§1.3.2.4). Next, I contest the suitability of conceptual possibility as the modal notion underpinning interventions, as it threatens to undermine the causal realist status of interventionism (§1.3.3).

1.3.2.4 Conceptual Possibility

The cost of imposing strict surgicality and arrow-breakingness on our model of the Moon-tides case study is high. Since there exists no strictly surgical and arrow-breaking physical process that doubles the relative distance between the Earth and the Moon, the possibilities considered as interventions are *conceptual*. This is the way the RPP* argument seen in §1.3.2.2 paves Woodward's modal story with respect to hypothetical interventions: the relevant notion of possibility the interventionist is after, Woodward argues, is conceptual possibility. Consider two variables, X and Y, which are presumably causally connected. Referring to a counterfactual claim about what would happen to Y were some possible intervention on X to be carried out, he says:

> **W2.** "[F]or this counterfactual to be true, I do not require that it be technologically or even nomologically or physically possible to manipulate X, either repeatedly or even once." (Woodward 2003, p.71)

That is, in W2 Woodward makes explicit that interventions need not be in practice possible, that is, constrained by technological means. Furthermore, interventions need not be physically or nomologically possible, an idea reinforced in a later passage when Woodward again says that "it doesn't matter that it may not be physically possible for those interventions to occur" (ibid., p.130).[39] This statement

[39] From W2 there seems to be a contrast between 'physical' and 'nomological,' yet it is not clear what this contrast is. To my understanding, nomological possibility is a subset of physical possibility, where the latter also included sets of initial, background, and boundary conditions. More on this topic in chapter 3.

is, however, prefaced with a proviso: "[...] as long as there is some basis for assessing the truth of claims about what would happen if various interventions were to occur" (ibid.). This basis is somewhat spelled out earlier in the same section of *Making Things Happen*:

> **W3.** "... if we cannot think of X as a variable that is capable of being changed from one value to a different value – if manipulation of X is not a logical possibility or if it is ill-defined for conceptual or metaphysical reasons – then claims about what will happen to Y under interventions on X will either be false or lack a clear meaning." (Woodward 2003, p.128)

According to W3 two sorts of cases are ruled out: logical impossibilities on the one hand, and conceptually or metaphysically ill-defined possibilities on the other one. Let us start with the first case: it is logically possible that 34 is not a number, but logically impossible that I am not identical to myself. Therefore, we cannot intervene to change logically necessary facts such as $A = A$.

What about metaphysically ill-defined variables? These are variables which represent metaphysically necessary properties, so interventions on them are metaphysically impossible. For instance, it is metaphysically possible that some physical particle moves faster than the speed of light, but metaphysically impossible that a gold miner would have bumped into massive blocks of gold – Au – with atomic number Z other than 79 while digging.[40] Along these essentialist lines, it is also metaphysically impossible to change an animal of one biological kind into an animal of another biological kind, so claims like 'N being a tiger causes N to run fast' won't be able to be evaluated via interventionism (Woodward 2010a, p.256). Of course, what counts as metaphysically (im)possible depends on one's modal metaphysical views.[41]

Finally, there is the case of conceptually ill-defined variables. Here, there are two plausible interpretations of what this conceptual vagueness might amount to. On one reading, one could think these are conceptual *impossibilities*, such as interventions that e.g. would change the number of angles of a triangle from 3 to 4. Since it is an analytic – and hence, conceptual – necessity that a triangle has 3 angles, this intervention is conceptually ill-defined. On another reading, conceptual vagueness is about there being multiple interpretations on how to intervene

[40] Among essentialists, at least, the latter is a metaphysical impossibility because it is part of the essence of gold – and so, a metaphysical necessity – that its atomic number is 79.

[41] For instance, whether it is metaphysically possible that I could have been born to different parents than the ones I have depends on whether one considers origin as metaphysically necessary.

on the variable in question. This is for instance the case of 'age,' 'gender,' or 'race' (see Woodward 2003, p.115; Woodward 2016, p.1052). Consider for instance the claim that being a woman causes salary discrimination. The idea is that 'gender' is conceptually ill-defined because there are several interpretations of what manipulation on gender in this context would amount to. One could think of a wide range of plausible interventions: suppressing gender information on applications, changing an employer's beliefs about the gender of the candidate, changing the specific discriminatory employment practice, etc. Thus, the vagueness here has nothing to do with conceptual impossibility – all the aforementioned interventions are conceptually possible – but with the fact that the counterfactual in question can be evaluated via any of the multiple causal relations at play, and hence the set of values the variable in question may take is underdetermined.

In short, we now have a first sketch of what is needed to evaluate interventionist counterfactuals: that interventions are logically/conceptually/metaphysically possible. This idea is floated in yet another passage:

> **W4.** "... an intervention on X with respect to Y will be "possible" as long as it is *logically* or *conceptually possible* for a process meeting the conditions for an intervention on X with respect to Y to occur." (Woodward 2003, p.132)[42]

Again, the role of both logical and conceptual possibility is mentioned.[43] In the Moon-tides case study this would amount to the claim that it is conceptually possible to adjust the distance between the Moon and the Earth while holding everything else in the universe fixed. There is a trivial sense in which this is a conceptual possibility: we can imagine this happening, an other-than-actual distance between the Moon and the Earth is not a conceptual impossibility. The biggest concern with Woodward's modal picture is, however, that it is not clear how conceptual possibility latches onto the causal structure of the world as characterised by interventions. In the following section I aim to identify the tension between this proposal and interventionism's spirit.

1.3.3 Causal Realism

Woodward's interventionist theory of causation presumably endorses a "metaphysically modest and noncommittal kind of realism" (Woodward 2003, p.121).

[42] Italics mine.

[43] Italics mine. Woodward uses 'conceptual' and 'logical' interchangeably, despite these being usually conceived as two separate notions – logic concerns propositions' form, concepts their content, their meaning. Hereafter I shall call it 'conceptual' possibility to simplify the reference.

This minimal metaphysical commitment is the idea that causal relations are "features of the world", "out there in nature" (ibid., p.23), that have their source "out there in the world" (Woodward 2014, p.698). Our role as agents is to uncover these "independently existing" relationships (Woodward 2003, p.120). While heavy metaphysics is not required to make sense of this causal realism, the mind-independent existence of patterns of causal relations does commit us to the existence of "facts of the matter, independent of facts about human abilities and psychology" (Woodward 2003, p.121) that ground these causal relations. In fact, much criticism of Woodward's project has centered on its alleged lack of truth conditions for causal claims (Strevens 2008, pp.244-247). I will not dwell on this criticism here, but I will show that presupposing causal realism poses a challenge to Woodward's resolution of the PQ.

Recall that Woodward's proposal is that in hypothetical scenarios like the one described by the **Moon-tides** causal claim, interventions need not be physically but only conceptually possible. The main concern about conceptual possibility is that it concerns relations between the content of our concepts, not things. Since dependencies between concepts' contents and dependencies between worldly things are distinct types of dependencies, if causal realism is to be sustained it must be a matter of dependence of worldly things on one another rather than dependencies amongst concepts; therefore causal relations can't be determined by mere conceptual possibility. More specifically, the problems conceptual possibility poses are as follows:

The explanatory asymmetry problem: interventionism identifies the direction of explanation with the direction of causation and thereby identifies the explanatory asymmetry with a causal asymmetry. As a matter of mere conceptual possibility, a change in the position of the Moon translates to a change in the magnitude of the tides, but the converse also holds. To get the relevant explanatory asymmetry, conceptual possibility needs to give us the causal asymmetry (Jansson 2014, Gebharter 2017).

The causal problem: if interventions are a regulative ideal, it seems like Woodward says we shouldn't really care how conditions (IV) are met. Reference to conceptual possibility undermines the rationale for using interventions, which is to identify those counterfactual claims which are causal. To test whether conceptually possible interventions are genuine interventions, i.e. satisfy criteria (IV), one needs to already be familiar with the relevant causal structure, but then the notion of an intervention is viciously circular (Jansson 2014).

Given the explanatory asymmetry and causal problems, one might deem conceptual possibility wholly unsuitable for interventionist causation. However,

these problems might be (at the very least) ameliorated provided conceptual possibility is construed as an objective type of possibility or tracks an objective possibility of some kind. In fact, something more than mere conceptual possibility seems to be at work in the process of elucidating the meaning of interventionist counterfactuals:

> **W5.** "The sorts of counterfactuals that cannot be legitimately used to elucidate the meaning of causal claims will be those for which we cannot coherently describe what it would be like for the relevant intervention to occur at all or for which *there is no conceivable basis* for assessing claims about what would happen under such interventions." (Woodward 2003, p.132)

W5 points at some sort of reasonably detailed conceptual possibility that also gives robust or selective traction on counterfactual evaluation. And so does W1 (see §1.3.2.2), where Woodward explicitly appeals to Newtonian theory and rules about the composition of forces in order to evaluate the **Moon-tides** causal claim. For counterfactual evaluation to be reliable, reality must be constraining our conceptual possibility judgements in some way. The burden of proof is with Woodward to account for how concepts are a guide to the causal structure of the world and, therefore, how conceptual possibility is the minimal relevant sense in which interventions are possible.

1.4 Conclusion

In this chapter, I introduced the possibility question concerning interventions, and thoroughly examined Woodward's argument against the physical possibility of interventions, as well as his conclusion that hypothetical interventions need only be conceptual possibilities. Given that is not clear how conceptual possibility latches onto the causal structure of the world, one might deem conceptual possibility wholly unsuitable for interventionist causation. I argued that to be a suitable modal notion for interventions, conceptual possibility ought to correspond to a suitably objective type of modal notion. This way, the objective modal notion being represented by – or underlying – conceptual possibility places constraints on the relations between concepts, and causal realism can be sustained.

While in this chapter I already disclosed my sympathies for some kind of physical possibility as the right answer to the PQ, a line of argument which will be picked up in chapter 3, in chapter 2 I try to interpret as charitably as possible the hypothesis that it is conceptual possibility which correctly captures the modal

aspect of interventions and that causal relations have an objective existence. In order to do so, I entertain the hypothesis that conceptual possibility is underlain by (at least) metaphysical possibility, seen as "the [most inclusive] sort of real possibility" (Rosen 2002, p.16) and "the maximal objective modality" (Williamson 2016, p.459).[44] The hypothesis I thereby suggest is to interpret interventionism via a rationalist lens: the lens of David Chalmers' Weak Modal Realism. In the next chapter, I discuss the plausibility of this hypothesis.

[44] As cited in Clarke-Doane (2021, p.1862). Here real or objective are understood interchangeably, as opposed to epistemic or deontic modal notions. I shall not discuss here the thesis that metaphysical possibility is the most inclusive or maximal type of objective modality; for more details, see Clarke-Doane (2021).

Chapter 2

Possibility via Conceivability: A Rationalist Account of Interventionism

2.1 Introduction

In the previous chapter, we saw that Woodward claims interventions need not be physically possible; all that is required is that they are *logically* or *conceptually* possible (Woodward 2003, p.132). As I argued, conceptual possibility understood as mere compatibility with conceptual truths is insufficiently objective to conform to the causal realism to which the interventionist subscribes.

A possible justification for the use of conceptual possibility in Woodward's project might be to interpret it as an objective modality, albeit accessed via conceivability. Consider thus conceptual possibility as possibility obtained through conceivability, our capacity to represent scenarios to ourselves using words, concepts, or sensory images (T. Gendler & Hawthorne 2002, pp.1-2). Supposing that conceptual possibility is interpreted this way, there is a project out there which connects it with metaphysical possibility, a suitably objective alethic modal notion and, as such, potentially able to give us the independent grip on reality the interventionist seeks. This is none other than modal rationalism, most influentially championed by David Chalmers (2002; 2010). In this chapter, I investigate whether Chalmers' weak modal rationalism (WMR) can help us satisfactorily answer the possibility question (PQ) concerning the sense in which interventions are possible.

The plan for this chapter is the following. In §2.2, I introduce WMR, a view which at its core has a thesis linking conceivability and possibility, the CP thesis. In §2.3, I characterise conceivability and possibility by explaining all the key notions that feature in the CP thesis, and dispel potential concerns about the a priori nature and fallibility of conceivability so characterised. In §2.4, I introduce Chalmers' conceivability argument (CA) against materialism, which has the CP thesis as one of its central premises, and which I use as a setting against which I discuss whether WMR might be able to provide interventionism with an objective notion of possibility that does justice to the causal realism which the theory promises. In §2.5,

I present the application of the CA to the Moon-tides hypothetical scenario seen in chapter 1, which results in the Moon-tides conceivability argument (MTCA). In §2.6, I argue that the soundness of the MTCA – and, consequently, whether WMR offers the resources to answer the PQ – depends on whether one holds what I respectively call 'functionalist' or 'non-functionalist' views of physical properties, two interpretations of properties which different aspects of interventionism invite. While on ontological parsimony grounds functionalism is generally preferable, I note that a certain version of non-functionalism is immune to arguments from parsimony. In §2.7, I argue that unless these views about the nature of physical properties are filled in with further details, there is no basis for preferring functionalism over non-functionalism, or vice versa. The answer to the PQ is thus conditional: if the interventionist adopts a functionalist view of properties, 1-possibility answers the PQ, but if the interventionist adopts a non-functionalist view, since the relevant kind of possibility is 2-possibility and it is not entailed from 1-possibility, WMR can't help answer the PQ. In light of this result, I suggest a pragmatic response: given the epistemic benefit functionalism offers over non-functionalism with respect to the PQ – because it is only in the context of the former, but not the latter, that the PQ can be answered – the interventionist may want to endorse functionalism about physical properties. In §2.8, I conclude by critically examining the answer to the PQ that this view about the individuation of properties backs.

2.2 What is Weak Modal Rationalism?

Chalmers' Weak Modal Rationalism (WMR) is based on the following main thesis, which establishes a link between conceivability and possibility (Chalmers 2002, p.171):

> **Conceivability-Possibility (CP):** *ideal primary (positive or negative) conceivability* entails *primary possibility*.

The CP thesis builds upon three distinctions: prima facie vs ideal conceivability, primary vs secondary conceivability/possibility, and positive vs negative conceivability.[45] First, regarding the prima facie and ideal conceivability distinction, I explain how Chalmers characterises ideal conceivability, the kind of conceivability which underwrites the CP thesis, in terms of what appears to be the case to a subject (prima facie conceivability) plus the idealising assumption that no

[45] Note that the CP thesis in fact consists of two thesis, a positive version (CP^{+}) and a negative version (CP^{-}).

better reasoning is possible, where these rational notions are taken as primitive (Chalmers 2002, p.148). After that, I review the distinction between positive and negative notions of conceivability, where both positive conceivability – characterised as modal imaginability – and negative conceivability –characterised as what cannot be ruled out a priori – give rise to plausible conceivability-possibility theses. Finally, I shall extensively examine the distinction between primary and secondary conceivability, a distinction that arises in the context of 2D semantics and which plays a central role in the inquiry of whether Chalmers' WMR can be of any use to the interventionist. But before I explain these pairs of conceivability types, let me zoom out and contextualise Chalmers' view.

It is important to note that WMR rests on two crucial and interrelated assumptions: an epistemic understanding of the two-dimensional semantics framework (E-2D), and modal monism (MM). Starting with the latter, as the name already indicates, modal monism (MM) is the idea that there is only one modal domain. In Chalmers' words: "[u]ltimately, there is just one circle of modal concepts, including both the rational modal concepts… and the metaphysical modal concepts" (Chalmers 2002, p.194). According to MM, what cannot be ruled out a priori (logical possibility) gives us insight into how things could have been (metaphysical possibility).[46] It is because reason is the main source of modal knowledge that this (modal) monism is a version of (modal) rationalism.

MM is challenged by a posteriori necessity and a priori contingency (see Putnam 1972 and Kripke 1980); this is the reason why Chalmers supplements it with E-2D semantics. But what is E-2D and how does it serve its purpose?[47] 2-D semantics is a formal framework which, in contrast with standard possible worlds semantics, assigns extensions and truth-values to expressions relative to *two* intensions (primary and secondary) instead of one, which capture two different dimensions of meaning. 2-D semantics has been used for different explanatory purposes; for instance, David Lewis, Frank Jackson, and David Chalmers used it to isolate an a priori aspect of meaning. This provides us with an account of how speakers can assign truth-conditions to their own utterances on a priori grounds (Schroeter 2019).

Chalmers' understanding of 2-D semantics is epistemic – reason why we named it 'E-2D'– because the first dimension in the 2D framework is filled by intensions which represent epistemic possibilities, where these are broadly logical

[46] In the literature MM is often formulated as the coincidence of logico-conceptual possibility (e.g. Mallozzi 2021) or broadly logical possibility (e.g. Vaidya 2008) with metaphysical possibility, respectively.

[47] The following rather brief introduction to E-2D semantics will be expanded in §2.3.3, but for more details one may check Chalmers (2004).

possibilities, and thus connected to the a priori. This is how Chalmers closes the gap between the rational and epistemic domains. The second dimension in the 2D framework corresponds to intensions which represent possibilities constrained by the world, that is, regular metaphysical possibilities. This way, and as we will see in §2.3.3, the E-2D semantic apparatus allows Chalmers to draw the desired distinctions between aspects of meaning while advocating MM, and thus deal with a posteriori necessity. There is no modal dualism; there are two types of evaluations of sentences over one single space of rationally accessible, and epistemically and metaphysically possible, worlds.

2.3 What is Conceivability?

2.3.1 Prima facie vs ideal conceivability

The most natural characterisation of 'ideal conceivability' is as a priori knowledge only available to perfect reasoners, "free of all contingent cognitive limitations" (Chalmers 2002, p.148, in reference to Menzies 1998). As Chalmers himself admits, the existence of an all-knowing figure seems incongruous, for it leads to an infinite regress: nothing seems to preclude the existence of an even better possible reasoner for every possible reasoner. Furthermore, I would add that the notion of ideal conceivability constitutes, in Nagelian terms, a 'view from nowhere.' While interventions need not be carried out by humans, we humans will naturally be interested in those causal relations which we can manipulate and are, therefore, accessible to us. Hence, a notion of conceivability with such an alienating nature, while certainly objective, won't be able to accommodate these manipulability considerations.

Chalmers contrasts ideal conceivability with 'prima facie conceivability,' defined as conceivability for a subject on first appearances.[48] After some consideration, a non-ideal reasoner finds that a statement S passes the tests that are criterial for conceivability, e.g. the subject cannot a priori find a contradiction in the hypothesis expressed by S (Chalmers 2002, p.148). Even though prima facie conceivability doesn't share the problems of ideal conceivability, it is problematic in another respect: prima facie conceivability judgements can be undermined by continued rational reflection, or by the judgement of less cognitively limited reasoners (whose existence is logically possible). Additionally, should one adopt

[48] Prima facie conceivability is here taken as a contrast class for ideal conceivability. The further reflection of a *secunda facie* conceivability would be fine. The crucial point here is that ideality is a problematic assumption.

a notion of conceivability defined in terms of what subjects judge to be possible instead of what is ruled out on a priori grounds, there is danger of ending up with a subject-relative notion of conceivability, which certainly would not meet the objectivity requirement.

In view of these challenges on both ends, Chalmers offers an in-between notion which weakens the formulation of ideal conceivability in a way that i) it does not require the figure of an all-knowing reasoner, and ii) it strengthens the notion of prima facie conceivability so that it is not undermined by rational reflection. This middle ground position characterises ideal conceivability as a sort of *rational undefeatability* or, as he calls it, 'undefeatability by better reasoning' (Chalmers 2002, p.148). Undefeatability by better reasoning establishes that for statement S to be ideally conceivable it is sufficient that there exists a possible subject for whom S is prima facie conceivable, and who has enough evidence or justification that such prima facie conceivability cannot be defeated by any further reasoning (ibid.). Chalmers here takes these rational notions to be primitive, trusting that we have enough of an intuitive grasp of what (un)defeatability by better reasoning and good reasoning are: he notes that this idealising assumption is the one usually provided in epistemology as a warrant or stability condition for knowledge.

How, or rather why, is ideal conceivability connected to apriority? To avoid a trivial characterisation of conceivability as what a subject judges to be possible, undefeatability by better reasoning ought to be understood as what cannot be ruled out a priori, a characterisation to which I now turn. The idealising assumption thus has an "in-built" connection with apriority (Chalmers 2010, p.552).

2.3.2 Positive vs negative conceivability

'Negative conceivability' holds that statement S is conceivable when it is not ruled out a priori, or when there is no apparent contradiction in it (Chalmers 2002, p.149). With respect to ideal conceivability, S is ideally negatively conceivable when it is not a priori that $\neg S$.

At the other end, 'positive conceivability' holds that S is conceivable when one can form a positive conception, when one can *imagine* a situation in which S is the case.[49] There are two ways of positively imagining a situation that verifies S: through *perceptual* imagination and through *modal* imagination. To perceptually imagine that S requires that the subject has a perceptual image that represents

[49] Chalmers' notion of positive conceivability follows that of Yablo's (1993) in terms of coherent modal imagination. Chalmers' negative vs positive mirrors Van Cleve (1983)'s distinction between weak/strong conceivability.

S as being the case, e.g. one imagines a two-headed giraffe by forming a visual image of a two-headed giraffe. Nonetheless, imagining need not be constrained by perception. One can imagine a world or situation where such imagining is not a perceptual experience, e.g. DNA's nucleobases, or historical events that did not occur. In these cases, one has "an intuition of" (Chalmers 2002, p.151) a world or a configuration of objects or properties within a world. This sort of imagination which goes beyond perception but shares with it its mediated objectual character is modal imagination, and it is the one involved in positive conceivability.[50] Chalmers makes the notion of modal imagination more precise by requiring that it be coherent, that is, that it does not involve contradictions when filled in by additional details (Chalmers 2002, p.153). Hence, S is positively conceivable when it is *coherently* modally imaginable.[51]

Note that there is a CP thesis for both types of conceivability, positive and negative. If I can rule out S on a priori grounds, I will definitely not be able to coherently imagine a situation that verifies it, so the negative thesis entails the positive one. In other words, positive conceivability implies negative conceivability. If I cannot rule out S on a priori grounds, it is not clear whether I will be able to coherently imagine a situation that verifies S, and so, whether positive conceivability entails negative conceivability. Therefore, ideal primary negative conceivability seems a stronger guide to primary possibility. Despite this apparent difference in strength between (ideal primary) positive and negative notions of conceivability, Chalmers argues that since both seem to withstand all of the well-known counterexamples to the CP thesis, they both seem reasonably suited to back the CP thesis (Chalmers 2010, pp.147-148).

2.3.3 Primary vs secondary conceivability and possibility

The primary vs secondary conceivability distinction is motivated by the need to accommodate Kripke's a posteriori necessities (Kripke 1980, Chalmers 2002, p.156), which are not knowable a priori. This is also the reason why Chalmers resorts to 2D semantics. Consider Putnam's famous Twin Earth thought experiment, an illustration of his argument for semantic externalism (Putnam 1975). The thought experiment invites one to suppose that somewhere in the galaxy there is a planet which is like Earth in all respects except that there the liquid we call water is not

[50] That imagination has a 'mediated objectual character' means that it involves an intermediate mental object – the imagined world or scenario – towards which we form a certain attitude (Chalmers 2002, p.150). This is the case of make-believe, which in chapter 5 I will argue is the sort of imagining involved in counterfactual reasoning in science.

[51] As we will see in chapter 5, while make-believe generally allows logical inconsistencies, counterfactual reasoning in science ought to be coherently modally imaginable.

H_2O but XYZ. XYZ is "indistinguishable from water at normal temperatures and pressures" and "it tastes like water and it quenches thirst like water" (Putnam 1975, pp.9-10). This planet is Twin Earth. In Twin Earth, what fills oceans, lakes, and seas, and composes rain, is not H_2O but XYZ.

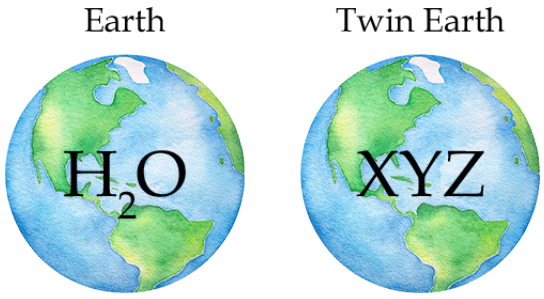


Figure 2.1: Putnam's Twin Earth thought experiment.

In the context of the Twin Earth thought experiment, consider the statement S: 'Water is H_2O.' Is $\neg S$ conceivable? According to E-2D, there is a sense in which it is, and one in which it isn't: the answer hinges on whether $\neg S$ is considered with respect to its primary or its secondary intension. Before dwelling on the distinction, let me briefly introduce the notion of 'intension.'

Traditionally, the meaning of words and terms was supposed to consist of an extension – the set of things the term applies to, or referent – and an intension – the concept associated with a term. Frege and Carnap resisted the *psychologism* implied by this definition of intensions and instead defended the view that meanings are publicly available (see Putnam 1975, pp.5-6). On this note, Carnap defined a word's or sentence's intension as "a function from possibilities to extensions relative to these possibilities" (Carnap 1947, as cited in Chalmers 2004, p.156). What does this mean?

In E-2D, there are two ways in which the extension of an expression might depend on possible states of the world, or in other words, two ways of thinking about possibilities. One might consider a world as actual – i.e. as what it may turn out to be the case for all we know – or as counterfactual – i.e. as what could have been the case (Davies & Humberstone 1980). The function that selects the world considered as actual is the *primary intension* (1-intension) of an expression, and the function that selects the world considered as counterfactual is its *secondary intension* (2-intension). A word's or expression's meaning is constituted by both its 1-intension and 2-intension, and because these are functions to extensions with respect to notions of *possibility* (actual and counterfactual), meaning is thereby tied to modality.[52]

[52]In a way, 1-intensions and 2-intensions mirror Frege's (1892) notions of a name's sense and

The counterfactual reading of possibility is the usual way of thinking about metaphysical possibility: as what could have been the case had the world been other-than-actual in some respects. The actual reading corresponds to epistemic possibilities, i.e. possibilities evaluated from the perspective of a speaker who hypothetically locates herself in the world entertained and imagines that such world is actual. It is for this reason that worlds which are relevant to primary intensions are 'centered' (Chalmers 2004, p.160): they describe possibilities evaluated from a center, namely, a spatiotemporal point with the coordinates of the particular individual making an utterance and a concrete time (Chalmers 2010, pp.146-147). For instance, the primary intension of the word 'I' picks out Chalmers at a world centered on Chalmers, but Woodward at a world centered on Woodward. Conceivability tracks these two ways of understanding possibility. 'Primary conceivability' (1-conceivability) is a type of conceivability that concerns ways the world might turn out, possibilities mapped by 1-intensions ('1-possibility' hereafter).[53] Analogously, 'secondary conceivability' (2-conceivability) concerns ways the world might have turned out, with respect to which 2-intensions map words' and expressions' extensions ('2-possibility' hereafter).

We are now in the position to understand the sense in which $\neg S$: 'Water is *not* H_2O' is conceivable and the sense in which it is not. We can imagine a Twin Earth world in which water, the drinkable liquid which fills up rivers and oceans, is made up of the chemical molecule XYZ instead of H_2O. Such a world is epistemically or 1-conceivable, for the possibility of a world filled with watery stuff with a chemical base XYZ is not ruled out a priori.[54] How do we evaluate this world considered as *actual*? We entertain the hypothesis that the scenario or world entertained is one's own world (Chalmers 2004, p.177) and then consider whether statement $\neg S$ would be true if that were the case. If so, we say that the scenario in question *verifies* $\neg S$.[55] At a Twin Earth, where water is XYZ, the word 'water' would then have a different meaning than the word's meaning on Earth. Its primary intension would pick out XYZ, and so $\neg S$ would there be true. The scenario verifies $\neg S$, and so $\neg S$ is 1-conceivable.

What about 2-conceivability? The crucial difference with 1-conceivability is that a word or statement's 2-intension is whatever thing a word or expression happens to pick in the actual world, its actual reference, which is kept fixed. This

reference.

[53] For those familiar with the literature on conceivability, 1-conceivability corresponds to Yablo's (1993) $conceivability_{ep}$.

[54] Unlike statements like 'Some bachelors are married' (Chalmers 2004, p.181), which are ruled out a priori because the statement 'All bachelors are unmarried' is analytic.

[55] Verification is a relation that captures our daily linguistic uses.

is why the evaluation of possibilities takes a *counterfactual* form: if the conceived world in question had obtained, would $\neg S$ have been the case? If so, such world *satisfies* $\neg S$. The secondary dimension of the E-2D framework thereby accommodates Krikpean a posteriori necessities. In the Twin Earth case, given that the chemical compound of watery stuff in the actual world is H_2O and identity is a necessary relation, 'water's' 2-intension is H_2O in every world.[56] Hence, when $\neg S$ is considered according to its 2-intension, since it must be necessarily true that S, its negation is not 2-conceivable. The fact that 2-conceivability requires that we keep the reference of words fixed explains why it is a posteriori; one needs empirical information that water is H_2O in order to judge whether S is true. On the contrary, when making judgements about 1-conceivability, "the true character of our world is irrelevant" (Chalmers 2002, p.158); one brackets empirical information about our world and then takes the perspective of a speaker at the world considered, supposed to be actual (Mallozzi 2021, p.6). The idea is that whether S is true or not will be assessed based solely on considerations of logical and conceptual coherence, so it is in this respect that 1-conceivability lines up with apriority.

It is now hopefully clear how the E-2D framework accommodates a posteriori necessities: they are those sentences which have necessary 2-intensions but contingent 1-intensions. Analogously, a priori contingencies – such as 'I exist', 'the length of S at t_0 is 1m' (Kripke 1980, pp.56-57) – are those sentences which have contingent 2-intensions but necessary 1-intensions; this implies that 1-conceivability is at best an imperfect guide to 2-possibility. In §2.6.2, I revisit the consequences of this lack of perfect entailment from one to the other.

Let me first focus on the suitability of the CP thesis, which establishes a link between 1-conceivability and 1-possibility, to accommodate the needs of the interventionist. Since interventionism aims to establish causal claims, one may worry that physical scenarios are not the sort of thing one may know non-trivial facts about in an a priori manner. Furthermore, one may worry that even if these are known a priori, given we may not know whether better evidence for a statement is available, conceivability may lead to erroneous possibility judgements. In short, one may be concerned about the characterisation of conceptual possibility as potentially fallible a priori reasoning.

[56] The necessity of identity is one of Kripke's essentialist bridge-principles. There is disagreement among metaphysicians as to whether the principle holds; see e.g. Lowe (2012). Nonetheless, the overwhelmingly accepted view is that identity is necessary.
For the sake of brevity I also omitted the terminology of 'rigid' vs 'non-rigid' designators. Rigid designators (such as proper names, indexicals or natural kind names) are terms that have the same extension in all possible worlds where the thing selected exists, whereas non-rigid designators (such as descriptions) don't. That's why 'water' – provided one grants it's a natural kind term – picks out H_2O in all worlds, but 'watery stuff' picks out H_2O in Earth and XYZ at Twin Earth.

2.3.4 Concerns with the CP thesis

The CP thesis says that ideal primary conceivability entails primary possibility. There are, I believe, two main concerns regarding the rationalist nature of CP which might undermine its application to interventionism. I should address before I entertain whether WMR sheds light on the PQ.[57] The first concerns the a priori nature of WMR: the modal rationalist seeks to vindicate the objectivity of modal facts by appealing to idealised a priori reasoning. But if 1-conceivability is purely an a priori notion, thus based on considerations of ideal logical and conceptual coherence, how can it inform us about causal structure? The second concerns how ideal conceivability, qua rational undefeatability, is implemented in practice.

Let me start with the concern that WMR characterises conceivability as a priori reasoning. Lack of sympathy for rationalism is not difficult to come by. Philosophers like Block & Stalnaker (1999) have argued that modal claims are *partly* justified by background knowledge about the world: chemists had to carry out experiments to find out water's composition. On the most radical end of the spectrum, philosophers of physics and philosophers of science like Redhead (1995), Esfeld (2012); Ladyman & Ross (2007), French & McKenzie (2012) or McKenzie (2020), among others, have even completely rejected the use of a priori methods in modal epistemology. The clearest articulation of such a complaint is found in Ladyman's and Ross' *Everything Must Go*. In the opening pages, for instance, they deny:

> "[...] that a priori inquiry can reveal what is metaphysically possible. Philosophers have often regarded as impossible states of affairs that science has come to entertain. For example, metaphysicians confidently pronounced that non-Euclidean geometry is impossible as a model of physical space, that it is impossible that there not be deterministic causation, that non-absolute time is impossible, and so on" (Ladyman & Ross 2007, pp.16-17).

More recently, in the particular context of physics, McKenzie has provocatively (and rhetorically) asked: "for if physics is replete with metaphysical entailments, and clearly physics is an a posteriori venture, then in what sense could a purely a priori method, one that bypasses those empirical constraints, be expected to reliably arrive at those same conclusions?" (McKenzie 2020, p.3).

I won't attempt to counter this completely dismissive attitude towards a priori metaphysics, for as Guay & Pradeu (2020) argue, this *radical* naturalised

[57] Recall that it is short for 'possibility question.'

metaphysics is not, and should not be, the only way of doing metaphysics of science.[58] However, regarding Block's and Stalnaker's objection, I concede that WMR does not *solely* rely on a priori reasoning, for words' and statements' 2-intensions are indeed fixed by their referents in the actual world, which we discover in an a posteriori manner – this is not only compatible with Chalmers' modal rationalism but the reason why it has been described as 'weak.'[59] The key point here is that while the composition of water is something we discovered empirically, the conditional inferences we make in order to establish 1-possibility claims are justified a priori. For instance, from the supposition that the actual world is the H_2O world we don't need to empirically check that this is the case to infer that S is 1-possible.[60]

As an analogy, think of scientific theories: while a posteriori knowledge is needed in the early stages of theory construction, we can and, in fact, do dispense with it later on. This way of making scientific modal inferences is what Hirvonen et al. (2021) call 'relative modality (RM):' modal inferences are made and evaluated a prori relative to a system. The contrast class for RM is 'actualist-manipulationist modality' (AM): what is observed to be actual informs what is possible. These two ways of making modal inferences are clearly linked: in scientific investigation, what is kept fixed in a system is what we previously found out to be the case via actual manipulations (Hirvonen et al. 2021, pp.13824-13825). With respect to the intervention MT-I (see §1.3.2.2, reproduced in upcoming §2.4), changing facts about relative distances in the Solar system is way beyond our reach, and yet we constantly and confidently assert possibility claims about such hypothetical physical scenarios on the basis of inference from the content of the theory – which we acquire a posteriori – and mathematical and logical knowledge – which we acquire a priori – alone.[61] Thus, WMR does not endorse a priority from the ground up.

Having addressed the a priority concern about modal rationalism, let me now discuss the fallibility concern. It is undeniable that judgements about what

[58] Guay & Pradeu (2020, pp.1848-1851) distinguish between 'scientific metaphysics,' which is metaphysics done building upon scientific results, and 'metaphysics applied to science,' which is the reverse type of inquiry from metaphysics to science. The trend started by Ladyman and Ross is an instance of the former type. They label it 'radical' (p.1851) because it postulates that good scientific metaphysics should not only be universalist and realist, but also reducible to physics.

[59] In comparison to the 'strong' rationalism of Plato, Descartes, etc. for whom a priori inquiry is the only route to knowledge.

[60] Chalmers responds to Block and Stalnaker in *The Character of Consciousness* (Chalmers 2010, pp.567-568) saying that empirical facts don't justify but "play a causal role in determining the relevant patterns of inference." I must confess I find this response cryptic.

[61] Note as well that even though the dynamical content of physical theories is determined a posteriori, the kinematical constraints on theories, which are preconditions for their applicability, are a priori (see Curiel 2016).

is conceivable are sometimes a bad guide to judgements about what is possible; e.g. in the face of new phenomena, which has not yet been theorised, or when there is limited access to the phenomena. In these circumstances, metaphysicians, like scientists, make mistakes, as erroneous past pronouncements about e.g. non-Euclidean geometry or indeterministic causation have shown. However, the charge of fallibility can be countered on the following grounds. First, that a priori reasoning is fallible doesn't undermine its general reliability; it only highlights the fact that the knowledge so produced is revisable (Tahko 2011, pp.35-37). These mistakes do not point at any flaw in the methodology, so to speak, but only at the uses humans make of it. The point that conceivability, even if fallible on certain occasions, is generally reliable in everyday and applied scientific contexts, has also been made by Van Inwagen (1998). He nicely illustrates the point via an analogy between modal judgements and judgements of distance made by the eye, such as "that mountain is about 30 miles away" (Van Inwagen 1998, p.70). This judgement won't be reliable when it's very foggy or the speaker has a very bad vision, not corrected by glasses or contact lenses. However, in normal weather and health circumstances these judgements will be pretty accurate. Along these lines, the idea is that even if imperfectly, undefeatability by better reasoning tracks 1-possibility sufficiently well.

The important question now is: is 1-possibility a notion of possibility the interventionist can make use of? In the rest of this chapter, I argue that provided the interventionist accepts the CP thesis, whether she can make do with 1-possibility depends on one's view of physical properties.[62] Unfortunately, interventionism as Woodward develops it is silent with respect to the metaphysics of the properties causal models aim to represent. In an attempt to fill this gap, I argue that different aspects of interventionism inform different views of properties, which in turn lead to different prospective answers to the question of whether WMR can offer the interventionist the resources to answer the PQ. This is why I now turn to the conceivability argument (§2.4) and its respective application to the Moon-tides hypothetical intervention discussed in chapter 1 (§2.5), in the context of which the previously-raised questions will be investigated.

2.4 The Conceivability Argument (CA)

The **conceivability argument (CA)**, also known as '2D argument against materialism,' is a very well-known argument against the physicalist view that phenomenal

[62] I believe this result generalises beyond interventionism: the modal knowledge one may gain via 2D semantics crucially depends on one's metaphysical commitments.

properties supervene on physical properties. However, what I am here interested in is not its conclusion – namely, 'materialism is false' – but rather the strategy followed to get there: the gist of the argument is showing that conceivability entails 1-possibility (epistemic possibility), which because of modal monism entails 2-possibility (metaphysical possibility). In particular, I use the CA as a background against which to investigate whether WMR is able to furnish interventionism with an objective *conceptual* notion of possibility (see §1.3.2.4).

Having outlined the road map, let me now have a look at the CA, which reads as follows (Chalmers 2010, p.149):

CA:

P1. $P \wedge \neg Q$ is conceivable.

P2. If $P \wedge \neg Q$ is conceivable, then $P \wedge \neg Q$ is 1-possible.

P3. If $P \wedge \neg Q$ is 1-possible, then $P \wedge \neg Q$ is 2-possible.

P4. If $P \wedge \neg Q$ is 2-possible, then materialism is false.

C. Materialism is false.

In the CA, P is the conjunction of all the microphysical truths about the universe, and Q is any phenomenal truth about the actual world, such as the truth that someone is phenomenally conscious. There conjunction $P \wedge \neg Q$ describes a *zombie* world, that is, a world physically identically to ours but where no one is phenomenally conscious (Chalmers 2010, p.142).

P1 is just a conceivability statement, where conceivability should be understood as primary positive ideal conceivability (see §2.2 and §2.3).

P2 is none other than the CP thesis: if $P \wedge \neg Q$ is a priori coherent, there is an epistemically possible scenario relative to which the 1-intension of $P \wedge \neg Q$ is true, and so $P \wedge \neg Q$ is 1-possible.

P3 is the application of Chalmers' E-2D framework, and is admittedly the trickiest premise. Materialism requires the ordinary metaphysical impossibility that P is true without Q also being true. Hence, materialism requires 2-(im)possibility rather than 1-(im)possibility: we are not talking about the (im)possibility that a world verifies $P \wedge \neg Q$, but about the (im)possibility that it satisfies $P \wedge \neg Q$. For P3 to be true, both the 1-intensions and 2-intensions of P and Q ought to coincide. If they don't, as we will see in §2.6.3, it is because Russellian monism – a dualist view of consciousness which explains this metaphysical and explanatory gap – is true.

P4 is just an interim conclusion, irrelevant for our current explanatory purposes: as just said, materialism requires that it could not have been the case that P

without Q. In other words, materialism requires that zombie worlds are impossible. From P1-P4, Chalmers concludes that materialism is false.

2.5 The Moon-tides Conceivability Argument (MTCA)

In §1.3.2, I introduced the PQ via the example of a hypothetical intervention, MT-I, which would double the relative distance between the Moon and the Earth (in variable talk, the intervention changes the value of the variable r_{M-E} from r_{M-E_A} to $2r_{M-E_A}$) and thereby affect the magnitude of the Earth's tides, represented by $TIDES_E$. I there entertained a couple of realistic ways MT-I could come about, both of which violate $SURGICALITY$ and $ARROW\text{-}BREAKINGNESS$. In both cases the intervention variable I corresponds to the change in the relative distance between the Moon and an astronomical body other than the Earth, be it Mars (r_{Ma-M}) or a planetoid (r_{p-M}). Either of these variables changing value incurs a change in the Moon's position with respect to the Earth when the respective bodies move with respect to or impact with the Moon. The variable intervened upon, r_{M-E}, which as a result of the intervention changes value, causes $TIDES_E$ to change its value accordingly.

Consider MT-I_1, the implementation of the Moon-tides intervention via Mars. $SURGICALITY$ is violated because a change in r_{Ma-M} affects $TIDES_E$ not just via its effect on the Moon but also directly, as such change will alter Mars' gravitational pull on the Earth. $ARROW\text{-}BREAKINGNESS$ is also violated, for the intervention does not sever causal connections between $TIDES_E$ and the position of other astronomical bodies other than the Moon. One way to preserve $SURGICALITY$ and $ARROW\text{-}BREAKINGNESS$ is to abandon realism about *how* interventions may in fact come about.

Recall that interventions need not be *actual* physical processes. Evaluating the possibility of a hypothetical intervention that doubles the orbit of the Moon amounts to none other than envisioning that ICs at some earlier time had been such that evolving the dynamics of the system forward according to the actual laws would have made the distance between the Earth and the Moon to be $2r_{M-E_A}$ instead of r_{M-E_A}, in such a way that the value of $TIDES_E$ would have been a well-defined, yet other than the actual, one. In short, when evaluating interventionist counterfactuals, one brackets information about what could do the job of the intervention variable I and which values it could take, and focus instead on the dependence between X and Y holds under non-actual values. This is where WMR and the CA come in.

Consider MT-I once again. One can conceive that the Moon is now at twice its original distance with respect to the Earth (r_{M-E_C}=2r_{M-E_A}), without this change having been caused by the impact or displacement of any other body. This is a situation in which the relative distance between Mars and the Moon has also changed value (r_{Ma-M_C}<r_{Ma-M_A}), but the relative distances between bodies other than the Moon (r_{Ma-E_C}=r_{Ma-E_A} among them) have remained intact. The fact that a change in the value of the relative distance between the Moon and the Earth is accompanied by a change in the value of the relative distance between Moon and Mars is the result of the triangle inequality theorem in Euclidean geometry, which tells us that the sum of any two sides of a triangle is greater than or equal to the third side. If one considers the triangle formed by the Moon, the Earth, and Mars, given that it has been stipulated that the value of r_{Ma-E} is kept fixed, the value of r_{M-E} cannot be wiggled without – at least – a change in the value of r_{Ma-M}.[63] Figure 2.2 represents this situation:

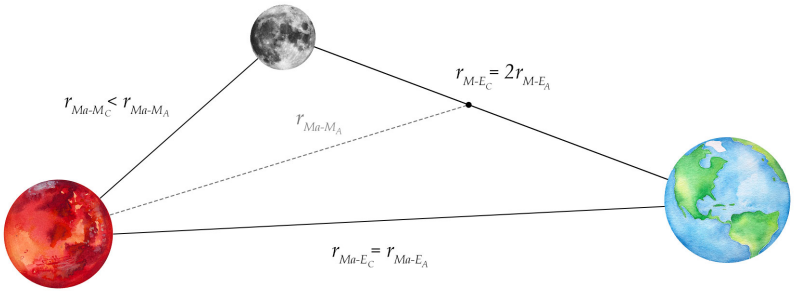


Figure 2.2: Representation of the relative distances between the bodies in the three-body system Mars-Moon-Earth pre (r_{i-j_A}) and post intervention (r_{i-j_C}). The point between the Moon and the Earth represents r_{Ma-M}'s position pre-intervention.

To simplify notation, I will hereafter drop the subscript '*A*', which indicates the actual value of the relative distances.

We are now in the position to apply Chalmers' CA to the hypothetical scenario described by **Moon-tides**. Recall that the goal is to investigate whether WMR could afford the interventionist with the metaphysical possibility required to preserve the purported causal realism of interventionism. In analogy with Chalmers' original argument, I call the application of the CA to the Moon-tides example the **Moon-Tides Conceivability Argument (MTCA)**. It takes the following form:

[63]Since positive conceivability ought to be modally coherent, I won't entertain the possibility that the triangle inequality theorem can be violated. However, I acknowledge that this theorem might trigger questions about the distinctness of the relata and hence their causal status. It could be that different accounts of the metaphysics of space and time may differ on whether these variables are genuinely distinct, so I will proceed as though they are for the sake of the argument.

MTCA:

P1. $P \wedge \neg r_{M-E}$ is conceivable, where:

$r_{i-j} \subset P$, where $\forall r_{i-j}, r_{i-j} = k_{i,j}$ and $i, j \neq M$

P2. If $P \wedge \neg r_{M-E}$ is conceivable, then $P \wedge \neg r_{M-E}$ is 1-possible.

P3. If $P \wedge \neg r_{M-E}$ is 1-possible, then $P \wedge \neg r_{M-E}$ is 2-possible.

C. $P \wedge \neg r_{M-E}$ is 2-possible.

Note that I have removed P4, present in CA, as my aim here is not to defend any particular position in the mind-body debate. In this version of the argument there are also no phenomenal properties involved, all the properties are physical: P are all the microphysical truths about the universe except for r_{M-j} (or, equivalently, r_{i-M}).[64] The variable that plays the role of Q in the CA has been replaced by r_{M-E} in the MTCA. The MTCA thus concerns the conceivability of the scenario in which the relative distance between the Moon and the Earth is different from the actual but the relative distances between other bodies with respect to the Earth have remained unchanged. While MTCA does not specify it, we assume that r_{M-E} can only take physically possible values, namely, values which are possible according to physical theory (i.e. physically impossible values are ruled out). Furthermore, we should understand $\neg r_{M-E}$ not as the absence of said property – as is the case of Q in zombie worlds – but rather as an other-than-actual instantiation of said property. In variable-talk, we imagine a world where the distance between the Earth and the Moon is $2r_{M-E}$.

In §2.6, I use MTCA as a background against which I discuss what notion of possibility, 1-possibility or 2-possibility, the interventionist might be interested in. Unfortunately, interventionism is neutral with respect to causal metaphysics: in causal modelling, the relata are the variables' changing values, but what do these represent? Provided the values of variables describe properties or changes in properties, in the following section I argue that the relevant type of possibility is determined by whether physical properties are individuated by their structural features – in which case 1-possibility will suffice – or (also) by their intrinsic features – in which case it won't. As I show in §2.6, interventionism makes room for both views of properties, which unfortunately lead to incompatible results about whether we can obtain modal knowledge that supports substantive causal judgements via WMR.

[64] Recall that due to the triangle inequality, if the value of r_{M-E} changes and we suppose that no other object moves, all the other relative distances between the Moon and the other bodies will be affected.

2.6 Interventionism & Physical Properties

Interventionism does not offer any metaphysics for the properties represented via causal models. To remedy this, in the following I argue that different facets of interventionism warrant different theories of properties. In particular, I claim that the structuralism implicit in type-level causation licenses a functionalist view of properties, while the interventionist's interest in the counterfactual profile of tokens invites a non-functionalist view of properties.[65] I then look at the implications of endorsing each of these views for the notion of possibility interventionism is after. More specifically, I argue that the functionalist can make do with 1-possibility (§2.6.1), whereas the non-functionalist needs 2-possibility (§2.6.2). In §2.6.3, I discuss the prospects of both views about physical properties against the backdrop of the modified CA, where Chalmers complicates matters by introducing Russellian monism (see Chalmers 2010, pp.149-152). More specifically, I discuss at length the plausibility of a non-functionalist view about physical properties, what I call 'Russellian physicalism.' I argue that provided we accept that some versions of non-functionalism are ontologically and explanatorily on a par with functionalism, this in turn means that at best what we have is a conditional answer to the PQ, a conclusion I draw in §2.7.

2.6.1 Functionalism about Physical Properties

The interventionist theory of causation is concerned with causal relations among properties or event types – represented by variables – which might be instantiated by many tokens – represented by variables taking different values from a well-defined range. In this line, we might think that what defines physical properties is the functions or roles they play. Given that in the present case interventionism is concerned with type causal relations, the relevant characterisation is that for a property to be the property that it is, the property in question plays a certain causal role defined by its structure. I call this view of properties '(causal) functionalism about the physical'.[66] For instance, according to functionalists about properties, mass might be whichever property causes matter to change its speed or position upon the application of a force, or which attracts bodies in a certain way.[67] This view bears similarities to Hawthorne's (2001) causal structuralism, the view that a

[65] This tension will come up again in chapter 4.

[66] Hereafter I will drop the term 'causal' because it will be assumed that given interventionism is a causal theory it is clear the focus is on causal functions.

[67] Not necessarily an inverse square law, for according to Chalmers laws of nature are metaphysically contingent (see Chalmers 2002, p.190) for a discussion of this point.

property's essence is exhausted by its causal modal profile. Hawthorne's property causal structuralism may be seen as a variety of *pure dispositionalism*, the view that all properties are essentially dispositional (see e.g. Mumford 2004 or Bird 2007). In other words, properties *are* their dispositions, also known as 'causal powers,' and reason why the view is also called the 'Powers View' (see e.g. Shoemaker 1984). Since I don't think the interventionism ought to commit to the full metaphysical package these views offer, hereafter I shall stick to my proposed label: functionalism about the physical.

Functionalism has expanded beyond philosophy of mind, where it originated, and become popular within the philosophy of science community. For instance, in the philosophy of physics, Ney & Albert (2013) and Albert (2015) defend wavefunction realism in quantum mechanics – the view that the wave function, a field in high-dimensional space, is the fundamental ontology of the world. This is a functionalist view of spacetime, because spatiotemporal regions are picked out by the causal role they play. It is an ontological type of functionalism, but there is also a conceptual version thereof. In the context of spacetime functionalism, Knox (2019) has defended a *conceptual* version thereof; hers is a "functionalism about spacetime *concepts* that ties spacetime to a certain role within a broader theory" (p.121). Just like the concept of a kidney is the concept of an organ which filters the blood and maintains certain chemical balances, spacetime is a structure that plays the spacetime role, namely, the role of determining local inertial structure and the coordinate systems associated with it, in a physical theory. Even Chalmers (2012, 2021) has defended a version of conceptual spatiotemporal functionalism called 'spatial functionalism.'[68] Note that while ontological functionalism implies conceptual functionalism, the reverse need not obtain.

Back to functionalism about physical properties: if all that matters for a causal property to be the property that it is is that it has a certain causal function, then the reference of a word or modal claim about the property in question is completely fixed by its causal function. Therefore, the consideration of worlds as actual or counterfactual will make no difference to physical properties, and so the 1-intensions of words and statements will coincide with their 2-intensions. Hence, 1-possibility entails 2-possibility, and the MTCA is sound. If the content of possibility claims is exhausted by 1-possibility, then the CP thesis suffices to underwrite the sense in which interventions ought to be possible: if ideally primary

[68] The main difference with Knox's spacetime functionalism, an account of the concept of spacetime in physical theory, is that spatial functionalism is an account of the ordinary concept of space (Chalmers 2021, p.172). It need be so, for what Chalmers is interested in is closing the explanatory gap between the manifest and scientific image.

conceivable, a scenario is 1-possible.[69] In the context of the Moon-tides scenario, the CP thesis would guarantee the following inference: if the Moon has as an epistemic counterpart which is at twice its distance from the Earth, then **MT-I** is (1-)possible. Thus, the functionalist interventionist can endorse 1-possibility as an answer to the PQ.

2.6.2 Non-functionalism about Physical Properties

Besides the focus on type-causation, interventionism is characterised by another feature which may undermine a functionalist view of properties, and hence the response to the PQ along the lines of the CP thesis provided in the previous section. Interventionism is a *counterfactual* theory of causation, as as such is concerned with *what-if-things-had-been-different* questions. This suggests that causal reasoning is always centered in the actual world, whose character plays a key role when reasoning about alternative scenarios. For instance, Woodward & Hitchcock (2003, p.20) define hypothetical interventions as the "*hypothetical* values of *X* and *Y possessed by the very object o*." This claim might invite the view that the possibility claims interventionism cares about, as well as the claims these entail, are *de re* modal claims. In short, if the emphasis is on the counterfactual aspect of interventionism, the relevant possibilities are those evaluated with respect to worlds considered as counterfactual, in which case the identity of properties might not be exhausted or fixed by their causal roles. In this section I explore this vies, which in contrast to the previous I call (causal) 'non-functionalism about the physical.'

Note that non-functionalism aligns with *quidditism*, the view that there is more to properties than their causal powers, where 'quiddities' are these "natures," (Armstrong 1989, p.44, p.59), "intrinsic natures" (Schaffer 2005, p.1), or "intrinsic aspects" (Robinson 1993, p.19) above the causal powers of properties.[70] It is paramount to note the distinction between *pure quidditism* – also known as 'categoricalism' or the 'qualities view,' – according to which it is *only* the quiddities which are essential to properties, from both 'two-level' and 'powerful qualities' views of properties. On the one hand, the 'two-level' view of properties is a dualist view: there are both intrinsic and dispositional properties, and the former act as categorical bases which bring about the latter as higher-level properties (see e.g. Prior et al. 1982). On the other hand, the 'powerful qualities' view, most prominently cashed out in terms of identity (see Heil 2003), is the view that properties

[69] Note that ex-hypothesis the CA presupposes 2D-semantics.
[70] For a comprehensive characterisation, see Locke (2012, p.348).

have both dispositional and qualitative sides, which may or may not be seen as essential. Since I want to remain non-committal on this matter, I shall hereafter use the newly coined label: non-functionalism about the physical.

Consider the mass example again. A non-functionalist understanding of the property mass goes beyond its causal function. In such a scenario, one might reasonably hold that the 1-intension and the 2-intension of the word 'mass' capture different aspects of its meaning: its 1-intension is fixed by a description of its causal role and its 2-intension by the property that actually plays that role. Consider now a world *w* containing *schmass*, a property which satisfies the 1-intension of the term 'mass' – that is, which causally behaves like mass – but has a different composition than mass. At world *w*, the 1-intension of 'mass' would select the property schmass, but its 2-intension would not, for 'mass' refers to mass, the property with certain qualitative features that actually plays the mass role in Earth. Since the 1-intension and the 2-intension of the word 'mass' don't coincide, reasoning about an epistemic counterpart of mass like schmass doesn't tell us anything about mass, the property of interest in the actual world. Similarly, possibility claims about epistemic counterpart properties of the Moon's mass (like the Moon's *schmass*) or of the Moon's position with respect to the Earth (the Moon's *schposition* with respect to the Earth) are not possibility claims about the Moon, the possibilities we care about after all.[71]

In short, focusing on the counterfactual aspect of causation places more stringent demands on possibility, for it demands that it is 2-possibilities that feature in the truth conditions of causal judgements. However, the counterfactual aspect of causation might be interpreted along non-functionalist lines, under which 1-possibility does not entail 2-possibility; given it is the latter non-functionalists aim at, this is a fatal result for their prospects to answer the PQ.

2.6.3 Functionalism or Non-Functionalism?

Chalmers' WMR makes room for both functionalism and non-functionalism about the physical. As we shall see in §2.6.3.1, 1-possibility need not entail 2-possibility when physical properties are at stake.

In the context of the CA, Chalmers introduces Russellian monism, also known as 'type-F-monism,' to explain how there might be zombie worlds devoid of conscious beings (Chalmers 2010, pp.149-152): the actual and hypothetical zombie world differ with respect to their *intrinsic profile*, namely, the properties

[71] One might worry the distinction structural/intrinsic makes sense for mass, but not for the relative distance between the Moon and the Earth. This is a point I discuss in §2.6.3.2.

that fill the physical structure of the world in question (ibid., p.151). This explains why 1-possibility does not entail 2-possibility.

In §2.6.3.1, I introduce the modified CA which incorporates the possibility that Russellian monism is true, the CA*. In §2.6.3.2 I consider an analogously modified MTCA, the MTCA*, which incorporates a neutral kind of monism but for physical properties only, what I call 'Russellian physicalism.' I argue that provided one construes Russellian physicalism along the lines of 'austere quidditism' (Locke 2012; Smith 2016), there is no obviously compelling reason to prefer either functionalism about the physical to non-functionalism about the physical, or vice versa. This result leaves us with a conditional answer to the PQ.

2.6.3.1 Russellian Monism: The Modified Conceivability Argument (CA*)

In *The Character of Consciousness*, Chalmers acknowledges that while the inference from 1-possibility to 2-possibility in P3 in the CA (see §2.4) holds for *Q*s about consciousness, where there doesn't seem to be "such a strong dissociation between appearance and reality" (Chalmers 2010, p.149), the case is not as straightforward when it comes to physical properties, which are considered to have a mind-independent existence. Therefore, he suggests an amendment to P3 in the CA, amendment which takes the form of a disjunct. The resulting argument is:

> **CA*:**
> P1. $P \wedge \neg Q$ is conceivable.
> P2. If $P \wedge \neg Q$ is conceivable, then $P \wedge \neg Q$ is 1-possible.
> P3.* If $P \wedge \neg Q$ is 1-possible, then $P \wedge \neg Q$ is 2-possible or Russellian monism is true.
> P4. If $P \wedge \neg Q$ is 2-possible, then materialism is false.
>
> ———————
>
> C. Materialism is false or Russellian monism is true.

I have renamed this modified form of the original conceivability argument 'CA*.' CA* incorporates 'Russellian monism,' as a disjunct in P3. Russellian monism is based on Russell's ideas regarding physics and consciousness in *The Analysis of Matter* (Russell 1927): physics only describes relations among physical bodies, but says nothing about the intrinsic nature of physical properties (Chalmers 2010, p.133). One way to fill this gap is to embrace the view that there are quiddities which categorically *ground* both the structural features of physical properties *P* and phenomenal properties *Q*. There is disagreement about the nature of these intrinsic properties: panpsychists claim that these are phenomenal, panprotopsychists claim

that these are *protophenomenal* (Chalmers 2010, pp.151-152; Chalmers 2016), i.e. properties that without being phenomenal themselves constitute phenomenal properties when collectively organised in an appropriate way.

The idea behind Russellian monism is that consciousness is necessitated by some conjunction of structural and intrinsic properties, and this is why $P \wedge \neg Q$ is not 2-possible: hypothetical worlds differ from the actual with respect to their intrinsic profiles. If one thinks that it is the protophenomenal intrinsic properties what physical terms refer to, they will qualify as physical. Then, zombies won't be full physical duplicates, and so it is in this sense that this view can be seen as a sort of materialism – though non-reductive and hence distinct from the reductive materialism Chalmers aims to reject with the CA and CA*.[72]

Why is Russellian monism of our interest? Despite it being a view about consciousness, it is relevant to us because it provides an explanation of why in some cases relevant to interventionism 1-possibility might not imply 2-possibility: conceived worlds might differ from the actual with respect to the intrinsic properties which undergird physical structure. While zombie worlds differ from the actual with respect to phenomenal properties, the Moon-tides hypothetical world differs from the actual with respect to *some* physical properties, namely, the distances between the Moon and other bodies. Can we modify Russellian monism to account for scenarios involving only physical properties, à la D. K. Lewis (2009)? In §2.6.3.2, I show how an extension of the CA* to the Moon-tides hypothetical scenario would look like, and discuss the prospects that such Russellian physicalism holds. I argue that while dualist conceptions of non-functionalism can be defeated on parsimony grounds, this is not the only non-functionalist view available. If construed as 'austere quidditism,' it is no longer clear the interventionist ought to choose functionalism over non-functionalism about the physical.

2.6.3.2 Russellian Physicalism: The Modified Moon-tides Conceivability Argument (MTCA*)

Just like CA* incorporates Russellian monism as a disjunct in P3 of the CA, the MTCA can be analogously modified to incorporate a version of Russellian monism for conceivability exercises exclusively about physical properties (recall that the Moon-tides hypothetical scenario differs from the actual world only with respect

[72] Alluding to the non-reductive nature of Russellian monism, Chalmers writes that "while the view arguably fits the letter of materialism, it shares the spirit of antimaterialism" (2010, p.134). Russellian monism shares the spirit of anti-materialism because it postulates phenomenal or protophenomenal properties as fundamental and distinct from the structural properties of physics. In which sense non-reductive physicalism is different from property dualism is not clear, as Chalmers himself admits (ibid., p.152).

to r_{M-E}) which I shall call 'Russellian physicalism,' as follows:

MTCA*:

P1. $P \wedge \neg r_{M-E}$ is conceivable, where:

$r_{i-j} \subset P$, where $\forall r_{i-j}, r_{i-j} = k_{i,j}$ and $i, j \neq M$

P2. If $P \wedge \neg r_{M-E}$ is conceivable, then $P \wedge \neg r_{M-E}$ is 1-possible.

P3. If $P \wedge \neg r_{M-E}$ is 1-possible, then $P \wedge \neg r_{M-E}$ is 2-possible or Russellian physicalism is true.

C. $P \wedge \neg r_{M-E}$ is 2-possible or Russellian physicalism is true.

Since it is an adaptation of the modified conceivability argument CA*, I shall analogously name this extended version of the Moon-tides Conceivability Argument 'MTCA*.' Russellian physicalism is the view that the structural features of physical properties are grounded in intrinsic properties that serve as their categorical bases, and the differences in the intrinsic profiles of the actual and hypothetical worlds are what explain the differences in r_{M-E} (and other r_{M-i}).

In the following, I present a twofold objection that can be levelled against the view that Russellian physicalism is a coherent view, as well as a counter-objection to it. In other to contextualise the objection and counter-objection, I first shed light on the relationship between physical properties and physical theory: physical theory informs us about the extrinsic or relational character of physical properties. While non-functionalists might equally share this view, I first outline a parsimony objection to the existence of intrinsic properties: why assume the existence of quiddities when these have no clear explanatory benefit? (see e.g. Hawthorne 2001; Locke 2012; Smith 2016). While Russellian monism has a clear explanatory benefit, namely, that it explains how consciousness fits the physical world, it is not clear what the distinctive explanatory import of Russellian physicalism is, with respect to functionalist views of physical properties. Furthermore, if physical theory informs us about extrinsic properties, intrinsic properties can't be physical; but then it is a mystery how they can make worlds qualitatively distinct with respect to their physical features.

Let me start by outlining the relationship between physical properties and physical theory. Chalmers takes physical properties to be "[...] the fundamental properties that are invoked by a completed theory of physics. Perhaps these will include mass, charge, spatiotemporal position [...]" (Chalmers 1996, p.33). I believe the invocation of a completed theory of physics is epistemically problematic, for it implies we won't be able to settle any physical truths until we know what the

most fundamental physical theory is; this is precisely the sort of scenario where ideal conceivability is fallible (see §2.3.4). Fortunately, in the context of the Moon-tides, no fundamental physical theory is required to understand what the relevant physical properties are; therefore, I suggest we loosen Chalmers' definition of physical properties as those invoked in our best available physical theory.

In short, Chalmers takes physical properties to be what Stoljar has named 't-physical' properties, that is, "*either* the sort of property that physical theory tells us about or else is a property which metaphysically (or logically) supervenes on the sort of property that physical theory tells us about" (Stoljar 2001, p.256).[73] This 'theory-based' conception of the physical is contrasted with the 'object-based' conception of the physical, which understands physical properties as 'o-physical' properties: either "the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents," or else "a property which metaphysically (or logically) supervenes on the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents" (Stoljar 2001, p.257). In short, physical theory informs us about t-physical properties.

Furthermore, pace Russell, Chalmers tells us that "[...] the basic elements of physical theories seem always to have come down to two things: the structure and dynamics of physical processes" (Chalmers 1996, p.118).[74] In short, physical theory informs us about the structural-dynamical character of the physical world, and so about *extrinsic* rather than *intrinsic* properties of the physical world. In fact, r_{M-E} is unarguably an extrinsic property, whose instantiation – represented by a variable taking a certain value – is fully described by Newtonian mechanics and rules about the composition of forces (see W1 in §1.3.2.3). Note that this idea that physical theory only informs us about the extrinsic or relational nature of properties, however, is not incompatible with non-functionalism. For instance, Lewis views structural properties as "role-occupants" in scientific theory and everyday life; this leaves open whether fundamental physical properties also have some non-structural features, even if these cannot be studied by science (see H. Taylor 2017 for a similar argument).

Many arguments have been raised against what I have here called 'Russellian physicalism.' In the following, I focus on what I believe is the most forceful

[73] As mentioned in Stoljar (2001, p.257, fn.9), this notion of the physical had been made by a few authors since Meehl & Scriven (1956).

[74] Whether physical theory refers to dispositions or dispositions and their categorical bases – provided one grants that dispositions need categorical bases – and the implications this has for the t-physical/o-physical distinction is something that has been discussed by Stoljar in 'Two Conceptions of the Physical' (Stoljar 2001, pp.260-261).

argument: why invoke what you don't need? (Hawthorne 2001, p.268).[75] This is an argument for ontological parsimony. In Locke's words, "[...] a metaphysician ought to try her best to appeal only to the ontological resources of empirical science" (2012, p.357). Think about the hypothetical intervention described by **Moon-tides**: we can make full sense of the consequences of intervening on the distance between the Moon and the Earth by appealing exclusively to Newtonian mechanics and rules about the composition of forces. Russellian monism clearly has an explanatory benefit, which is to explain how consciousness arises, though it is unclear what the explanatory benefit of positing intrinsic properties or grounds for structural properties has for physical explanations and modal claims exclusively referring to these structural features. This brings me to a second, related concern, which is that quiddities are a "will-o'the-wisp" (Hawthorne 2001, p.362): if physics does not inform us about them, their nature is spooky.

In the literature on the quidditism vs dispositionalist debate about the individuation of properties, both Locke (2012) and Smith (2016) have attempted to counter this two-fold objection by suggesting that quiddities are *not* something *distinct* from the property itself. This line of reasoning has led them to defend an alternative type of quidditism to traditional quidditism they call 'austere quidditism,' which is the view that properties are individuated by numerical identity (instead of the individuation via quiddities of traditional quidditism, which they call 'extravagant quidditism').[76] This sort of "quidditism *without* quiddities" (Locke 2012, p.351) evades parsimony concerns because it does not require the postulation of superfluous intrinsic natures/aspects. I won't go into the intricacies of Locke's and Smith's views; what interests me here is the observation that there is at least one plausibly coherent construction of non-functionalism under which there is no reason to prioritise functionalism over non-functionalism about physical properties.

2.7 Does WMR Answer the Possibility Question?

To what extent can WMR provide interventionism with a notion of *conceptual* possibility which secures causal realism is an open question, for Woodward's interventionism is neutral with respect to the metaphysics of causal properties. Therefore, in §2.6 I suggested two ways in which one may fill this gap. These

[75] For a summary of all the arguments that have been levelled against quidditism, see Smith (2016, p.238).

[76] Smith defends what I believe is a version of austere quidditism, which she calls 'moderate.' However, note that she insists that it should not be classified as a type of quidditism (Smith 2016, p.244).

two proposed theories of property individuation are based on two aspects of interventionism, apparently in tension. On the one hand, the focus on type-level causation – (IV) are about the structural and relational aspect of causation – encourages a functional view of properties. In a functional view of properties, given that the 2-intensions of the words by which we refer to them coincide with their 1-intensions, 1-possibility is all that is needed for the truth of hypothetical causal claims about e.g. how changes in r_{M-E} affect $TIDES_E$. Fortunately, provided the concerns about the a priori nature and fallibility of ideal primary conceivability (positive or negative) discussed in §2.3.4 have been assuaged, 1-possibility is a suitably objective notion of alethic possibility the interventionist can adopt.

On the other hand, the focus on causation as a relation of counterfactual dependence might encourage a contrasting view, where properties are not exclusively individuated by their causal functions but also by their intrinsic features. In this case, we are not only interested in claims about an epistemic counterpart of the property whose modal profile we study, but rather in claims about the modal profile of the actual property the claim in question is about. Therefore, it is 2-possibility the interventionist is after, and given the relevant possibility claims (like **Moon-tides**) are about physical properties, unfortunately there is no guarantee it is entailed from 1-possibility.

Which of the two views obtains? On parsimony grounds, functionalism seems undoubtedly preferable: a scientifically-responsible metaphysics ought not posit any ontological entity or feature not posited by empirical science. However, as usual, the devil is in the details. I argued that, at face value, there are versions of non-functionalism – more specifically, Russellian physicalism qua austere quidditism – which evade parsimony concerns. In the face of this result, the answer to the PQ can at best be conditional: should one adopt a functionalist view about the physical, 1-possibility is a feasible answer to the PQ. Should one adopt a non-functionalist view about the physical, WMR falls short of resources to help the interventionist clarify the modal aspect of interventionism. Of course, on the basis of these different epistemic outcomes with respect to the PQ, there is still a case to be made in favour of functionalism about the physical. This is a pragmatic yet reasonable move, as I will now argue.

2.8 Conclusion

Should we take Woodward's tenet of causal realism seriously, what follows is that the notion of possibility the interventionist is after ought to be objective. In the

previous chapter, I introduced his argument that the notion of possibility at work in interventionism is conceptual or logical (in the broad sense). In this chapter, I explored one way one might secure the truth of conceptual possibility claims via Chalmers' WMR, which underpins the objectivity of the conceivability-possibility link via rationalism. Furthermore, I argued that for modal rationalist claims about physical properties, only if one endorses a functionalist view about the physical can WMR provide a satisfactory answer to the PQ. In this case, 1-possibility is identified as the notion of possibility of interest to the functionalist interventionist.

Functionalism squares nicely with interventionism: interventions are cashed out in terms of variables of causal models changing values. As discussed in §1.2.3, if apt these models correctly describe the causal structure of the systems they represent. Nonetheless, while one might hold a functionalist view of properties and think that the meaning of causal claims is completely fixed by the causal roles the properties they describe play, I concede that one might still be suspicious of the rationalist flavour of 1-possibility. The fear is that with WMR Chalmers has reduced modality to a property of statements: whether something is 1-possible or 2-possible depends on whether an intension is verified or satisfied (Mallozzi 2021, p.1398). While this notion of possibility might be sufficient to ground a notion of causation "independent of facts about human abilities and psychology" (Woodward 2003, p.121) like Woodward wants, the worry remains that 1-possibility is not sufficiently worldly; not in the sense of not being part of this world, for by definition all non-actual things are not part of this world, but in the sense of not being sufficiently constrained by the world. After all, recall that WMR endorses modal monism: it unifies the epistemic and metaphysical possibility domains, which are rational. Moreover, a non-worldly type of possibility would not be able to accommodate actual interventions – the result would thereby be a dis-unified account of the possibility of interventions.

For those who are suspicious of the rationalist interpretation of interventionism here offered, in the following chapter I explore an alternative: that the relevant notion of possibility the interventionist is interested in is not conceptual but physical.

Chapter 3

Possibility via Invariance I: A Physical Account of Interventionism

3.1 Introduction

Woodward's interventionism is a realist theory because it presupposes that there are causal relations *out there* in the world, which are invariant under at least some interventions. Along these lines, Woodward defines lawhood as high invariance under interventions. Accordingly, I call his invariance-based view of laws the 'Invariance View.' The present chapter will engage directly with the Invariance View, the nexus of many familiar and unfamiliar issues concerning the realist commitments of interventionism. The chapter is organised along three axes – one methodological, one metaphysical, and one epistemic – which structure the various goals of this chapter. They are as follows.

Regarding methodology, in §3.3.3 I show how the Invariance View aims to respect scientific methodology. This commitment motivates the characterisation of invariance in terms of three independence conditions: independence among laws, between laws and ICs, and among ICs. Building upon this motivation to take not only scientific methodology but also scientific reasoning seriously, in §3.5 I propose the *Extended Invariance View* of laws. The Extended Invariance View explains a specific subset of counternomics: counterfactuals which involve changes in the laws themselves. Building upon the independence conditions in terms of which invariance is characterised, I propose to understand changes *to* laws as *higher-order interventions*, i.e. changes to the values of what can be seen as ICs with respect to higher-order laws (which I call 'ICs^n'). Thus, the Extended Invariance View provides a framework to make sense of these counterfactuals as describing higher-order patterns of invariance.

Regarding the epistemology of the Invariance View, in §3.3.4 I try to dispel an epistemic concern around the 'Possibility Question' (PQ). In chapter 1, we saw that the PQ is the question of what sort of possibility characterises interventions. There is a worry that this cannot be physical possibility (understood as nomological

possibility in significant part) because if it is laws that tell us what interventions are possible but in order to assess the degree of invariance of a law one needs to know which changes are possible, this reasoning is epistemically circular (Psillos 2004, p.300). In §3.3.4, I show how this worry might be assuaged, though this does not clear out the following related metaphysical circularity concern.

Regarding the metaphysics of the Invariance View, in §3.3.5 I argue that the Invariance View is affected by a metaphysical circularity concern regarding modal facts. The Invariance View is a non-reductive view of laws because it does not reduce laws, which provide modal information, to non-modal facts. Nonetheless, it suffers from a circularity concern regarding the constitutive dependence between laws and the possibilities described by interventionist counterfactuals, all modal facts. I dub this concern the 'Modal Priority Question' (MPQ). In §3.4, I contend that the methodological motivation of the Invariance View does not exempt Woodward from providing an account of what makes laws and counterfactuals true and offer a solution: Woodward could break this metaphysical circularity by explicitly supplementing the Invariance View with a primitivism about invariance à la Maudlin (2007), which is perfectly compatible with a philosophical account of laws that stays true to scientific methodology.

The three axes which provide the backbone of the Invariance View come together in §3.6, where I argue that the PQ is best answered in terms of physical possibility, cashed out as invariance. Since invariance is worldly dependence, understanding hypothetical interventions as physical possibilities allows us to see why they are objective possibilities. More specifically, given that interventions need only be hypothetical, I argue that Weak Physical Possibility (WPP, see chapter 1) is the relevant notion of possibility for most interventions. That said, WPP does not accommodate higher-order interventions. This poses a generalised PQ, which can be answered via an even weaker notion of physical possibility: Very Weak Physical Possibility.

3.2 The Possibility Question Revisited

In chapter 1, I identified the 'Possibility Question' (PQ) as one of the undeveloped cornerstones of Woodward's interventionist theory of causation. Interventions have a formal and a modal aspect: they have to satisfy formal criteria (*IV*) and (*IN*) and be possible in the relevant sense. The sense in which actual interventions are possible is trivial (at least under most views of modality). However, since hypothetical interventions do not take place, it is unclear what sorts of possibilities

these are. Thus, the PQ essentially inquires into the relevant sense of 'possible' for hypothetical interventions.

In connection to the PQ, there are a couple of points that need further attention. One is that the PQ is a metaphysical question about the modal status of interventions, not an epistemic question about *how* these possibilities are identified, though these metaphysical and epistemic aspects are interconnected, as we shall see in §3.3.4 and §3.3.5. The other is, as previously highlighted, that criteria (*IV*) for interventions remain silent on the modal metaphysical status of interventions, hence the question of whence causal realism.

The causal realist aspirations get somewhat blurred by Woodward's assertion that "it is built into the notion of an intervention that any change in the value of Y will occur only through the change in the value of X, so that the details of how both I and X come to have their values do not matter to the assessment of [interventionist] counterfactuals" (Woodward 2003, p.128).[77] Interventionist counterfactuals are those whose antecedents specify interventions, but "any more detailed specification of the way their antecedents are realized seems inappropriate and unnecessary" (ibid.).[78] I shall call the outcome that it does not matter how the intervention takes place, i.e., how I and X get their values, *PROCESS IRRELEVANCE*.

Given *PROCESS IRRELEVANCE*, the focus seems to shift from there being a possible process $I - X$ to there being an intelligible process $X - Y$. But what does it mean that there is an intelligible possible process that wiggles Y's value only via X? As seen in §1.3.2.4, a constraint on interventions is that they are metaphysically and conceptually well-defined. Note that this might suffice regarding intelligibility, but given the tenet of causal realism, more is needed: the relevant possibilities ought to be objective, i.e. mind-independent.

This is why in chapter 2 I proposed to view interventionist possibilities as metaphysical possibilities entailed from a priori ideal conceivability. Nonetheless, this strategy requires that we concede to some controversial premises of weak modal rationalism and furnish interventionism with a specific metaphysics for the individuation of properties. Hence, in this chapter I discuss an alternative

[77] Square brackets mine.

[78] Woodward has even suggested that I gets value i via "some minor miracle" (Woodward & Hitchcock 2003, p.15; see also Woodward 2003, p.145). I find this reference to the work of D. Lewis (1979) potentially misleading, for as we will see in the following section, Woodward's invariance-based view of laws is built upon the assumption that the value of a law's ICs can be varied independently of the law in question, and as such does not require any local violation of the law. The dissimilarities between the two accounts are stark; for Lewis laws are exceptionless regularities – reason why miracles are understood as similarity relations between the possible worlds which describe local violations of the actual laws and the actual world – for Woodward exceptionlessness does not characterise laws.

answer to the PQ in terms of physical possibility via invariance, which is worldly dependence and therefore accounts for the causal realism of interventionism in a much more straightforward and scientifically-realist-friendly way. This is, I believe, the right answer to the PQ.

3.3 The Invariance View

3.3.1 What is Invariance?

Laws undoubtedly play a central role in science, especially in physics.[79] Woodward acknowledges the centrality of laws, yet rejects a familiar conception of them as exceptionless generalisations, as well as their monopoly of this conception of our explanatory practices. For this reason, he develops an alternative view, which defines explanatoriness in terms of *invariance under interventions*. In this section, I introduce the main elements of his invariance-based view of laws, which I call the 'Invariance View.'

Invariance describes a property of relationships "which continue to hold under some appropriate range of variations in initial conditions, where these include variations in the values of the variables figuring in the generalisation" (Woodward 2020, p.1910; see also Woodward 2013, p.61). Imagine the Moon-tides scenario described in chapter 1: Newton's law of universal gravitation is invariant under variations in the values of the masses and the distance between them, which may be characterised as initial conditions (ICs hereafter).

The concept of invariance under ICs is attributed to Wigner (1967), who claims that physical information can arguably be split into laws of nature, "the domain of regularities," and initial conditions, nature's "complications" (ibid. p.3).[80] Such split between laws and ICs follows the methodological motivation that a proper account of laws ought to describe *actual scientific practice*: "our target should be conceptions of law and physical modality that figure in current and past science" (Woodward 2020, p.1908).

The reconstruction of physical theorising presented above is an oversimplified version thereof. Note, for instance, that most laws are stated in the form of differential equations. The solutions of differential equations are functions that

[79] Lange (2000) offers a good defence of this claim. There are sceptics to this idea – e.g. Cartwright (1983), Van Fraassen (1989), or Giere (1999) – but most philosophers agree on the centrality of laws in science.

[80] See Bursten (2021) for a criticism of the view that initial (and boundary) conditions are mere contingent facts, a common view in the literature on scientific explanation. She focuses on boundary conditions and argues that they are components of models which specify structure. As such, they should not be viewed as having less explanatory import than laws.

satisfy the equations plus the constraints imposed on them by the modelling context. In physics, it is in the context of differential-equation problems that boundary conditions (BCs hereafter) come into play. In such contexts, BCs are characterised as those conditions imposed on different values of an independent variable, where ICs are those conditions imposed on the solutions and their derivatives at the same value of an independent variable (usually time t).[81] As a matter of example, consider the strings of a guitar, whose wave behaviour is described by the following 2^{nd} order partial differential equation:

$$\frac{1}{v^2}\frac{\partial^2 y}{\partial^2 t} = \frac{\partial^2 y}{\partial^2 x} \tag{3.1}$$

In eq. (3.1) v is the velocity of the longitudinal wave. The ICs would be the values of the solution $y(x,t)$ and $\partial y(x,t)/\partial x$ at $t = 0$, whereas the BCs would be the values imposed on the solution at its 'extremes', e.g. since the string is fixed on both ends, supposing the string has a length L we know that at both x=0 and x=L the vertical displacement $y(x,t)$ will be null.[82] See table 3.1 for a comprehensive summary of both sets of conditions for the present example:

ICs	BCs
$y(x,0) = f(x)$	$y(0,t) = 0$
$\frac{\partial y(x,0)}{\partial x} = g(x)$	$y(L,t) = 0$

Table 3.1: ICs and BCs for a longitudinal wave with fixed endpoints.

Hopefully, this toy example gives a flavour of the difference between ICs and BCs, at least in physics. Note that since BCs are "conditions at more than one point on a spatial or temporal boundary characterizing some system" (Woodward 2020, p.1913), ICs are a subclass thereof. Thus, invariance does not only characterise a certain modal robustness under ICs, but a modal robustness under BCs more generally.

Besides ICs and BCs, causal generalisations can also be invariant under background conditions (BaCs), where these are "conditions that do not figure in the laws themselves – roughly because variations in them are irrelevant to the relationship described by the law" (ibid.). The colour of the string or the number of guitars in my bedroom would be examples of BaCs. Woodward calls invariance under BaCs 'stability' to distinguish it from the invariance under ICs and BCs; this

[81] The number of conditions has to be the same as the order of the ordinary (ODE) or partial differential equation (PDE) in question.

[82] This type of BCs which specify the value of the function itself are known as Dirichlet BCs.

distinction is important, for as we will see in §3.3.2, it is invariance under ICs and BCs that is key for causal status.[83] In short, laws are those generalisations that are *invariant* under a substantial range of BCs (including ICs) and *stable* under a wide range of BaCs.

As seen in the example of the wave equation for the fixed-ends string, differential equations like eq. (3.1) do not incorporate information about the values ICs or BCs take; these constraints are imposed by the modelling context (whether the string is fixed or loose, at what time we want to describe its behaviour, etc.). It is only when combined with specifications of these conditions that differential equations can explain the behaviour of physical systems. That is, the interventionist does not assimilate modal features *exclusively* to laws, but also to ICs and BaCs. For this reason, I believe the best representation of physical modal space is:

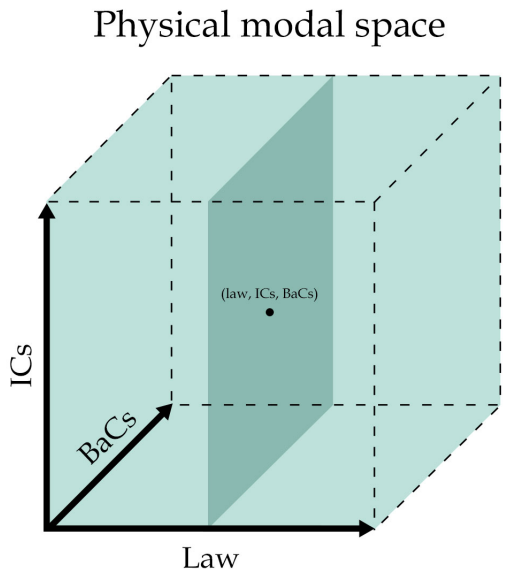


Figure 3.1: Representation of the physical modal space.

In fig. 3.1, every actual law is a plane in the space, and each point (law, BaCs, ICs) represents a physical possibility. ICs correspond to all the variables featuring in the laws, each of which might take different physically possible values. A higher-dimensional space would be needed to correctly capture this feature, but hopefully fig. 3.1 provides a schematic impression of the physical modal space.

Let me focus next on the meaning of 'laws.' Consider once more the example of the guitar's strings: it is not clear whether eq. (3.1) *represents* or *is* a dynamical law. In earlier works, Woodward refers to the generalisations representing

[83] This distinction between invariance and stability first appears in Woodward's 'Causation with a Human Face' (Woodward 2007, p.77).

repeatable relationships in nature, and not these relationships themselves, as laws (Woodward 2013, p.59). Nonetheless, this earlier reading presupposes a minimal causal realism: "[...] if a generalization is a law, then it must accurately represent (at least up to some level of approximation or within a certain range or regime of circumstances) how matters stand in nature, even though the law itself is not 'in' nature" (ibid.). In this reading, laws are representational devices that capture objective causal relations. In latter work, however, he talks of laws as "represented by differential equations" (Woodward 2020, p.1919), alluding to the most common ontological reading of lawhood. This is a similar controversy to that discussed in §1.2.3.1, where we wondered whether the relata of causal models are the variables or the properties these variables represent. From these interpretations, we can see that Woodward is interested in the *laws of science*, which are expected to map onto some kind of objective patterns, what metaphysicians would call the *laws of nature*.

Having looked at the main ingredients of the Invariance View, in §3.3.2 I lay out how these relate to one another in order to explain what makes the Invariance View an attractive view of lawhood.

3.3.2 What is the Invariance View?

In the Invariance View laws are characterised as *change-relating generalisations* which are *highly invariant* under *testing interventions*. Before we proceed to engage with the details of the view, let me analyse the key terms in this definition.

A 'change-relating' generalisation G is a generalisation which maps the values of some dependent variable Y to the values of some independent variable X such that $G{:}Y = f(X)$. Some change-relating generalisations are invariant, and some are accidental. We discern the former from the latter via testing interventions.

A 'testing' intervention for G is an intervention that changes the value of X from x_1 to x_2. If and only if $f(X)$ changes according to G, we can say that G is invariant under said intervention. The intervention thus *tests* whether G is invariant (Woodward 2013, p.63). For instance, 'All the coins in Bill Clinton's pocket on January 8, 1999 are dimes' (Woodward 2000, p.208) is not invariant under any testing intervention that would introduce a non-dime coin in Clinton's pocket, for that would not turn it into a dime.[84]

In the Invariance View, laws are just the *highly* invariant causal generalisations, namely, those generalisations which hold in a sufficiently large or substantial

[84]There is a certain ambiguity about whether invariance is generality with respect to different objects, or possible properties (of the object the explanation targets). For instance, is Newton's 2^{nd} law *F=ma* invariant under different types of objects, – such as balls, cars, or rockets – or invariant under changes to, say, the mass (of a given ball)? This issue will be examined in chapter 4.

range of ICs and BCs, including BaCs, and "which (at least in many cases) are integrated with other laws as part of a coherent theory" (Woodward 2020, p.1915). For instance, generalisations in fundamental physics or chemistry tend to be highly invariant. Nonetheless, except for the range of conditions under which they are true, they are of the same nature as the change-relating generalisations in the special sciences.[85] It is for this reason that Woodward says that:

> "[T]here is nothing more to lawhood than, so to speak, de facto invariance under some appropriately large range of changes in initial/boundary conditions, including changes involving interventions." (Woodward 2013, p.65)

The Invariance View is thus formulated in terms of the following necessary condition (*N*) for lawhood (ibid., p.64):

> **(N):** If a change-relating generalization *G* is a law, then it is invariant under some testing interventions.

And can be illustrated as fig. 3.2 below:

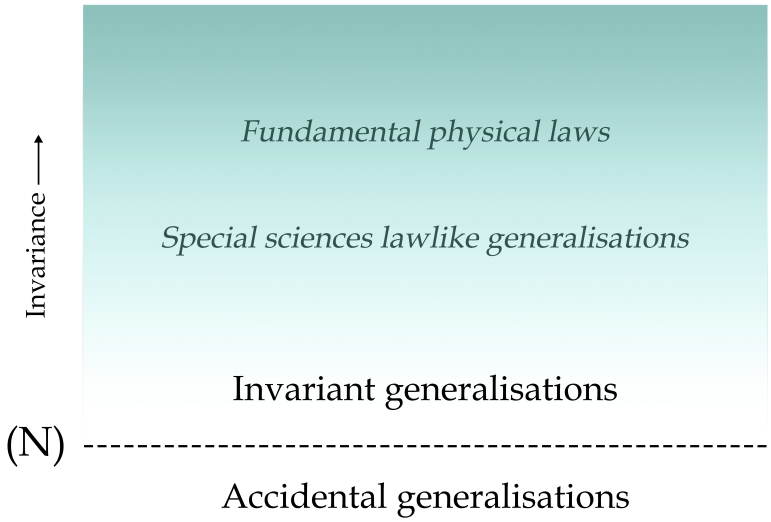


Figure 3.2: Invariance View in terms of (*N*).

Figure 3.2 is an improved version of the illustration of the Invariance View in Woodward (2013, p.64). Note first that condition (*N*) draws a threshold between accidental and genuine change-relating generalisations. Any change-relating generalisation above (*N*) is invariant to some extent, and hence explanatory. Its

[85] Though causal generalisations within and outside physics might be different in other respects, e.g. being more or less probabilistic.

degree of explanatoriness is defined by the width of the range of conditions which can be obtained via testing interventions under which it holds.[86] On the contrary, the generalisation 'All the coins in Bill Clinton's pocket on January 8, 1999 are dimes' (Woodward 2000, p.208) is, as mentioned, not invariant under any testing intervention. This generalisation is accidental, and thus relegated below (N), outside the realm of causal explanation.[87]

Note that the Invariance View avoids the explanatory black-and-white dichotomy between laws (understood as exceptionless generalisations) and accidental generalisations: the only sharp boundary is the one between accidental and invariant generalisations. As fig. 3.2 illustrates, this leads to a substantially bigger set of generalisations which count as explanatory than the set of exceptionless generalisations, as generalisations may have exceptions or hold only in a limited spatio-temporal region and still be invariant and hence explanatory. This *gradation* of explanatoriness is the most appealing feature of the Invariance View, for it it depicts actual explanatoriness in science: there are plenty of causal generalisations which, despite not holding universally, are considered explanatory. [88]

Let us look at physics: most (if not all) laws are 'effective,' i.e., they hold to a high degree of approximation within a certain 'domain' or 'regime,' assumed to be specifiable independently from the laws.[89] Here examples range from Newton's law of gravity or the simple harmonic oscillator to more modern theories such as effective field theories or even general relativity. Newton's law of gravity and the simple harmonic oscillator are classical laws which describe macroscopic objects; the former breaks down when close to massive objects, for which the curvature of spacetime is important, and the latter when the oscillation of the system in question is affected by damping or friction. General relativity breaks down at the Planck length, when the effects of quantum gravity become salient. That laws hold only within a specific domain is even more evident when we pay attention to the special sciences (e.g., biology, psychology, sociology, economics, etc.), where invariance is even more local. Examples of special science laws – or *lawlike* generalisations – are the laws of supply and demand in economics, Mendel's law of segregation or Hardy-Weinberg law in biology, or Snell's law in optics (Cartwright 1983, fn.46).[90]

Before proceeding any further, let me address a couple of potential vagueness

[86] This results in Woodward and Hitchcock's (2003) account of explanatory *depth*: the depth of an explanation is a measure of how invariant the causal generalisation at its core is.

[87] And explanation simpliciter if explanation reduces to causal explanation.

[88] The explanatory benefits of characterising laws as invariant generalisations over laws as universal generalisations will be discussed more carefully in chapter 4.

[89] The domain is delimited by the sets of ICs, BCs and BaCs.

[90] Some more examples of generalisations which are believed to be lawlike even though they are not universal can be found in Reutlinger's 'A Theory of Non-universal Laws' (2011).

concerns regarding the Invariance View.

The first worry is that (*N*) draws a somewhat vague threshold between invariant and accidental generalisations, for it does not specify how many *some* interventions are. Nonetheless, I think this is a pseudoproblem: presumably *some* is more than 1, which is enough to consider the phenomena described by invariant generalisations *repeatable*. Think of the Bill Clinton – dimes accidental generalisation discussed above: it is not invariant under *any* interventions. Hence, the vagueness in the formulation of (*N*) does not undermine the Invariance View.

The second worry, however, deserves some more careful attention. This is the concern that the Invariance View does not offer sufficient conditions for lawhood. The fact that (*N*) is a necessary condition for lawhood means that if a change-relating generalisation *G* is a law, then it must be invariant under some testing interventions. There is no sufficient condition because the reverse is not true: if *G* is invariant under some testing interventions it need not be a law. That's the case of most generalisations in the special sciences, which are invariant yet not *highly* invariant. Here, a second type of vagueness crops up: what is considered a sufficiently large number of testing interventions on causal relationships for these to be called 'laws'? It can't be invariant under any intervention possible, for that would recover the view of lawhood as exceptionlessness which the Invariance View opposes.[91]

In view of this result, I believe Woodward is faced with a choice in order to underpin his Invariance View against the second type of vagueness mentioned above: either to interpret (*N*) as a necessary and sufficient condition (all invariant generalisations are laws), or to replace (*N*) with a necessary and sufficient condition for invariance along the following lines:

> **(N&S):** Iff a change-relating generalisation is explanatory, then it is invariant under some testing interventions on variables represented in the relationship.

The first option would enlarge the domain of laws from a subset of invariant generalisations to the whole set thereof. He considers, though doesn't seem convinced by this option: "We can, if we wish, stipulate that the word 'law' must be used in such a way that all invariant generalisations are laws" (Woodward 2000, p.198). The second option would displace laws altogether. The choice is then between altering the meaning of 'lawhood' in a very expansive way, or disposing of the concept altogether. In fact, even though these two sources of vagueness

[91] Note that the Invariance View does not deny that causal generalisations can be exceptionless; it only establishes that lawhood does not reduce to exceptionlessness.

pervade the exposition of the Invariance View elsewhere, in at least one place in *Making Things Happen* we find them addressed as suggested: "Invariance under at least *one* testing intervention (on variables figuring in the generalization) is *necessary* and *sufficient* for a generalization to represent a causal relationship or to figure in explanations" (Woodward 2003, p.259).[92] The choice is thus clear: it is invariance under at least one intervention what makes a generalisation explanatory, and given that lawhood is nothing more than invariance, it is invariance under interventions that is key to (causal) explanation.

3.3.3 The Methodology of Invariance

For Woodward, a good account of laws should respect their usage in scientific practice. The Invariance View is thus a *methodology of laws* because it takes into account how scientists use laws and constraints on laws for learning and reasoning about the world.[93] On this basis, Woodward's view is somewhat prescriptive insofar as it encourages philosophers of science to put our efforts in, e.g. understanding state or phase spaces – used to represent physical possibilities – and differential equations – used to represent dynamical laws (Woodward 2020, p.1928). To this end, the Invariance View is supplemented with three main independence assumptions regarding laws and ICs.[94] These independence conditions are the 'worldly infrastructure of causation,' insofar as they are the empirical presuppositions for the applicability of causal notions (Woodward 2021).[95]

In the following, I introduce the three main independence conditions underpinning invariance: (1) independence between laws and ICs, (2) independence among ICs themselves, and (3) independence among laws. These conditions constitute preconditions for our assessment of the truth-conditions of causal claims.

3.3.3.1 Independence Laws - ICs

Following Wigner (1967, pp.38-42), Woodward cashes out invariance in terms of an independence assumption between the change-relating generalisations (the laws, in Wigner's framework) and the sets of values the variables which figure in the generalisations may take (the ICs). This sort of independence is informational because "information about which ICs obtain for a system should not provide information about the laws that describe that system and conversely" (Woodward

[92] Italics mine.
[93] Well, arguably, mainly physicists.
[94] For simplicity, I focus on ICs instead of BCs, given that these are the conditions Wigner (1967) discusses.
[95] Thus, the worldly infrastructure of causation is not *constitutive* of causation.

2018a, p.164). Note that there is a connection between the laws and ICs, for ICs are the kind of things that the laws apply to, reason why laws do encode at least some information about ICs. The informational independence between laws and ICs ($INDEPENDENCE_{L-ICs}$ hereafter) is about the fact that the *values* of the ICs should be "freely assignable" independently of the laws, either by us or nature (ibid.).

What is the motivation behind $INDEPENDENCE_{L-ICs}$? Woodward interprets it as a regulative ideal for causal analysis and explanation which serves the functional goal of predicting similar invariance patterns. It also chimes with the constraints scientists impose to these invariance patterns: unlike ICs, laws are expected to satisfy symmetry conditions; e.g. physical laws ought to remain the same under spatial and temporal translations. While laws capture the order, the regularities, ICs define the space of possibilities of a system's behaviour and hence characterise the system's variability.

Of course, our ability to distinguish laws from ICs requires that nature behaves in a way that *supports* the distinction, namely, that the physics of the world gives rise to well-posed initial value problems like Cauchy surfaces. Cosmological models are particularly troublesome in this respect: the Past Hypothesis (PH), which states that the universe was at a low entropy state at the beginning of time, is treated by several accounts of laws as having nomological import despite the fact that it is a claim about ICs. So is the case for the anti-de Sitter/conformal field theory (CFT) correspondence, also called AdS/CFT duality, which establishes a relationship between theories of quantum gravity and quantum field theories. In particular, the correspondence establishes that the boundary of anti-de Sitter space (a type of spacetime with a different metric than the Minkowski spacetime) corresponds to the spacetime for a CFT. While AdS/CFT duality is a principle about BCs, it also has nomological import, for the duality alters what physical quantities count as BCs. What to make of circumstances where nature doesn't lend itself to a description in terms of laws and ICs is no cause of concern to the interventionist, for whom the impossibility to describe nature in such terms just signals the empirical, contingent nature of the split between laws and ICs. The PH, Woodward argues, should not be treated as a context in which both the notion of law and the notion of ICs apply, but rather as one in which the distinction does not make sense (see Woodward 2020, p.1921, fn.16). What matters is that there is a wide range of circumstances in which such distinction will be successful, and when it is successful, it is so because it tracks 'objective' features of nature (Woodward 2018a, p.166).

On the law vs ICs distinction, it is important to note that a combination of

laws and ICs does not always determine a system's behaviour. This is because not all the ICs we might envision for a system are combinable with laws or yield unique solutions for the behaviour of a system: e.g. there is no exact solution to Einstein's field equations of general relativity for a spacetime containing only two masses. And sometimes the converse occurs, like in the case of Maxwell's equations, where we have more equations than unknown variables, so the system's behaviour is underdetermined.

3.3.3.2 Independence IC - IC

The second independence assumption, which Woodward also draws from Wigner, is a probabilistic informational independence among ICs which characterises different properties of the same system ($INDEPENDENCE_{\text{IC-IC}}$ hereafter). The independence condition among ICs is justified on the following grounds. First, Wigner says that "there is reason to contend that these are, or at some time have been, as random as the externally imposed, gross constraints allow," (Wigner 1967, p.41) and that "... the existence of the regularities in the initial conditions is considered so unsatisfactory that it is felt necessary to show that the regularities are but a consequence of a situation in which there were *no* regularities" (ibid.). Wigner first cites Bode's law as an example of the dangers of assuming dependencies where there are none, to then defend the independence of the ICs on the basis of the explanatory success of the assumption.[96]

Note that $INDEPENDENCE_{\text{IC-IC}}$ is connected to the criterion of independent fixability: only probabilistically independent variables will be independently fixable.[97] That said, at the macro level most variables are not probabilistically independent: this is where interventions comes into play, by effectively randomising the value of variables so that they can be treated probabilistically as free variables. For instance, in the Moon-tides case study introduced in chapter 1, the gravitational force is the law of interest, which holds under testing interventions on the variables on the right hand side of the equation (the so-called 'independent' variables). The informational independence condition applied to this case translates into the fact that one might e.g. change the value of the relative distance between the Moon and the Earth, r_{M-E}, independently of the mass of the Moon, represented by M.

[96] The exact wording is as follows: "[i]n most cases, there is no reason to question the random nature of the uncontrolled, or non-specified, initial conditions, and the random nature of these initial conditions is supported by the validity of the conclusions arrived at on the basis of the assumption of randomness" (Wigner 1967, pp.41-42). He then proceeds to cite entropy-increasing examples.

[97] Recall that independent fixability was one of the criteria for variable choice (see §1.2.3.1).

3.3.3.3 Independence Law - Law

A last independence condition, commonly made by scientists in their physical theorising, completes the Invariance framework: an informational independence condition among laws ($INDEPENDENCE_{L-L}$ hereafter). For instance, alternative gravitational force laws would be consistent with the laws of motion, i.e., they would not imply alternative laws of motion (Woodward 2018a, p.173).

In Woodward's case, $INDEPENDENCE_{L-L}$ is not only motivated by pragmatic considerations of modularity, but also – and importantly – by metaphysical considerations of relative fundamentality. This fundamentality assumption permeates discussions of the autonomy of levels (see §1.2.3.3): the causal independence of higher-level variables on lower-level variables arises from the interaction of coarse-graining with facts about the underlying physical dynamics of a system (Woodward 2007, pp.87-88).[98] Even though upper-level causal claims are autonomous from lower-level causal claims, Woodward adheres to the canonical view that these bottom out in a fundamental microphysical level on which the others supervene: "given a true garden variety causal claim, there will be some associated in-principle physical explanation (or story or account, to use more neutral words) for its holding, and this will include, among other factors, appeal to fundamental laws" (ibid., p.103; Robertson forthcoming).[99] In §3.4, I will examine $INDEPENDENCE_{L-L}$ in the context of Lange's (2009) and Maudlin's (2007) views of laws, and suggest a way to similarly characterise fundamentality in the Invariance View. For now, all we need is a rough idea of $INDEPENDENCE_{L-L}$ as an asymmetrical modularity condition which establishes that changes to the content of more fundamental incur changes in the content of less fundamental laws, but the converse does not hold.

3.3.4 The Epistemology of Invariance: The Epistemic Circularity Concern

At the center of interventionism lies the PQ, which is accordingly deeply relevant to the Invariance View. The PQ seeks to answer in which sense interventions are possible (see §3.2). One of the main objections to using laws (i.e. physical possibility) to answer the PQ is the following epistemic circularity concern regarding evidence-conditions for counterfactuals: in order to assess whether a

[98] Coarse-graining is the process of mapping different values of a micro-variable into the same value of the macro-variable.

[99] As highlighted by List (2019), a locus classicus of this fundamental level assumption is Oppenheim & Putnam (1958, p.9).

change-relating generalisation is invariant one needs to know which changes are possible, and if it is the generalisations which we already know to be invariant that indicate which changes are permitted or permissible, this reasoning is circular (see Psillos 2004, p.300). As an example, he considers testing Ohm's law, $V = IR$. The idea is that to assess whether Ohm's law would remain invariant if the measurement of the intensity of the current was made on a spaceship moving at $v > c$ the law that nothing travels faster than light needs to already be in place. Hence, "Woodward's notion of *invariance under interventions* cannot offer an adequate analysis of lawhood, since laws are required to determine what interventions are possible" (idib.). In the following, I analyse whether this epistemic circularity threat is fatal.

First of all, let me me start with an observation: the epistemic circularity in question affects 'testing' interventions. 'Testing' interventions are the subset of interventions aimed at determining whether a generalisation is invariant, and hence, causal. On the basis of their specific function, I believe we should distinguish testing interventions from interventions more broadly. Consider Galileo's famous free-fall experiment: Galileo wanted to test Aristotle's hypothesis that the rate of fall of a body should be proportional to its density. To this aim, the story goes, he threw two balls of different weights from the top of the tower of Pisa, first strapped together and then each of the balls individually. To his surprise, except for some air friction they all (individual balls and composite system) fell at the same speed. This allowed him to uncover a generalisation for the average velocity of any falling body: $v_a = \frac{1}{2}gt$. Most people agree that Galileo did not in fact perform the experiment, but only calculated this result.[100] Whether it is a thought experiment concerns how invariance is *identified*, how we learn about worldly invariance: whether by throwing the balls from the top of the tower (actual intervention) or by calculating how they would fall (hypothetical intervention).

Once invariance has been identified, the breadth or scope of invariance can be investigated via further interventions.[101] Determining the scope of invariance alludes to the conceptually distinct and theoretically posterior phase of investigating the scope of the generalisation in question, so it presupposes the previous identification of a dependency relation as invariant. The interventions we may perform to determine the scope of an invariant generalisation are no longer *testing* interventions. In the case at hand, one could throw many more balls and from many more heights to establish the limits of the generalisation – or alternatively, calculating the effects these hypothetical interventions would have. For instance,

[100] See T. S. Gendler (1998) for a detailed reconstruction of the thought experiment.
[101] The notion of 'breadth' of invariance will be introduced in §3.4.

by actually throwing many and different balls we'd realise that the calculated ratio of fall is only valid if friction is discounted. And by considering stratospheric heights we'd notice that g can no longer be considered a constant. In practice, however, the two stages are intertwined, and whether we may call a regular intervention 'testing' will depend on whether the goal with which interventions are performed is specified.

After the above parenthetical remark, let us return to the epistemic circularity concern: to determine whether a generalisation G is invariant one needs to know under which interventions G is invariant, but this already presupposes some invariance. Woodward's reply to this objection is somewhat sneaky: possibilities can be identified via experiment without us knowing whether a given generalisation is invariant! After all, invariance "is something for which we can get empirical evidence" (Woodward 2018a, p.160). One can throw a rock and observe whether it falls (Woodward 2018a, p.171), or move a conductor in a magnetic field and observe whether an electrical current is induced (Woodward 2013, p.68). Actual interventions provide evidence for invariance: through experimental manipulations we are able to directly test whether a certain generalisation is invariant under a set of changes enacted by us or nature. For instance, the invariance of Coulomb's law's may be tested by altering the distance between opposite charges, which keep attracting each other no matter how far apart they are placed. It can also be that experiments cannot be performed because of ethical or technological reasons, but experimenters can profit from naturally occurring variations such as genetic randomisation. Either way, be it via experimental or observational exploration, it is quite clear how both sub-types of actual interventions provide evidence-conditions for causal claims.[102]

Nonetheless, empirical evidence is not directly available for all causal claims. For instance, cosmologists and astrophysicists would claim that the universe's accelerated expansion is caused by the presence of dark energy. But how do we know whether this causal claim is true? Clearly, we can't *actually* intervene on the presence of dark matter, but only wiggle its value via simulations using a general relativistic model. Woodward argues that in these cases the invariance of causal generalisations is checked against sets of hypothetical interventions on ICs that known laws and principles – together with domain-specific assumptions – allow (Woodward 2013, p.68), but that this does not need to lead to *vicious* circularity: to characterise the notion of an intervention on X with respect to Y all that is needed is information about the causal relationship between I and X. It is facts

[102]Interventions enacted by us or by nature respectively align with the experimental studies or observational studies (the natural experiments), as discussed in §1.3.1.

about *other* causal relationships, say G' and G'' involving I, X, and Y, together with counterfactual claims about how Y responds to interventions on X, which explain the causal relationship $X - Y$ (Woodward 2003, p.105).

By way of example, let us analyse one such example of a hypothetical testing intervention discussed by Woodward: the (near) uniformity of the cosmological microwave background (CMB) in all directions in space (Woodward 2013, pp.67-68). Woodward argues that it is not regarded as a law because this (near) uniformity is contingent on the ICs at the early universe. To reach this conclusion, "[...] one relies on previously accepted theoretical considerations to tell us that early initial conditions might have been different, but [...] one does not have to already know or presuppose that the cosmological uniformities themselves are physically contingent" (id., p.68). In this case, we know that the cosmological inflation model presupposes that matter and energy of all the observable universe were in thermal equilibrium before it started expanding, and this is what explains the even distribution of the CMB. However, it is possible that the early universe had not been in thermal equilibrium. Therefore, we know that the uniformity of the CMB is not invariant under other than actual ICs at the early universe, for in that case it would not have been evenly distributed.

The interventionist might have the resources to avoid vicious circularity about hypothetical interventions; this will hinge on whether the case at hand is a case of causal discovery or a case of model-based causal reasoning. If it is an instance of causal discovery, viciousness can be avoided by ensuring that finding about the invariance of G does not already presuppose its invariance (as in the case of the CMB). Nonetheless, one might still argue that the interventionist account of causation sets unreasonably high epistemic demands, for the evaluation of hypothetical interventions requires a substantial amount of background causal knowledge. This might be problematic when we are faced with complex causal structures, for which we might not know on what causal structure to rely on, as well as new phenomena, for which previous causal knowledge might be scarce.

If the hypothetical 'intervention' at hand is an instance of model-based causal reasoning, one might argue that the notion of an intervention is superfluous. I agree. However, let us not forget that interventions are needed for causal discovery: after all, invariance is ultimately empirically testable. This means that G might be hypothetically tested with the knowledge that G' and G'' are invariant, but these inferences must eventually bottom out in actual interventions. This is an idea I already discussed in the context of Chalmers' conceivability-possibility thesis (see §2.3.4), where I made reference to the distinction between actualist-manipulationist and relative modalities in science drawn by Hirvonen et al. (2021). Most modal

inferences are relative to the content of a model/theory, which we developed on the basis of knowledge acquired via actual interventions: the latter is what guarantees that causal models are empirically adequate, and the possibilities they encode are not merely conceptual.

In short, the possibility of hypothetical testing interventions is facilitated by previously accrued causal information about generalisations other than the one being tested, information which eventually bottoms out in empirically tested invariance. Vicious circularity can this way be avoided, though evaluating the associated causal claims requires a significant amount of causal information – this makes interventionism an epistemically demanding theory of causation and causal explanation.

3.3.5 The Metaphysics of Invariance: The Modal Priority Question

In the vicinity of the epistemic circularity concern lies another circularity worry. Since interventionism is a realist account of causation, the 'non-reductive' character of interventions is underwritten by a metaphysical circularity worry concerning the truth-conditions for interventionist counterfactuals, which, since interventions need not be causal, are not reducible to their evidence-conditions. The metaphysical circularity worry here is, phrased in terms of invariance, the following: do invariant relations make interventionist counterfactuals true, or do interventionist counterfactuals make invariant relations true? In short, what grounds what?

The Invariance View has been criticised on the basis that it provides no metaphysical story about truth-makers or foundations for laws, either in terms of reduction or special entities like dispositions or powers, so the invariance framework is like "Hamlet without the prince" (Woodward 2018a, p.160).[103] Mocking the critics, Woodward has self-proclaimed to be "an ontological philistine" (Woodward 2017, p.224). His dislike for metaphysics is no news. He has repeatedly and firmly argued that the task of postulating metaphysical underpinnings for laws is ill-posed from the start: reductionist accounts of laws are unclear about notions that figure in these, and about what it is they are trying to reduce (ibid., p.225).

But more importantly, even if there was a clear story about what the truth-makers for laws are, Woodward thinks these would not be useful for the epistemic or methodological purposes he is after, such as explaining the role the notion of 'law' plays in our reasoning and decision making, what evidence is needed to

[103] A similar widely-shared concern affects interventionism more broadly: Woodward provides no truth conditions for causal claims, which may lead to circularity worries (Strevens 2008, p.245).

establish law claims, or how this notion connects to other scientifically important ones (Woodward 2017, p.180). For instance, reifying chance as the thing that explains the behaviour of roulette wheels (Hicks & Wilson forthcoming), or even reducing it some something else (e.g. relative frequencies or some 'fit' to these), would not tell us anything about the behaviour of such systems (Woodward 2018a, pp.177-179). But the problem is that if nothing explains the facts about the invariance of the behaviour of systems under ICs, it is hard to see how one can endorse realism and, therefore, a (partly) ontic account of explanation, without getting into the metaphysics of causation and laws.[104]

The Invariance View is a non-reductive account of laws, in the sense that it doesn't aim to reduce laws, which have modal import, to non-modal facts. In this respect it is similar to Maudlin's (2007) and Lange's (2009) views of laws (to be introduced in upcoming §3.4). However, it differs from these in that it does not specify what the basis of the modal space is, whether it is laws, counterfactuals, or something else. Traditionally, laws have been considered to 'support' counterfactuals, that is, to act (partially, at least) as truth-makers for counterfactuals. If then counterfactuals are called for to distinguish invariant from accidental generalisations, this raises a metaphysical circularity concern: laws provide truth-making grounds for counterfactuals, and the latter provide truth-making grounds for laws. Maudlin's account of laws breaks this circularity by endorsing a truth-making relation from laws to counterfactuals, which is reversed in Lange's account of laws. Nonetheless, according to Woodward, "laws and counterfactuals are so entangled that there is no basis for claiming one of these is more fundamental than the other" (Woodward 2020, p.1913, fn.7). In order to avoid this metaphysical circularity concern, Woodward ought to answer the following chicken and egg question: what is metaphysically prior in the modal domain, facts about counterfactual dependence ('counterfacts'), or facts about lawhood, i.e. invariance? For future reference, I call this question the 'Modal Priority Question' (MPQ).

In order to get a sense of the MPQ, consider the kinds of counterfactuals which would make a law L true. These are counterfactuals which describe testing interventions on change-relating generalisations, and take the following general form (Woodward 2020, p.1911):

> (C) If such and such ICs reflecting possible values of the variables in L were to be realised, L would continue to hold.

Counterfactuals which fit the structure of (C) are true whenever there are possible

[104] It is *partly* ontic because it incorporates pragmatic and epistemic considerations (see Jansson (2018)).

interventions on ICs by nature or us that would realise the antecedents in such a way that the states of affairs described by their consequents would thereby be made to hold. Note that (C) restricts the classes of counterfactuals which are relevant to assessments of invariance to counterfactual suppositions *about ICs*. Counterfactual suppositions which are *not* about ICs are not relevant to assessments of invariance (Woodward 2020, p.1912); it is for this reason, I presume, that Woodward refuses to take counterfactual dependence as primitive and then use it to explicate the notion of a law. While his response to the MPQ is non-committal, in the upcoming §3.4.3 I will argue that adopting a methodological stance does on this issue not exempt him from taking an ontological stance, which I think should prioritise facts about invariance over counterfacts.

3.4 The Modal Priority Question Revisited

Woodward's Invariance View is non-reductive, in that it doesn't reduce laws to non-modal facts. In this respect, it is much like Lange's (2009) and Maudlin's (2007) accounts of laws. However, as I described in §3.3.5, the Invariance View differs from these because it doesn't *order* modal facts: what's at the bottom of the modal domain, facts about invariance or counterfacts? This is what I called the Modal Priority Question (MPQ). In this section I answer this question, by providing what I think is the best metaphysical completion of Woodward's Invariance View: a primitivism about invariance. In order to do so, I compare and contrast Lange's and Maudlin's accounts of laws. The results of this comparison bring to the fore an assumption that underlies all of these accounts, which builds upon $INDEPENDENCE_{L-L}$: laws are not only independent among themselves, but they are also vertically ordered, the reason why $INDEPENDENCE_{L-L}$ is an asymmetrical condition.

Woodward shares the idea that some invariant relationships are more fundamental than others: "... from the point of view of the assessment of invariance in physics [fundamental laws] are more important than invariance under other sorts of changes," they have a "privileged status" (Woodward 2000, p.220). I contend that Woodward can account for fundamentality in terms of the *breadth* of invariance: the more changes on ICs and BaCs a causal generalisation withstands, the higher its degree of fundamentality. This idea will play a key role in §3.5, where I will propose an extension of the Invariance View to counterfactuals which describe changes in the laws.

3.4.1 Marc Lange's View of Laws

Marc Lange's account of laws is very elaborate and technical, so in the following I will give a brief exposition of the elements which are relevant to establish a comparison between Lange's and Woodward's views (see Lange 2009 for full details).

Like Woodward, Lange connects lawhood to counterfactual stability. For instance, if it is a law that "At standard pressure (i.e. 1 atm) water boils at 100ºC," then the corresponding counterfactual "If this sample of water at standard pressure were heated at 100ºC it would boil" is also true. However, in contrast to Woodward, for whom counterfactuals and laws are ontologically on a par, and against the general view that laws are ontologically and conceptually more fundamental than counterfactuals, Lange takes counterfactual facts as primitive facts which *make* the laws: these counterfacts or 'subjunctive facts', as Lange calls them, ground the fact that the laws would have held even if various other things had been different. This stability under counterfactual suppositions is what confers laws their nomological – or natural – necessity. However, not all counterfactual suppositions are relevant for lawhood. In order to distinguish laws from accidents, he coins the notion of *NOMIC PRESERVATION* (*NP*) to capture the idea that the relevant counterfactual suppositions to consider are those which are consistent with *all* the laws taken together:

> **NP:** *m* is a law if and only if *m* would still have held under any counterfactual (or subjunctive) supposition *p* that is logically consistent with all of the laws (taken together). (Lange 2009, pp.13-14)

However, *NP* turns out to be circular, because laws are not picked out on external grounds. Lange amends NP by extending it to all truths about sub-nomic facts, namely, all counterfactual facts except for facts about which things are or aren't laws (Lange 2009, p.28). For instance, sub-nomic facts include the fact that all emeralds are green, or that all gold cubes are smaller than a cubic mile, but not the fact that the former is a law or the fact that the latter is an accident. These sub-nomic facts conform the basis of non-modal facts about the world, which laws might *govern* but to which they don't reduce (id., pp.17-18). It is thus *SUB–NOMIC STABILITY* (*SNS*) what distinguishes laws from accidents:

> **SNS:** "Roughly speaking, a set of sub-nomic truths is "sub-nomically stable" if and only if whatever the conversational context, the set's members would all still have held under every sub-nomic counterfac-

tual (or subjunctive) supposition that is logically consistent with the set – even under however many such suppositions are nested." (id., p.29)

For instance, classical mechanics conceives of $F=ma$ as belonging to a proper subset of truths that is sub-nomically stable because $F=ma$ would continue to hold under the supposition that various specific force laws do not hold. More generally, the structure Lange presents is the following hierarchy of nomic and sub-nomic facts: at the base there are the sub-nomic facts, where the facts on a given rung of the hierarchy are 'governed' by some of the facts one rung above, which constrain yet not fully determine them (e.g. first-order laws cannot fully determine the sub-nomic facts without ICs):

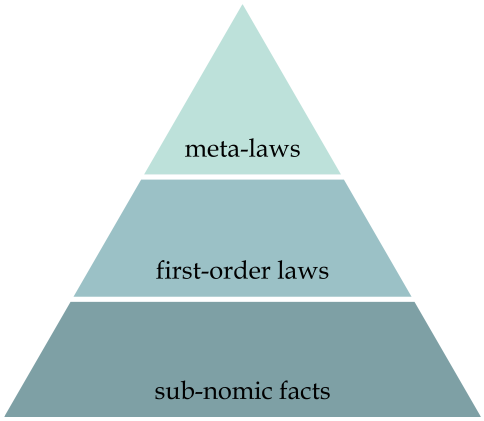


Figure 3.3: Lange's view of laws.

The facts in the top rung in fig. 3.3 specify meta-laws that govern the first-order laws. One rung below one finds the facts that the meta-laws govern, e.g. the fact that it is a law that all electrons have negative electric charge, or that it is not a law that all gold cubes are smaller than a cubic mile. Finally, the bottom rung are the sub-nomic facts, e.g. that all like charges repel. This stratified laws picture is motivated by Lange's reconstruction of past scientific practice:

> "In 1830, George Biddell Airy used Newton's second law of motion to figure out how bodies would have behaved had they been subjected to various weird hypothetical kinds of forces. His investigation presupposes that Newton's second law of motion would still have held, even if the force laws had been different. Similarly, Paul Ehrenfest in 1917 famously showed that had gravity been an inverse-cube force or fallen off with distance at any greater rate, the planets would eventually have collided with the sun or escaped from the sun's gravity. Ehrenfest's argument also presumes that Newton's second law of motion would

still have held, had gravity obeyed a different force law."[105] (id., p.39)

In short, both Airy and Ehrenfest showed that *F=ma* would have held even if the force laws had been different than the actual ones. And likewise for the law of the composition of forces, which via the parallelogram of forces specifies how component forces add to yield a 'resultant force' or 'net force.'[106] The same applies to conservation laws and Lorentz transformations: many believe they would have held had the force laws been different (see e.g. Lévy-Leblond 1976, p.271).[107]

There is a clear parallel between Woodward's $INDEPENDENCE_{L-L}$ and the independence assumption among laws in Lange's view. Furthermore, Lange is explicit about the one-directional nature of the independence between laws condition: as mentioned in §3.3.3.3, while this asymmetry permeates Woodward's laws' modularity as well, it is never made explicit. While Lange has been criticised on the grounds that he has said next to nothing about what primitive subjunctive facts are or how they are related to subjunctive conditionals (Woodward et al. 2011, p.37), his hierarchy of stable sets (i.e., stable sets stand in the subset relation) thus provides a more metaphysically robust laws picture than Woodward's. I will dwell on this observation in §3.4.3.

3.4.2 Tim Maudlin's View of Laws

Maudlin (2007, 2020) follows a converse strategy to Lange's: he takes dynamical laws of nature and the direction of time to be the primitive ingredients of our ontology, which ground the truth of counterfactuals about nomic facts.

Like Woodward, he thinks of laws as scientific equations, so as scientific laws. Laws are "mathematically formulated constraints" (Maudlin 2020, p.3), the content of which determines a class of models, distinguished only by their ICs and BCs.[108] These models represent "physically possible states of affairs" (ibid.). In other words, these models can be treated as 'possible worlds' consistent with those laws, which provide truth conditions for claims about nomological possibility and necessity: e.g. the actual laws turn out to be necessary because they obtain in all the models, where the necessity of the laws is "a simple by-product of the semantics of modality" (Maudlin 2020, p.6). Since the content of physical theories is exhausted by the models of their dynamical laws, formulated as differential equations – this is what M. Frisch (2014, p.126) calls a *quasi*-dogma in philosophy

[105] See Airy (1830) and Ehrenfest (1917).
[106] The parallelogram of forces is a corollary of Newton's second law (Lange 2009, p.40, fn.39), though Lange notes that not everyone treats it as such, e.g. Poisson doesn't.
[107] Cited in Lange (2009, p.41), as well as Lange (2013, pp.503-504).
[108] Note that provided that the laws are deterministic only one such model is generated.

of physics – dynamical laws are, according to Maudlin, the best candidates for the basic ontology of the world. Given that modality does not come from anything outside the laws, Maudlin's proposal is metaphysically 'modest' (Maudlin 2007, pp.5-49).

Regarding the vertical structure of laws, Maudlin's account of laws bottoms out in the Fundamental Laws of Temporal Evolution (FLOTEs), such as $F=\frac{d(mv)}{dt}$ in Newtonian mechanics, or Schrödinger's equation in non-relativistic quantum mechanics $i\hbar\frac{\partial|\psi\rangle}{\partial x}=H|\psi\rangle$. FLOTEs evolve physical states through time in some region of state space. These laws are aided by adjunct laws which provide information about the magnitude of the forces and the form of the Hamiltonian (such as Newton's law of universal gravitation), or about the physical states allowed (such as the ideal gas law $PV=nRT$). Analogously to FLOTEs, Maudlin also postulates Laws of Temporal Evolution (LOTEs), the dynamical laws in the special sciences guiding the evolution of higher-level systems (e.g. the Hardy-Weinberg law of population genetics). They take a similar form to the FLOTEs but are less law-like since they admit exceptions, and ontologically depend on the former. The primitivism of (F)LOTES stems from a pragmatic scientific standpoint: they make diachronic regularities in scientific explanations intelligible (Maudlin 2007, p.14).

Having expounded the primitivism of the dynamical laws of nature, let me say a few words about Maudlin's primitivism about time's arrow. The idea behind the postulation of a primitive arrow of time is that something ought to explain the time-irreversibility of the observed regularities, namely, the fact that e.g. dropping a ceramic tea cup causes it to shatter and the tea to get spilled, but not the other way around. One popular way to explain the irreversibility of macrodynamics is via Albert's (2000) Past Hypothesis (PH): as a low entropy constraint on the early universe.[109] The causal arrow of time is thus connected to the thermodynamical arrow of time via the 2^{nd} law of thermodynamics: the entropy state of the entire universe, considered an isolated system, will always increase over time.[110] Contrary to Albert, Maudlin thinks this asymmetrical treatment of the boundary conditions of the universe is a reflection of the passage of time; the entropy gradient does not explain the direction of time, but is explained by it. Hence, it is the *flow* of time that is primitive (Maudlin 2007, pp.108-110); the world has an irreducibly intrinsic asymmetry in its temporal structure.

Once we have the time-directed dynamical laws as primitives in our on-

[109] Another way to understand the PH is via Wallace's (2011) account, which characterises it as a no-correlations initial assumption: correlations can only be created as we move forward in time and causings occur.

[110] See Zeh (1999) for a discussion of the relations between the various arrows of time.

tology, we are in a position to construct and evaluate counterfactuals. These are evaluated by means of the laws via a three-stage recipe, which basically offers instructions as to how to construct a Cauchy surface intersecting the event which the counterfactual antecedent under study describes (step 1), save that the relevant physical properties are not given their actual values but the values specified in the counterfactual (step 2). Then, one allows the laws to operate on this Cauchy surface; this generates a model or set of models with the new BCs (step 3).[111] The counterfactual in question has a determinate truth value if the consequent obtains or fails to obtain in the model or models generated. This recipe can also handle stochastic cases, though since in these cases events are also affected by the random outcomes of stochastic processes, a restatement of the ceteris paribus condition of step 2 is needed.[112]

Important for the extension of the Invariance View which I shall propose in §3.5 is that Maudlin's recipe can also handle counterfactuals which describe what would have happened if the laws were changed. The idea is that when evaluating these counterfactuals, one constructs the relevant Cauchy surface but instead of altering the BCs one changes the laws which the antecedent puts on hold or violates, while keeping all else equal. For instance, in asking whether the Earth would continue to orbit around the sun should one set G=0 we ought to ensure that we're entitled to run Newtonian mechanics as usual, or that the laws of thermodynamics won't change. Note that the modularity assumption regarding laws which underwrites Maudlin's recipe here resembles Woodward's $INDEPENDENCE_{L-L}$.

3.4.3 Woodward in Perspective: The Modal Priority Question Answered

In §3.4.1 and §3.4.2, I introduced two other prominent non-reductive views of laws besides Woodward's Invariance View: respectively, Lange's and Maudlin's accounts of laws. In this section, I examine how Woodward's Invariance View compares – in the letter and the spirit – to these two views.

Like for Lange, lawhood is for Woodward related to counterfactual stability. Unlike Lange, though, he does not think that counterfactual stability is what

[111] See Maudlin (2007, pp.23-24) for more details. Just like the evaluation of interventionist counterfactuals requires that interventions be well-defined, Maudlin's recipe to evaluate counterfactuals requires well-defined initial value problems, where the relevant properties are described by independent variables.

[112] Such restatement of the ceteris paribus condition for stochastic cases is detailed in Maudlin (2007, pp.30-31).

confers laws their nomological necessity. As we saw in §3.3.5, it is only counterfactuals about ICs which are relevant to judgements of invariance. In this respect, his view aligns with Maudlin's: modality does not come from anything outside the laws (in combination with ICs and BaCs, that is). For instance, combining the laws of motion with an alternative to the inverse square law is a matter of logical 'consistency' or 'coherence' of the latter with the former (Woodward 2018a, p.173, fn.47).[113]

The independence between laws is precisely another commonality across the three accounts discussed, and in the three cases comes with a bonus fundamentality assumption. The way Lange and Maudlin justify fundamentality is by appeal to primitive facts – be these counterfactual or nomological – and relations to these, which define how fundamental the facts they stand in a relation to are. Woodward does not provide any explanation of what makes some laws more fundamental than others, but I believe he has all the tools needed to account for it: fundamentality is none other than *breadth* of invariance. What makes e.g. the laws of relativistic mechanics more fundamental than the laws of Newtonian mechanics is that the former are invariant under a wider range of changes in the velocity of moving objects than the latter (i.e. the former set includes changes to relative velocities comparable to the speed of light c, not described by the latter).[114]

Fundamentality qua breadth of invariance is a *relative* notion, for it requires at least two comparative elements about the same subject matter, say laws A and B, with some overlap in the conditions under which they are invariant. Note however that this does not make the notion epistemic: interventions are not agent-relative; they are objective possibilities. The fact that law A is more or less invariant than law B under interventions is mind-independent, that is, it does not depend on my knowledge of these two sets of interventions. Therefore, the resulting ordering of laws is objective.

However, as reiterated through this thesis time and again, Woodward does not provide metaphysical underpinnings for the causal realism of interventionism. At best, his argument for realism is a version of the 'no miracles' argument for scientific realism, which is an inference to the best (because of only) explanation type of argument: since our causal reasoning has proved successful in providing correct predictions and explanations, the argument goes, it must be so because it captures true causal facts. Since invariance is all there is to causal generalisations and laws, I believe this 'no-miracles argument for causation' generalises to invariance as

[113]This is much in line with the discussion of concerns about modal rationalism (see §2.3.4).

[114]This notion of fundamentality as breadth of invariance aligns with the resulting *depth* of the explanation (see Woodward & Hitchcock 2003).

follows: the existence of invariant relationships is the only hypothesis that does not make the predictive and explanatory success of laws a mystery. Should no facts about invariance underlie counterfactuals about ICs (i.e. counterfactuals of form (C), see §3.3.2), we wouldn't have any good basis for asserting their truth/falsity.

Despite Woodward's anti-metaphysical stance, I believe his brief attempts at providing a justification for the realism of interventionism points in the direction of a primitivism about invariant relationships ('primitivism about invariance' hereafter), which gets hidden under the rug because of Woodward's insistence that describing how the world is won't help us navigate it. I agree that the former need not entail the latter, for the latter has pragmatic overtones. However, as he acknowledges, how the world is does *explain* why we reason the way we do, and so under any ontic construal of explanation it ought to be ontologically prior too.

Furthermore, a primitivism about invariance would align with Woodward's rather pragmatic or 'functional' goals regarding causes and laws (see Woodward 2014) in two respects. First, assuming the existence of invariant relations as brute does not require that we provide grounds for them, so no full-blown metaphysical story is needed. Second, and relatedly, it does not require that one resorts to notions of metaphysical modality. One could make do with physical modality:

> "Invariance-based accounts provide a *naturalistic*, *scientifically respectable* and non-mysterious treatment of what non-violability and physical necessity amount to: this just amounts to the claim that within the domain of invariance of a law there are no initial and background conditions that might be realized – nothing that might be done by nature or an experimenter – under which the law will fail to hold."[115] (Woodward 2018a, p.160)

That is, the most highly invariant generalisations in science hold with physical, not metaphysical necessity. Physical necessity is total invariance within a law's domain, and this is all that is needed to explain their modal import. In this respect, the Invariance View is in spirit very close to Maudlin's project, whose view is also motivated by scientific definitions of laws and therefore identifies physical modality as the relevant kind of modal notion.

Supplementing the Invariance View with a primitivism about invariance would render Woodward's 'minimal' causal realism for interventionism a truly 'small m' project, or "a kind of tempered and restricted realism" (Woodward 2018a, p.167) like Maudlin's 'modest' realism for laws: it would explain what

[115]Italics mine.

makes counterfactual claims true.[116] The result would be an objectivist account of counterfactuals and causal explanation.

All in all, while Woodward's usual response to the MPQ is to shy away from metaphysical projects, I think he should rather answer it by adopting a primitivism about invariance (plus, following Maudlin, a primitivism about time's arrow), where invariance is physical dependence backing explanations. This view does better than Lange's in the sense that it restricts counterfactual explanatoriness to physical counterfactual dependence. Primivitism about invariance is a very low-key metaphysical commitment which not only aligns with practice but also safeguards the Invariance View against metaphysical and epistemic circularity threats.

3.5 The Extended Invariance View

3.5.1 What Would Have Happened Had the Laws Been Different

In this section, I argue that Woodward's Invariance View offers tools to address a certain kind of commonplace counternomic reasoning in science: reasoning about what would happen were the laws different. These hypothetical situations correspond to not too far-off possibilities which scientists themselves – such as Airy or Ehrenfest (see §3.4.1) – have taken seriously. This type of counterfactual reasoning trades in counterfactuals such as the following:

> C_{UG-N2}**:** Had Newton's law of universal gravitation been different, $F = ma$ would have continued to hold.

Recall that Newton's law of universal gravitation takes the following form:

$$F_{12} = G\frac{m_1 m_2}{r_{12}^2} \tag{3.2}$$

In eq. (3.2), F_{12} is the gravitational attraction between two bodies, of masses m_1 and m_2, r_{12} is the distance between their centers of mass, and G the universal gravitational constant, with a measured value $G \approx 6.67428 \cdot 10^{-11} m^3 kg^{-1} s^{-2}$.

As it stands, the original Invariance View accommodates counterfactuals which take the form (C), namely:

> (C) If such and such ICs reflecting possible values of the variables in L were to be realised, L would continue to hold.

[116] The 'm' refers to metaphysics.

Applied to the universal gravitation ('UG-L' hereafter), the Invariance View accommodates counterfactuals which consider whether UG-L would continue to hold had the values of m_1 and m_2, r_{12} been other-than-actual.

However, the Invariance View makes no room for counterfactuals like C_{UG-N2}, a counterfactual which considers what would happen were UG-L different. This seems problematic, for we can perfectly make sense of counternomic scenarios like these. When Ehrenfest entertained the hypothesis of gravity being an inverse-cube law (that is, space having more than three dimensions), this hypothesis being contrary to nomological fact did not preclude him from calculating that it would lead planets to collapse onto the Sun or "fl[y] away infinitely" (Ehrenfest 1917, p.200).[117] This counterfactual, just like C_{UG-N2}, are perfectly meaningful. I believe the Invariance View has the resources to deal with these counternomics that describe how changes to a law would impact other laws if extended as described in the following section.

3.5.2 What is the Extended Invariance View?

The Extended Invariance View is an invariance-based account of laws which draws on the objective ordering of invariant generalisations according to their *breadth* and thereby allows us to make sense of counternomics like C_{UG-N2} (see §3.5.1). Just like Maudlin's account of laws, which takes laws as ontologically primitive, the Extended Invariance View takes invariance as the primitive ingredient of our ontology. And like Lange's, it is a hierarchical view in which invariant relationships are ordered according to their modal robustness under interventions.

As we saw in §3.4.3, the breadth of invariance is a measure of the degree of invariance of causal generalisations and thereby accounts for their fundamentality.[118] This provides an objective ordering of generalisations as depicted in fig. 3.4:

[117] This prediction can be more easily made sense of if one recalls that UG-L was derived from Kepler's laws of planetary motion, in particular, the 2^{nd} (a radius vector joining any planet to the Sun sweeps out equal areas in equal lengths of time) and the 3^{rd} (also known as law of the areas, the squares of the sidereal periods of revolution of the planets are directly proportional to the cubes of their mean distances from the Sun), neither of which will hold. See 'Kepler's laws of planetary motion' for more details.

[118] And non-causal generalisations. I say some more about this topic in §3.5.4.1.

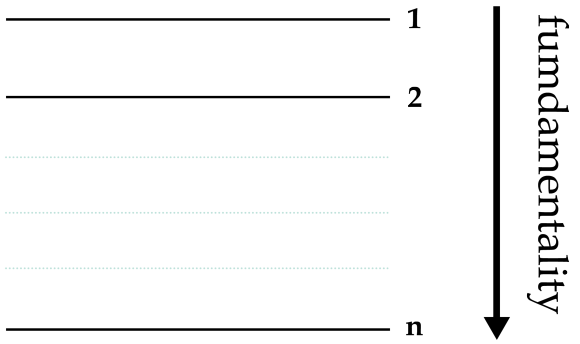


Figure 3.4: Breadth of invariance provides an objective ordering of generalisations.

The Extended Invariance View relies on the fundamentality picture depicted in fig. 3.4 to establish the truth-conditions of counternomics that describe situations in which some laws are false. It does so by acknowledging that while the counterfactuals in question are not supported by certain invariant relationships, they are nonetheless supported by invariant relationships which are *higher-order* with respect to them. It is in virtue of these higher-order invariant relationships that these counternomics are true.

Let us exemplify the view by focusing on the situation described by C_{UG-N2}. Given that C_{UG-N2} entertains a situation in which UG-L does not hold, it can't be made true by UG-L. Let us call UG-L a 1^{st} order invariant generalisation; the situation described by C_{UG-N2} does not preserve 1^{st} order invariance with respect to UG-L.

Nonetheless, $INDEPENDENCE_{L-L}$ leaves room for other types of invariance being preserved in such counternomic scenarios. Since UG-L is not derived from invariant generalisations like conservation laws, special relativity's second postulate, or even Newton's 2^{nd} law of motion, the non-invariance of UG-L does not preclude the invariance of these other relationships which have a broader range of invariance (i.e. they are more fundamental) than UG-L. Since I have labelled UG-L a 1^{st} order invariant relationship, relationships with a broader range of invariance than UG-L will be referred to as *higher-order* – i.e. 2^{nd} order and above – invariant relationships.[119] The Extended Invariance View enables us to say that, while the antecedent of C_{UG-N2} is not made true by the 1^{st} order invariant relationship whose content is considered other-than-actual, it is nonetheless made

[119]While the ordering of invariant relationships is objective, the label – 1, 2, ..., n – will depend on the explanatory context and the observer's state of knowledge. We will reserve the label '1' for the invariant relationship described in the antecedent, and n ≥ 2 for generalisations with a broader range of invariance with respect to it (i.e. more fundamental, see fig. 3.4).

true by invariant relationships which are higher-order with respect to it.

In §3.5.3, I analyse how counternomics like C_{UG-N2} can be interpreted as interventionist counterfactuals: I argue that they describe 'higher-order interventions,' that is, possible interventions with respect to higher-order invariant relationships. Next, in §3.5.4 I examine the metaphysical status of higher-order invariant relationships, as well as their explanatory role with respect to lower-order invariant relationships.

3.5.3 The Possibility of Higher-Order Interventions

What would have happened had UG-L been different? This sort of situation, I presume, invites one to consider alternative invariant generalisations describing facts about gravity. For instance, one might envision a law of gravitation with different values for the gravitational constant G, or exponents other than $n = 2$ for r_{12}^n.[120]

However, the possibilities described above don't straightforwardly correspond to other-than-actual ICs for UG-L, for these are not *variables* in UG-L. That said, there is reason to treat the gravitational constant G, or the n in r_{12}^n as ICs – or, more broadly, BCs – with respect to higher-order laws. Note that G also features in other basic laws of nature, such as Einstein's field equations in General Relativity, so it is independently fixable from UG-L. And gravity could have been an inverse-cube law instead of an inverse-square law had the dimensionality of space been different. I thereby propose to interpret these as possible ICs with respect to higher-order laws – to differentiate these from regular ICs, I shall call them 'ICs^n', where n indicates the order of the higher-order law in question.[121]

These sorts of counternomic possibilities are not consistent with *all* actual invariant dependencies, but since they are consistent with at least *some* actual invariant dependencies there is a sense in which they are *physical possibilities*. In line with the notions of strong physical possibility (SPP) and weak physical possibility (WPP) described in §1.3.2.1, I suggest to coin an even weaker type of physical possibility to characterise the possibility of changes to the content of invariant relationships which are nonetheless consistent with higher-order invariant relationships. I call it 'Very Weak Physical Possibility' ('VWPP' for short). A change on the ICs^n of invariant relationships is VWPP iff it is consistent with:

[120] As we will see in §5.3.2, research in cognitive psychology shows that when considering counterfactual scenarios we tend to imagine situations as close as possible to the actual one.

[121] Note that this n is not the same as the aforementioned parameter indicating the dimensionality of space.

Very Weak Physical Possibility (VWPP): the actual higher-order invariant relationships and some range of ICs^n consistent with the actual higher-order invariant relationships including the actual ICs^n.

In light of the critical analysis of §3.3.2, where I argued that Woodward should drop talk of laws altogether, I suggest to reformulate SPP and WPP as follows. An intervention is, respectively, strongly physically possible and weakly physically possible iff it is consistent with:

Strong Physical Possibility (SPP): the actual invariant relationships and the actual ICs.

Weak Physical Possibility (WPP): the actual invariant relationships and some range of ICs consistent with the actual invariant relationships including the actual ICs.

Note that VWPP is an even weaker notion of possibility than WPP, for VWPP only requires consistency with higher-order invariant relationships. At the same time, WPP is weaker than SPP, which requires consistency with all actual invariant relationships and actual ICs. These notions of physical possibility are centered on actuality as shown in fig. 3.5:

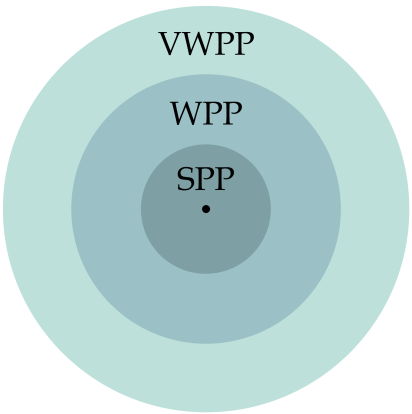


Figure 3.5: Notions of physical possibility (strong, weak, and very weak) centered on actuality.

With respect to VWPP, note that *PROCESS IRRELEVANCE* ensures that it is of no concern how these interventions come about: what matters is that they are consistent with higher-order patterns of invariance. Given that these are only possible with respect to higher-order invariant relationships, I shall call these 'higher-order interventions.'

Having established that changes on ICs^n are (very weakly) physically possible, the next question we ought to address is whether these are *really* interventions. Since the relevant variables are independently fixable, there is reason to expect they might be. However, should one think these changes to the values of ICs^n do not qualify as interventions because they are not *ideal* – i.e. they don't satisfy (IV) – this would pose no significant challenges to the present account. First of all, besides invariance under interventions and BaCs, Woodward's Invariance View accommodates invariance under changes on ICs which aren't interventions ('non-I-changes' for short, see Woodward 2003, p.248). One such example would be MT-I$_2$: an impact of a planetoid on the Moon would not only change its mass but also its distance to the Earth. Second, as I now explain, either characterisation is compatible with the non-causal explanatory role of the relevant dependence.

The sort of explanations these counternomics target aim to answer how interventions on the content of lower-order invariant relationships would affect higher-order invariant relationships. The dependence in question is not causal because such interventions leave the higher-order invariant facts intact. As I will argue in §3.5.4.1, higher-order interventions are best made sense of in the context of non-causal explanation. Intervention *liberalists*, who are sympathetic to using interventions in non-causal explanation, would find this view plausible. Intervention *puritanists*, for whom interventions are the signature of causal explanation, would find this view inconsistent but would argue that the interventions in question are ham-fisted, and as such demarcate non-causal explanations.[122]

The explanations characterised in terms of higher-order interventions have a non-causal nature because, as mentioned above, the counterfactual dependencies in question establish the stability of invariant facts under changes to lower-order invariant facts. Can we characterise said non-causal relationship more precisely? One tentative interpretation of the relationship would be as a 'fixing relationship', i.e. the values of these ICs^n are *fixed* by higher-order invariant relationships. For instance, one could think that the dynamical content of the force laws – UG-L, but also Hooke's law, or the electrostatic force, among others – is determined by e.g. Newton's 2nd law. However, it is unclear whether such a fixing function makes sense in the first place: e.g. UG-L is not derived from Newton's 2nd law. Another more plausible possibility would be a 'constraining relationship' from higher-order to lower-order invariant relationships, i.e. higher-order invariant relationships set

[122] These are terms coined by Emmerson in 'A Defence of Manipulationist Noncausal Explanation,' where he argues that there are distinctively non-causal explanations which can be characterised in terms of ideal interventions (Emmerson 2021). See also the summary of the relevant background debate in §1.3.2.2, footnote no.36.

constraints on the values of higher-order interventions. This proposal exempts us from determining what higher-order invariant relationships *do the fixing* and how such fixing looks like, and focuses instead on the (perhaps concatenated, perhaps totally overlapping) spaces of VWPP delineated by a given invariant relationship and the invariant relationships which are higher-order with respect to it. It is this second proposal the one I explore in the next section.

3.5.4 The Characteristics of Higher-order Invariance

The Extended Invariance View hypothesises the possibility of higher-order interventions, i.e. changes to the ICs^n of invariant relationships that are physically possible according to higher-order invariant relationships. Note that the notion of higher-order invariance applies to very heterogeneous relationships and is relative to explanatory context. Higher-order invariance may range from symmetry principles or conservation laws, to the laws of thermodynamics or fundamental dynamical laws, among others.

In §3.5.4.1, I claim that invariant relationships which are higher-order with respect to other invariant relationships which describe the same subject matter have a *constraining* explanatory role with respect to them.

Then, in §3.5.4.2 I spell out and address the potential twofold worry that the fact that higher-order interventions are only physically possible in a very weak sense indicates that i) higher-order invariant relationships are metaphysically necessary, and ii) higher-order interventions are metaphysically possible.

3.5.4.1 The Explanatory Status of Higher-order Invariance

Highly – or higher-order – invariant relationships are many and diverse, but one feature that unifies them is their non-causal explanatory role with respect to lower-order invariant relationships. In the following, I sketch the *constraining* explanatory function of highly invariant generalisations.

To begin with, consider special relativity's second postulate, which tells us that no massive object can be accelerated beyond the speed of light. The speed of light in the vacuum is a constant, which sets an upper limit on how fast objects can move. A similar case is the 3^{rd} law of thermodynamics, which states that the entropy of a system approaches a constant value as its temperature approaches the absolute zero, $0K$. This renders the generalisations 'G_{SR}: No material object can be accelerated from a velocity less than that of light to a velocity greater than light' (Woodward 2000, p.206; Woodward 2003, p.254) and 'G_T: No body can be cooled from a temperature above $0K$ to a temperature below $0K$' non-change-relating,

for there are no well-defined interventions on variables such as $v \geq c$ and $T \leq 0K$. There is nothing "we or nature can do that will render [them] false" (Woodward 2003, p.254).

Since there are no possible interventions that would test the invariance of G_{SR} or G_T, Woodward considers these non-change-relating – i.e. non-causal – generalisations. However, G_{SR} and G_T are highly stable, i.e. highly invariant under BaCs (see §3.3.1), and "... it is because [they are] invariant in this sense that we think of [them] as laws" (ibid.). Presumably, while they aren't causally explanatory, high invariance under BaCs renders them *non-causally* explanatory.

However, non-causal explanatoriness is not exclusive to total invariance under ICs. Consider a somewhat less invariant relationship than G_{SR} and G_T: Newton's 2^{nd} law of motion $F{=}ma$ ('2N-L' hereafter). As a dynamical law, it is a change-relating generalisation, which admits interventions on its ICs, i.e. interventions on m, a, or even F. However, with respect lower-order invariant relationships such as force laws, among which we find UG-L, 2N-L has a non-causal explanatory role. This is because it acts as a kinematical constraint on the content of physical theories: it functions as a "necessary precondition for the appropriate application of the theory" (Curiel 2016, p.1), i.e. it specifies the states a physical system can be in. N2-L may be seen as a precondition for the applicability of Newtonian mechanics: that is, to accept Newtonian mechanics is to accept that N2-L is true (ibid., p.10). It is in this respect that N2-L is a constraint: it tells us what cannot vary in Newtonian mechanics. There is a lack of consensus regarding whether kinematical constraints are physical, merely mathematical, or whether they are physical but what is distinctive about them is that they feature in mathematical explanations about the world. This is a complex matter, which would require a thesis of its own. Here I follow Curiel (2016), who does not think that laws having physical content precludes their functioning as kinematical constraints. On this note, consider G_{SR} and G_T anew. They belong to the category of negative laws, i.e., laws which prohibit states of affairs: think e.g. of Pauli's exclusion principle, which prohibits that two identical fermions (i.e. particles with half-integer spin) simultaneously occupy the same quantum state. Together with more positive principles like conservation laws, they may be seen as (physical) kinematical constraints on lower-order laws.

To wrap up the present discussion, let me add the following sidenote. Note that Noether's theorem relates symmetries to conserved quantities (though the converse need not be true). Note as well that the category 'symmetry' is very heterogeneous, so not all types of symmetres might be covered by the Extended Invariance View. For instance, Wigner (1967, pp.38-50) differentiates between

general symmetry principles, like Euclidean transformations, from *dynamical* symmetries of specific interactions, such as electromagnetic or strong force interactions. The latter clearly have physical content, but what about the former?

Among general symmetry principles, we find some which are covariant (like the crossing relations of particle physics, which *do* change events), but most of the times what we have in mind are invariant symmetry principles (like transformations or rotations, which do not change events but only their location in space and time and their state of motion). Insofar as their validity transcends any particular physical theory, Wigner calls these invariant symmetry principles "geometrical" (ibid., p.44). As such, one might argue that they fall outside of the Extended Invariance View, for recall that invariance is physical dependence.[123] That said, the conservation laws that might be derived from e.g. the Lagrangians of physical systems (which is what Noether's theorem says) seem good instances of higher-order invariant relationships. The scope of the Extended Invariance View is a complex topic that needs to be further investigated.

3.5.4.2 The Metaphysical Status of Higher-order Invariance

We have seen that the Extended Invariance View, just like the Invariance View, has been motivated by a desire to a respect scientists' intuitions with respect to nomological reasoning. As such, it claims to bypass debates about the metaphysically contingent vs necessary status of laws of nature between Humeans – for whom laws are just summaries of regularities, and hence contingent – and anti-Humeans – for whom laws are either metaphysically necessary or metaphysically necessitate their instances, a necessity conferred by either relations between universals, powers, or dispositions.[124] In the Invariance View, and by extension the Extended Invariance View, all lawhood (i.e. invariance) amounts to is physical necessity (see §3.4.3).

Nonetheless, I grant that one might still worry that the Extended Invariance

[123] That's why, in 'Some varieties of Non-Causal Explanation' Woodward argues that changes to the dimensionality of space are not physically but mathematically or conceptually possible (Woodward 2018b, p.123). To this end, one may question whether changes to the dependence of the gravitational force on the distance between the two masses is in fact a very weak physically possible or only mathematically/conceptually possible. However, I hope the case of higher-interventions on non-dimensionless constants like G is a much less controversial case of very weak physical possibility.

[124] Humean supervenience is the doctrine, put forth by David Lewis, that "all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another" (D. K. Lewis 1986c, ix). According to Lewis' Best System Account of Laws of Nature (BSA) laws are the true generalisations that best systematise our scientific knowledge, where the best systematisation is one which achieves the best combination of simplicity, informativeness, and fit (D. K. Lewis 1973, p.73)

View sneakily reclaims metaphysical possibility. The worry, I take, is twofold. First, one might think that metaphysical necessity is what justifies the fact that we keep some invariant relationships fixed in counterfactual reasoning about the physical world. Second, and in relation to the first worry, one might be concerned about the nature of VWPP. Since the changes discussed in §3.5.2 seem to be nomologically – and hence physically – impossible to some extent, for they require changes in some laws, does this make VWPP a variant of metaphysical possibility? In short, the worry is that metaphysical modality is part of the modal import of highly invariant relationships and, as a result, metaphysical possibility underlies higher-order interventions. In the following, I try to dispel any such fears.

Let me start with the worry that the invariant relationships which are kept fixed hold with a higher degree of necessity than nomological necessity: they hold with metaphysical necessity. (Metaphysical) necessitarianism about laws has, in fact, often been defended on the basis that some of these more fundamental laws of nature look much like metaphysical truths. Metaphysical necessities are rooted in the identity of things, and some of these fundamental laws are definitional. Fine (2002) argues e.g. that the law 'electrons have negative charge' is metaphysically necessary, because it defines what it is to be an electron. Wolff (2013) also thinks that some laws of nature, such as conservation laws, or symmetry principles, are like that: conservation laws define what it is to be a certain type of system (isolated, closed, etc.), while symmetry principles define physical or geometrical properties of systems. In the same way, to some extent NL-2 defines what it is to be a force in a non-relativistic classical setting, and the Schrödinger equation defines what the wave-function of a quantum-mechanical system is.

However, the Extended Invariance View, just like the original Invariance View, does not need metaphysical necessitarianism about laws to get off the ground. Science is an epistemic activity, so our access to objective modal facts will go through these epistemic means. Science gives us access to the space of physical possibility, but not to the space of metaphysical possibility (should we grant the legitimacy of such a notion). As Hirvonen et al. argue, "[e]ssences may or may not exist. We merely point out that one does not need to know that those properties, which render teacups shatterable, are essential to them. It is sufficient that teacups simply have those properties" (Hirvonen et al. 2021, p.13829). In short, metaphysical facts do not come in at the practical level: in order to know how electrons behave we need to know that they have negative charge, whether they have negative charge essentially has no practical significance. For this reason, metaphysical necessity is typically not assigned any significant theoretical role

either by most philosophers of science.[125]

The Extended Invariance View takes these practical considerations into account: it makes do with physical modality. Higher-order invariant relationships are not held fixed in virtue of their metaphysical necessity, but only in virtue of their physical necessity – this is why it bypasses debates about the metaphysical necessity vs contingency of the laws. In §3.4.3, we saw how Woodward defines physical necessity as absolute or total invariance within a law's domain of invariance: a physical dependency is necessary if there are no changes in ICs or BaCs under which it fails to be invariant. This, together with the proposed picture of fundamentality as breadth of invariance (which leads to a hierarchy of objectively invariant relationships) suggests the following relative understanding of physical necessity: an invariant relationship is physically necessary with respect to a lower-order invariant relationship iff there are no changes in the lower-order relationship under which it will fail to hold. For instance, N2-L is physically necessary with respect to UG-L because it is invariant under any changes on UG-L. This proposal results in a nested view of physical necessities, where higher-order invariant relationships may be physically necessary with respect to lower-order laws, but physically contingent with respect to even higher-order laws.[126]

The nested view of physical modalities just described also addresses the worry that VWPP is a variety of metaphysical possibility. Recall that this is a worry about the sense in which higher-order interventions are possible: given that these correspond to changes to the content of invariant relationships which are only so with respect to higher-order invariant relationships, are these physical? Even if the lower-order features described by these counternomics are not physically necessary, this need need not suggest these describe metaphysical possibilities. The key idea is that they are physical possibilities because they are consistent with at least some invariance facts of our world.

Note that necessity vs contingency per se are one thing and our judgements of necessity vs contingency are another, just like the physical possibility of interventions is one thing, and our knowledge of interventions to unravel worldly structure is another. What determines the breadth of invariance of invariant relationships under interventions and other changes in ICs and BaCs are mind-independent facts. However, these will be relativised to our current state of knowledge about

[125] Or at least philosophers of the special sciences. Necessitarians about laws will of course insist that the modality involved in VWPP may be both physical and metaphysical. Note that my view can accommodate this result; all I have argued is that science is only concerned with physical possibilities.

[126] This excludes the *highest-order* invariant relationships, whatever these may be, which are physically necessary simpliciter.

the world – an epistemic stance that requires a substantial accrual of information about worldly invariance – and dependent on explanatory context – an explanatory stance which will determine which invariant relationships are the most fundamental given our explanatory goals. For science is, after all, an epistemic endeavour.

3.6 The Possibility Question Answered: Whence Causal Realism?

The PQ is the question of what relevant notion of possibility characterises interventions. Actual interventions are possible in the sense that they are actual. However, it is unclear what possibility notion characterises hypothetical interventions, i.e. non-actual but possible interventions. Considering that interventionism is a realist theory, in what sense are interventions possible?

I believe that the right thing to say is that interventions ought to be physically possible. In this chapter, I provided a characterisation of physical possibility in terms of Woodward's Invariance View of laws. In a nutshell, the Invariance View presupposes that the physical world is a network of invariant relationships. These invariance patterns are all there is to lawhood, and the possibilities these encompass what the interventionist aims to uncover and understand. Thus, physical possibility is none other than invariance under ICs as well as BaCs. Given that interventions need not be actual, I believe WPP is the right answer to the possibility question, for recall that it encompasses actuality and SPP. Hence, it can account for both actual and hypothetical interventions.[127]

Recall that hypothetical interventions are interventions that set the relevant ICs to some non-actual values which, regardless of whether they can be realised, are *admissible* values – i.e. consistent with the relevant physical theory. While Woodward emphasises that interventions are broadly logical or conceptual possibilities in his early work, physical possibility is the most natural way to understand the possibility of interventions within the Invariance View: we could say that invariant relationships are "the patterns that nature respects" (Maudlin 2007, p.15), and physical possibilities are the possibilities these patterns allow.

But, most importantly, physical possibility is the most natural way to provide minimal metaphysical underpinnings for all of Woodward's commitments. In this respect, I argued that one way to strengthen the Invariance View and thereby block the MPQ concern would be to acknowledge the ontological primitivism of

[127] Or VWPP if one considers the *generalised* PQ arising from higher-order interventions, as I shall discuss shortly.

invariance (see §3.4.3). Nothing makes invariant generalisations invariant, they just *are*. Only such a move would render Woodward's Invariance View a truly 'small m' project that does justice to the purported causal realism of interventionism. Furthermore, this move would respect interventionism's practice-based motivation, and favour the traditionally-held intuition that it is laws that support counterfactuals, and not the other way around.

In fact, in Woodward's later work we find suggestions that support this answer to the PQ in terms of physical possibility. As an example, there is Woodward's concession that "[l]aws are what remain stable or invariant as initial conditions are different, and this is what gives us the basis for assessing counterfactual claims about what would happen if initial conditions were different from what they actually are" (Woodward 2004, p.47). Note that he does nowhere establish that interventions ought to be realised; the only condition is that they are consistent with the laws, and this is none other than physical possibility as here defined. In fact, realisation can't be a requirement if we think that hypothetical possibilities not only include what could be the case, but also what could have been the case.[128]

This leads me to the following point, which I believe is what has triggered the belief that hypothetical interventions are mere conceptual possibilities. With respect to causal models, what interventions do is cut a given variable X off from any other causal influences and set it to a new value. This is captured by *PROCESS IRRELEVANCE* which, as explained in §3.2, establishes that it doesn't matter how interventions come about, for it is embedded in the notion of an intervention – in the criteria (*IV*), so in the formalism – that any change in the value of Y will go via $I - X - Y$. In the context of interventionist explanations, *PROCESS IRRELEVANCE* is applied as follows: a variable (IC) in the generalisation which represents the invariant relationship of interest is set to a new value, according to which the system is evolved forward with the use of the generalisation in question. They key point is that hypothetical interventions, by definition, involve no actual physical process that brings the system to the new ICs. Hypothetical interventions don't occur, so they don't involve 'jumps' in physical history. Hence, the suspicion that for hypothetical interventions the relevant notion of possibility is conceptual is unfounded. We may interpret hypothetical interventions as alternative physical histories that could have led to the other-than-actual ICs in question. Only if the possibilities represented by the causal model in question correspond to alternative physical histories can we deem them objective in the sense required by

[128] So, by definition, they can't all be actualised. In chapter 2 we saw that sometimes the former are referred to as 'epistemic possibilities,' while 'counterfactual possibilities' is reserved for the latter.

interventionism.

Having outlined why WPP is the best answer to the PQ, let me now exemplify it via the Moon-tides scenario discussed in §1.3.2. The interventions on the distance between the Moon and the Earth are clearly physical possibilities because they are just changes in the ICs of UG-L. Since this intervention is consistent with all the actual invariant relationships and non-actual but possible ICs of UG-L, the relevant sense of possibility in this scenario is WPP.[129] In other words, an intervention that would double the orbit of the Moon around the Earth, is weakly physically possible, and not merely conceptually possible, because within Newtonian mechanics – a well-confirmed theory – it is possible that the Moon had been at twice the distance from the Earth. One further question to consider in relation to this case study is whether this is an intervention at all.[130]

The Moon-tides case study has given rise to a *generalised* PQ: in which sense are higher-order interventions, i.e. changes to the content of invariant relationships such as changes to G or the form of the dependency of the gravitational force on the distance, possible? In §3.5.3, I argued that provided there are relationships which are more invariant than (i.e. higher-order with respect to) the relationship whose invariance is being tested – say UG-L – higher-order interventions are physical possibilities in the sense of VWPP. This is because these are possibilities which are consistent with actual invariant relationships which are higher-order with respect to the actual relationship whose invariance is violated.

Whence causal realism, then? The reason Woodward calls the Moon-tides intervention 'conceptual' instead of 'physical' (see §1.3.2.4) is because the relative distance between the Earth and the Moon cannot be manipulated *ideally*, i.e. there is no such change of its value which satisfies all the intervention criteria, yet we know how to manipulate a Newtonian model of the two bodies. There is a trivial sense in which this possibility is conceptual: it is a possibility given by a model. But models are largely used to learn about the world, and are therefore (epistemic) representations thereof.[131] Insofar as models are constructed upon empirical information about invariance, these possibilities are not merely conceptual: they represent objective possibilities. And, insofar as these models use actual invariance patterns to generate possibilities, there is reason to think of these possibilities as

[129] Recall that whether and, if so, how criteria (*IV*) ought to be reformulated in order to count such a case as causal is a question – the applicability question – I have set aside in the current thesis.

[130] Provided it is a case of model-based causal reasoning, the causal problem discussed in §1.3.3 resurfaces. Physical possibility might provide some objective basis for interventions, but still can't guarantee that the case at hand satisfies (IV). One might think that this case study is an example of model-based causal reasoning; see Jansson (2014) for an interpretation of this sort.

[131] Models are epistemic representations because they afford information about their targets (see Nguyen & Frigg 2022, p.6).

ways the world could be or could have been. In short, a spectrum of physical possibilities of the type I explored in this chapter can replace conceptual possibility and thereby provide a much more robust metaphysical basis, from the scientific realist perspective, for interventionist explanations in full generality.

3.7 Conclusion

In this chapter, I closely analysed Woodward's Invariance View of laws, which characterises lawhood as invariance under changes in ICs, including interventions, with several aims.

First, I argued that supplementing the Invariance View with a primitivism about invariance satisfactorily answers the MPQ, a metaphysical circularity concern regarding laws and counterfactuals: supplementing the Invariance View with a primitivism about invariance prioritises invariance over counterfactuals. This, I believe, is the best metaphysical completion of the Invariance View, which not only renders it a 'small m' project but, importantly, stays true to its methodological motivation to focus on what we can do with laws instead of providing an account of what they are. Laws – or rather, invariance patterns – just *are*, and they are what supports counterfactuals and enables scientific explanations.

Second, I also addressed a circularity concern in the vicinity: the epistemic worry that if it is laws that tell us which interventions are possible and in order to assess whether a generalisation's degree of invariance, one needs to know which changes are possible, this reasoning is circular. In hypothetical interventions, we rely on invariant relationships and other theoretical considerations that we had already identified on straightforward empirical grounds in order to assess whether the interventions in question are possible. This move does not already presuppose the invariance of the relationship being tested; hence, vicious epistemic circularity can be avoided.

Third, I argued that given invariance relations are empirically supported (and thereby objective), the key notion of possibility for interventions ought to be a notion of physical possibility. Since interventions are actual or hypothetical manipulations on worldly invariance, I argued that WPP provides a satisfactory answer to the PQ posed in chapter 1. While it is trivial in which sense actual manipulations of the world are physically possible, in the hypothetical cases it is only via models that we can gain access to the space of physical possibility. While models encode epistemic possibilities, the primitivism of invariance ensures that they are representations of objective possibilities; what guarantees their latching onto the

world is the empirical information about invariance input in their construction.

Finally, inspired by Woodward's conviction that an account of laws should do justice to actual scientific methodology – which motivates the independence conditions $INDEPENDENCE_{L-ICs}$, $INDEPENDENCE_{IC-IC}$, and $INDEPENDENCE_{L-L}$ – in the same spirit I developed the Extended Invariance View. By giving an account of the fundamentality of worldly dependencies, the Extended Invariance View aims to do justice to scientific reasoning: scientists often confidently assess counterfactuals which describe what would happen had the laws been different. Provided the worldly infrastructure is such that these counterfactuals admit an understanding in terms of laws vs ICs, we can make sense of changes to the laws by understanding them as interventions on their ICs^n consistent with higher-order laws. This explains why we cannot make sense of changes to totally invariant relationships: not only do they rarely admit a description in terms of ICs but there are no physical facts with respect to which they may vary.

Higher-order interventions posed a generalised PQ, which I proposed to answer via VWPP: higher-order interventions are physical possibilities because they respect the most fundamental substrate of the actual world, though only very weakly so because they are not consistent with its lower-order features. I also argued that the kind of dependence between higher-order and lower-order invariant facts is non-causal because of the constraining role of higher-order invariant generalisations exhibit towards lower-order invariant generalisations, yet still physical because it relates invariance facts. The plausibility of non-causal versions of interventionism is a topic that will be further explored in the upcoming chapter 4.

Chapter 4

Possibility via Invariance II: The Object of Interventionist Counterfactuals

4.1 Introduction

In Woodward's and Hitchcock's (*W-H* hereafter) interventionist account of causal explanation (Hitchcock & Woodward 2003; Woodward 2003; Woodward & Hitchcock 2003), causal generalisations support *same-object* counterfactuals (SO-CFs), counterfactuals that provide modal information about *"the very object o"* (Woodward & Hitchcock 2003, p.20). The contrast class for SO-CFs are *other-object* counterfactuals (OO-CFs), which are associated with generalisations that apply to different objects. Hence, according to the interventionist, generalisations are causally explanatory when they support SO-CFs. Provided *W-H*'s account of causal explanation is all there is to scientific explanation, this renders SO-CFs explanatory and OO-CFs wholly non-explanatory. By focusing on the roles counterfactuals play in scientific explanation, in this chapter I identify two challenges to this dichotomous taxonomy of counterfactuals. These challenges highlight the limitations of *W-H*'s account of causal explanation as an exhaustive account of scientific explanation.

The first challenge is the *individuation challenge* (§4.3), which is the challenge of determining the objects of interventionist counterfactuals. I argue that in scientific explanations SO-CFs typically provide information about *classes of systems*. This helps us understand what particulars are the objects of hypothetical interventions: non-actual members of the classes of objects specified by the causal models in question. Moreover, I argue that the fact that SO-CFs typically provide information about classes of systems challenges the view that explanatoriness is just a matter of *depth*.

The second challenge, the *idealisation challenge* (§4.4), concerns counterfactuals which appeal to idealised models and are therefore invariance-violating, i.e. counternomics, as a result. The challenge is that, qua form, these counterfactuals are OO-CFs but, if so classified, *W-H*'s account of scientific explanation can't account for their somewhat explanatory character. This is why I propose

to understand them as '*same-object-enabling* counterfactuals' (SO-enabling-CFs, for short), OO-CFs with a distinct explanation-enabling function: the function to delimit the scope of causal generalisations. In particular, the counternomics which give rise to the idealising challenge have the function to enable causal explanations with idealised models. Next, I propose an interventionist strategy to identify them from the broader class of OO-CFs. I then discuss the limitations of this strategy and the resulting implications for the view that scientific explanation reduces to causal explanation.

Last, I discuss why abstraction does not raise an analogous challenge to the idealisation challenge: unlike idealisations, which are misrepresentations, abstractions are detail-poor representations. Given that abstractions are detail-poor representations, there is no need to justify the scope of abstract models, so no abstract model-based explanation-enabling counterfactuals in play. Next, I show why abstraction is not, in *W-H*'s account of causal explanation, constitutive of non-causal explanation. Finally, I look at the renormalization group (RG) explanation of universal behaviour at criticality; for many, an archetypal example of non-causal explanation. I argue that we should not confuse inter-system RG explanation (Sullivan 2019) with what I call 'alternative RG explanation:' it is alternative RG explanation, not inter-system RG explanation, that is non-causal by the interventionist's lights.

4.2 Interventionist Counterfactuals & Scientific Explanation

Counterfactuals play a central role in the metaphysics and the philosophy of science. They have traditionally been thought to be supported by laws of nature, understood as exceptionless generalisations. As seen in chapter 3, interventionism breaks this connection of explanatoriness with exceptionlessness via an alternative view of what makes causal generalisations explanatory: their invariance under changes in BaCs, BCs, and ICs, including interventions. This view of explanatoriness as invariance motivates a redefinition of the notion of generality involved in explanations from universality, or 'scope,' to 'depth,' where depth characterises explanations in terms of the degree of invariance of the generalisations featuring in these explanations (Woodward & Hitchcock 2003). The key idea is that while *scope* captures a generality with respect to objects or systems other than the one the explanation targets, *depth* captures a generality with respect to other properties of the very object or system the explanation is about. Thus, the type

of counterfactuals interventionists deem explanatory are not those that involve many objects or systems, but those that concern only the object or system that is the focus of an explanation. This is why interventionist counterfactuals are also known as 'same-object' counterfactuals (SO-CFs). In this first section, I motivate and introduce these.

4.2.1 The Standard View of Scientific Explanation: Other-Object Counterfactuals

Laws have traditionally been understood as universal generalisations of the form *All As are Bs.* In the philosophy of science this standard view of laws is connected with the deductive-nomological (DN) model of scientific explanation (Hempel 1965; Hempel & Oppenheim 1948), which asserts that explanations are deductive arguments.[132] The premises of such deductive arguments are theoretical statements, statements of initial conditions, and at least one law of nature, where laws are characterised as exceptionless generalisations describing regularities, and are expressed as universal generalisations. This nomological component is central to the DN model: explanation is *essentially* a matter of deduction from laws. Therefore, only counterfactuals which are *supported* by – i.e. deduced from – laws are explanatory; those which follow from accidental generalisations, law-like assertions that hold by coincident or accident, aren't. As such, DN-model supporters are committed to a dichotomous view of scientific explanation, where explanation is exceptionlessness or nothing. As a matter of example, consider the following universal generalisation and counterfactual pair (Chisholm 1955, p.97):

> **(1)** All gold is malleable.
>
> **(1-C)** If that metal were gold, it would be malleable.

Counterfactual **1-C** naturally follows from generalisation **1**, because **1** is a law. Contrast **1** with the following famous example (Van Fraassen 1989, p.27):

> **(2)** All solid spheres of gold have a diameter of less than one mile.
>
> **(2-C)** If this billiard ball had been made out of gold instead of polyester, it would have had a diameter of less than one mile.

While counterfactual **2-C** seems to follow from **2**, **2-C** is not explanatory because while **2** may be true, it is an accidental generalisation, not a law. There is nothing

[132] Strictly speaking, the standard view of scientific explanation is identified with the covering-law model, which comprises deductive-nomological explanation on the one hand, and probabilistic-statistical or inductive-statistical on the other.

in the laws of nature that would prevent solid spheres of gold having a diameter of less or more than one mile, so **2** is accidental because it is not *necessarily* the case. Therefore, an inference from **2** to **2-C** is not a deduction, which a scientific explanation should be according to the DN model. Only law-like generalisations like **1** are allowed to feature as premises in scientific explanations, arguments with the logical form (Woodward & Hitchcock 2003, p.18):

> **(3)** All *A*s are *B*s.
>
> Object *o* is an *A*.
>
> Therefore, *o* is a B.

Law-like universal generalisations trivially support counterfactuals that describe the behaviour of objects other than *o*, call these *o**, were they to be *A*s; this is the reason why Woodward calls these counterfactuals '*other-object* counterfactuals' (OO-CFs). They take the following form (Woodward 2003, p.281):

> **OO-CF**: If some object *o**, different from *o* and that does not possess property *A*, were to be an *A*, then it would be a *B*.

As we shall see in the following section, the interventionist argues against the explanatoriness of OO-CFs and proposes an alternative account of explanation to the DN model based on another type of counterfactuals, which act as the contrast class for OO-CFs. These are *same-object* counterfactuals (SO-CFs), and the interventionist maintains that these are the *only* explanatory counterfactuals.

4.2.2 The Invariance View of Causal Explanation: Same-Object Counterfactuals

Woodward judges the widely accepted DN model inadequate to describe causal explanation on the grounds that i) not all true causal claims are underlain by laws, if laws are exceptionless generalisations, and ii) in the event that they were, the DN model would not offer necessary and sufficient conditions for lawhood, and so it is unclear how it could help demarcate laws from accidental generalisations, and consequently explanations from non-explanations.[133] While he acknowledges that laws play an important role in explanations (Woodward 2003, p.185), he disagrees with the idea that all explanations work by showing that their explananda are nomologically expectable. For this reason he moves away from this traditional

[133] By causal generalisations which are not laws I mean exception-ridden generalisations like 'Smoking causes lung cancer' or 'Penicillin cures tuberculosis.'

view of scientific explanation and proposes an alternative account of explanation in terms of *invariance* under ICs, including testing interventions (this is the Invariance View introduced in chapter 3).

As we saw in §3.3.2, a generalisation G is *change-relating* when it maps the values of some dependent value Y to the values of some independent variable X so that $G{:}Y{=}f(X)$. G is causal if it is change-relating, i.e. invariant, under interventions. Recall that interventions are actual or hypothetical changes in the value of an intervention variable I which cause an exogenous variable X to change its value from, say, x_1 to x_2, such that $f(x_1) = y_1 \neq f(x_2) = y_2$. A change-relating generalisation's degree of invariance is proportional to the amount of testing interventions it withstands, and 'laws' are the name by which we refer to the highly invariant change-relating generalisations.

In order to better understand in which sense the Invariance View diverges from the standard account of laws, and by extension, which kinds of counterfactuals turn out explanatory, let us look at the following example:

Coulomb's Law:

$$\vec{F} = k\frac{q_1 q_2}{|\vec{r}|^3}\vec{r} \qquad \text{where} \quad k = \frac{1}{4\pi\epsilon_0} \tag{4.1}$$

Equation (4.1) states that the magnitude of the electrostatic force of attraction or repulsion between two point charges is directly proportional to the product of the magnitudes of the charges and inversely proportional to the square of the distance between them, where k = Coulomb's constant (Coulomb 1785). Note that eq. (4.1) is about the electric force $\vec{F}$. The electric field $\vec{E}$ can be obtained by dividing the electric force per unit charge:

$$\vec{E} = \frac{\vec{F}}{q} \tag{4.2}$$

The charge q in the denominator corresponds to the charge placed in the electric field, and so the subscript assigned to it will depend on one's labelling of the charge which creates the field, and the one placed in the field. Suppose that $q = q_2$, and q_1 the conductor generating the electric field. Upon plugging eq. (4.1) in eq. (4.2), the latter equation becomes:

$$\vec{E} = k\frac{q_1}{|\vec{r}|^3}\vec{r} \tag{4.3}$$

Let's call eq. (4.3) 'Coulomb's Law*.' When analysing the relation between the expression's variables, we shall only consider their magnitude, not their direction.

What makes eq. (4.3) explanatory? To investigate this question, consider the following counterfactual (adapted from Woodward 2003, p.282):

> **CONDUCTOR**: If Bill Clinton, the government of Italy, or a neutron were a conductor with such and such a geometrical configuration and charge distribution, he/she/it would produce an electric field of strength E.

There are no possible interventions on, say, Bill Clinton that would change the value of the strength of E because E does not depend on anything like Bill Clinton's *identity*.[134] Certainly, eq. (4.3) does *not* support CONDUCTOR. This is no surprise: CONDUCTOR is an OO-CF, for it informs us that any non-conducting object (Bill Clinton, the government of Italy, or a neutron) would behave like a conductor if they were one. This is a trivial truth, and the reason why CONDUCTOR is non-explanatory.

Having seen that causal generalisations do not draw their explanatory power from their supporting OO-CFs, what sorts of counterfactuals do they support? Consider eq. (4.3) anew. Given that it expresses a dependence of the electric field a conductor produces on its geometrical configuration and charge distribution, it supports counterfactuals such as:

> **CONDUCTOR'**: If a conductor's charge were to increase an amount δq, the electrical field produced would increase by an amount δE.

> **CONDUCTOR"**: If a finite straight uniformly charged conductor were coiled into a solenoid or twisted into a circle, the value E of the electrical field produced would change.

While the change described by CONDUCTOR goes beyond the scope of Coulomb's Law*, counterfactuals like CONDUCTOR' or CONDUCTOR" do fall within the law's domain of application: geometrical configuration and charge distribution do affect the value of E. In short, what makes a generalisation explanatory, according to the interventionist, is the extent to which it conveys information about an actual system's 'internal' causal dependencies.[135]

[134] This sort of intervention would seem to require metaphysically impossible changes to identity facts.
Note that the manipulative element is contained in the counterfactuals, not the laws. In our example, Coulomb's Law* does not tell us how to change the amount of charge so that the electrical field changes, but only that the amount of charge changes the electrical field's strength. Interventions need to be specified in the counterfactual: e.g. whether the amount of charge along the wire increases, the field is measured at a different distance from the wire, or the initial conductor (a finite straight wire) is twisted into a circle or coiled up into a solenoid.

[135] I use 'system' interchangeably with 'object.'

One way to understand what 'internal' means is via Hüttemann's account of laws (Hüttemann 2021), inspired by Scheibe (2001). Hüttemann's view is that law statements involve two kinds of generalisations:

- *system-external* generalisations: they define the domain of applicability of an equation over systems of a certain kind.
- *system-internal* generalisations: they define a system's 'state space,' namely the space of all possible states in which a system can be.

Law statements of the external type help us identify all the systems of kind A that behave according to B. In other words, they identify the law's scope or domain of applicability. For instance, a system-external generalisation may be: 'All ideal gases behave according to $PV{=}nRT$'.

Law statements of the internal type tell us exactly what being a B depends on, e.g. $p{=}p(V,T)$, $V{=}V(p,T)$ and $p{=}p(V,T)$ for the ideal gas law.

Why is this distinction important? OO-CFs presuppose system-external generalisations. When we say 'If b were/weren't an ideal gas, then... ' we take b *outside* of its state space. Given that the interventionist takes the possibility space for interventions to be a system's state space, it is the system-internal generalisations and the counterfactuals about them which are explanatory. It is for this reason that interventionist counterfactuals are called '*same-object* counterfactuals' (SO-CFs). They take the following form (Woodward 2003, p.281):

> **SO-CF**: If the value assigned by the variable X to o were to be changed via an intervention (e.g. from $X(o){=}0$ to $X(o){=}1$), then the value assigned by Y to o would change in some way predicted by the generalization.

The definition **SO-CF** makes explicit the fact that causally explanatory counterfactuals must, according to the Invariance View, describe what values the explananda variables take once interventions change the values of the explanantia variables (Woodward & Hitchcock 2003, p.17). In the present example, CONDUCTOR' and CONDUCTOR" are examples of SO-CFs because they describe how changes in the explanantia variables, δq or $\frac{\delta r}{r^2}$, affect the explanandum variable δE according to Coulomb's Law*.

4.2.3 Why Same-Object Counterfactuals?

In the following, I discuss the motivations behind *W-H*'s dismissal of OO-CFs' explanatoriness and the resulting identification of interventionist counterfactuals with SO-CFs.

The first motivation concerns well-known problematic features of traditional accounts of scientific laws and explanation. While attracted by some other features of the DN model, such as the fact that it is an objective and unifying account which highlights the role of laws and derivations from laws in scientific explanation, *W-H* oppose the idea that explanation reduces to nomological 'expectability;' e.g. if the generalisation 'All ravens are black' establishes a nomological relation between ravenness and blackness, the fact that *this raven is black* is explained by the fact that *it is a raven*. Just like ravenness implies blackness, generalisations such as 'All inertial bodies have no acceleration' apply universally without exceptions. However, a quick look at the special sciences should suffice to convince us that this is barely ever the case. Generalisations may hold in specific domains and/or have exceptions and yet be explanatory. Even within physics universality is usually off the table: Newton's laws correctly describe bodies moving at low velocities, and apply in inertial frames of reference. Similarly, Maxwell's equations break down when quantum effects become significant, and so does general relativity below the Planck length (Woodward 2013, p.62). Explanatoriness cannot be a matter of nomological expectability because it requires that laws be exceptionless generalisations, which are rare – if not mythological – creatures in science.

How is this criticism of the DN account reformulated as a criticism of OO-CFs? The argument goes as follows: if explanatory generalisations are to support OO-CFs, they must specify conditions that are sufficient to ensure that any object will behave according to the generalisation in question, ergo they must be exceptionless. But, as just seen, explanatoriness does not reduce to exceptionlessness, so we don't have sufficient conditions for lawhood. Therefore, we must relinquish the idea that laws are exceptionless generalisations. The Invariance View *loosens* the explanatory demands on laws by offering a necessary condition (*N*) for lawhood in terms of invariance under interventions (see §3.3.2).[136] This framework accommodates the fact that generalisations may have exceptions and yet be explanatory. For instance, Coulomb's Law does not apply to quantum electrodynamical systems, but it correctly describes the electrodynamical behaviour of classical systems.

Thus far, we have seen that Woodward & Hitchcock (2003) reject the idea that it is in virtue of their exceptionlessness that generalisations are explanatory, and propose instead that their explanatoriness comes from their invariance under testing interventions. Invariance under interventions is not universal in character: to count as explanatory a generalisation must be invariant under some – though not necessarily all – interventions. Nonetheless, this does not imply a dismissal of

[136]Which, as I argued there, could be reformulated as a (*N&S*) for invariance.

exceptionless generalisations, for very high invariance corresponds with exceptionlessness. The interventionist's view regarding exceptionlessness is that causal generalisations are neither explanatory in virtue of their exceptionlessness, nor that their exceptionlessness is a sufficient condition for their non-explanatoriness.

The second more positive motivation SO-CFs is their epistemic privilege over OO-CFs. Recall that interventionism is a manipulability theory, where interventions are handles on causal relations. This means that, even though there are objective causally explanatory facts, only those accessible to us are of our interest. In other words, causal relations are out there in the world, and so enjoy a mind-independent existence, yet we are interested in them insofar as we can *use* them in our benefit for purposes of manipulation and control, hence the epistemic access requirement. Unlike OO-CFs, Woodward argues that "same-object counterfactuals involving interventions are typically clear enough in meaning, and we often have or can obtain scientific evidence that is relevant to their truth" (Woodward 2003, p.282). They have a clear meaning because their antecedents specify interventions; they are what Woodward calls 'active' counterfactuals (Woodward 2000, p.199).[137] Active counterfactual dependence is a necessary condition for empirical testing, which is what provides the evidence-conditions for SO-CFs' truth: often we can evaluate the content of these counterfactuals through actual interventions. Having said that, recall that within interventionism possibility is not reduced to actuality, and manipulability is not reduced to practical possibility, so in great part this epistemic privilege is contingent on possible actual interventions.

To conclude, SO-CFs are motivated on the following grounds. First, they are active counterfactuals which specify interventions and thereby fit the view that explanation is not subsumption under laws but active counterfactual dependence. Second, their truth can sometimes be experimentally tested.

4.3 The Individuation Challenge

In §4.2, we saw that *W-H*'s account of causal explanation differs from the DN model in the kind of generality it attaches to explanations: the DN model of explanation emphasises *scope*, i.e. generality with respect to objects other than the one the explanation targets, *W-H*'s account emphasises *depth*, defined as "generality with respect to *other possible properties of the very object or system that is the focus of explanation*" (Woodward & Hitchcock 2003, pp.2-3). This distinction between types

[137] Though recall that later on Woodward relaxes the view that explanation is *active* counterfactual dependence when he extends interventionism beyond causal explanation (see chapter 3). This topic will be discussed again in §4.5.3.

of generality translates to a distinctive standard for the explanatory potential of generalisations, which the interventionist believes is determined by their depth, not their scope, and therefore, by their ability to support SO-CFs – like CONDUCTOR' and CONDUCTOR" (see §4.2.2), – which are SO-CFs rather than OO-CFs because they describe properties of "*this very conductor*" (ibid., p.20).

The main aim of this section is to examine what the *very object* of SO-CFs is. This is what I dub the 'individuation challenge,' which I introduce in some more detail in §4.3.1. I illustrate it via a case study in §4.3.2. Finally, in §4.3.3 I offer a solution to the challenge.

4.3.1 What is the Individuation Challenge? Tokens vs Types

Interventionists use models to represent causal dependencies out there in the world. The patch of the world under study is the model's 'target system,' whose properties are represented by variables. Since values of variables represent determinations of properties, and properties are always possessed or instantiated in particular individuals, there is a clear sense in which the *very object* always refers to a particular individual *o*. At the same time, since the relata of causal relations are variables' changing values, often the causal models wherein they feature represent classes of structurally identical physical systems. What sorts of objects are then the focus of interventionist explanations: particular systems, types of systems, or both? I dub the 'individuation challenge' the challenge of characterising what is meant by *the very object* or *the same object o* with respect to which interventions are enacted to test the invariance of the causal generalisations which characterise their behaviour.

The tension between particular systems and types of systems highlighted by the individuation challenge is at the core of Woodward's interventionist framework. On the one hand, Woodward is interested in type-level causal relations: (direct, contributory, etc.) causal notions, which criteria (*IV*) describe, are type-level notions. On the other hand, the truth of type-level causal claims depends on relationships that do or would obtain at the level of *particular, spatiotemporally located individuals* (Woodward 2003, p.40). The definition of an intervention, (*IN*), is, after all, a token-level causal notion (see definitions in §1.2.2).

Note that for Woodward token-level and type-level causal claims are two ways to describe causal relations, which obtain among tokens. Even if tokens are not causes of types, and types are not causes of tokens, there is a sense in which type causal relations like 'The alarm going off at 8 am causes me to get out of bed every morning' or *S* :'Smoking causes cancer' depend on relationships that do

or would obtain at the level of particular individuals: the values of the variables these type-level claims connect – namely, {alarm on 8 am, alarm off 8 am} causing {goes out of bed, stays in bed}, and {smokes, does not smoke} causing {develops lung cancer, does not develop lung cancer} – are always realised in particular individuals. In this sense, the explanatoriness of type-level causal relations *flows* from the explanatoriness of token-level causal claims. At the same time, all token-level causal relations commit us to the truth of some type-level causal relation, even if we are unaware of the relevant type-level generalisation (Woodward 2003, p.72). Therefore, in this other sense the explanatoriness of token-level causal relations presupposes the explanatoriness of some type-level causal relation.

The lack of clarity regarding the direction of the truth-making relation between type-level and token-level causal claims – which backs the resulting explanatoriness – is the source of the individuation challenge: what are most fundamental, causal facts about concrete systems or classes of them?[138] Fortunately, I believe one need not give a full-fledged metaphysical account of the truth-making relation between tokens and types in order to highlight some of the problems that pervade *H-W*'s account of causal explanation.

4.3.2 A Case Study: Sea Water's Electrical Conductivity

As an intuitive introduction to the individuation challenge, consider the following scenario. Imagine you are walking on the beach shore with your family. Your son jumps into the water to swim, and soon after comes back with a wry face: his eyes hurt, and he exclaims 'the water is *so* salty!' You smile and tell him the differences between sea water and fresh water. This event triggers a more intellectually-stimulating conversation with your teen daughter, who has recently taken up chemistry at school. You, a proud chemistry teacher, explain to her that the sea water's salinity is mainly due to two ions, chloride and sodium, and that this gives it very interesting properties, such as its electrical conductivity. You tell

[138] Given that causation obtains at the level of particular individuals, one would expect tokens to be truth-makers for type-level causal claims; S is true if there is at least one actual intervention on someone's smoking habit that has the effect that said individual develops lung cancer as a result. However, that still doesn't guarantee that this causal relation will obtain in other instances. Consider a probabilistic scenario where S constitutes a truth-maker for more than 70% of instances of lung cancer and the respective token-level causal claims (Maria's smoking caused her lung cancer, Ian's smoking caused his lung cancer, etc.). S, e.g., does not ground the claim that "Jones' smoking caused his lung cancer" because despite being a smoker, his lung cancer was caused by asbestos exposure (Woodward 2003, p.75). At the same time, given the reproducibility assumption implicit in S, one could argue that the characterisation of actual causal relations depends on type-level causation (ibid., pp.70-71). See Hausman (1998, 2005) for a discussion of this point: he defends a truth-making relation from tokens to types, where types are generalisations over actual and possible tokens (Hausman 2005, p.20).

her:

> **DISTIL**: If sea water were distilled, it would stop conducting electricity.

DISTIL describes the behaviour of sea water more generally, and you know is made true by the following type-level causal generalisation, which links number of ions to electric conductivity:

> **H_2O – CONDUCTIVITY**: Changes in the quantity of ions in water cause changes in the electric conductivity of the water.

What type of counterfactual is DISTIL? It is a SO-CF, since its content can be re-described as an interventionist counterfactual in the following way:

> **DISTIL'**: If *a* sample of sea water with volume V_{SW} were distilled, it would stop conducting electricity.

The antecedent of DISTIL' describes an intervention whereby *a* sample of sea water with volume V_{SW} is deionised through distillation; the mineral ions (anions and cations) present in the water are removed with the help of ion-exchange resins. The object of said counterfactual is thus *any* member of the class '(sample of) sea water.'

Envisage now a situation where you are at high school, giving a lecture on this very same topic. The summer vacation is over, so you are not on the beach (alas!) but in the lab. Explaining H_2O-CONDUCTIVITY to your students, you tell them, while pointing at the container in front of you:

> **DISTIL''**: If *this* sample of sea water with volume V_{SW} were distilled, it would stop conducting electricity.

DISTIL'' is also clearly a SO-CF: it captures the behaviour of a concrete system, a certain sample of sea water with volume V_{SW}. In this case, the counterfactual is about *this* very sample of water, and not just *any* sample of water. DISTILL'' can be shown to be the case via an actual experiment, if when one intervenes on the concrete system in the way described by its antecedent the consequent results.

Thus, DISTILL'' is a counterfactual that belongs to the class specified in DISTIL', supported by the generalisation H_2O-CONDUCTIVITY (which, as a type-level causal generalisation, also underwrites many other interventionist counterfactuals). Consider now an intervention that instead of distilling the original sample of sea water V_{SW} replaces it with a sample of distilled water of volume V_{DW}. Assume that nothing except for the number of ions distinguishes the

values of V_{SW} and V_{DW}, so $V_{SW} \approx V_{DW}$. Such an interventions can be described as follows:

> **REPLACE'**: If *a* sample of sea water V_{SW} in a container were replaced by a sample of distilled water V_{DW}, there would no longer be an electrical conductor in the container.[139]

Call the original system with volume V_{SW} '*o*', and the one it is replaced by, with volume V_{DW}, '*o**'. Unlike DISTIL', REPLACE' describes an intervention on a system *o* which transforms it into another *o**, at least if one views it through the lens of identity as material constitution. This intervention alters the system's explanans property, N_i = no. ions, by replacing the original system with one identical except for the value of N_i. Under a token identity view, REPLACE (or any more localised version thereof along the lines of DISTILL") is an OO-CFs. However, this doesn't seem right, for REPLACE' is a perfectly well-defined SO-CF. Here is why.

First, it satisfies the criteria for intervention variables (*IV*). Considering an intervention variable I on X with respect to Y, there is a direct causal path from I to X, I switches off other interferences with X, X has no other influences on Y other than through X, and I is statistically independent of any other variables that also cause Y. By replacing V_{SW} with V_{DW}, where these two samples of water only differ in N_i, we ensure that changes in water's E are only due to a change in the value of N_i. The intervention is, therefore, surgical.

Second, its meaning is perfectly clear. As concluded in §3.6, interventions ought to be physically possible, in the sense of being describable by empirically-supported scientific theories, regardless of how they could come about. In this respect, the material discontinuity of the intervention captured by REPLACE poses no difficulties, because the end-state of the intervention corresponds to one of the possible values N_i = could take, and is in this sense identical to the post-intervention state described by DISTIL'. One could thus conclude that REPLACE' is also a SO-CF. Given this result, how should we meet the individuation challenge?

4.3.3 The Solution to the Individuation Challenge

At the beginning of this section, I defined the individuation challenge as the challenge of understanding what is meant by the *very* or *same object o* with respect to which interventions are enacted in order to test the explanatoriness of the internal generalisations which characterise their behaviour.

[139] Distilling the water minimally changes the fluid's volume. The volume of the ions is so small compared to V_{SW} that it is negligible. Therefore, $V_{DW} \approx V_{SW}$ to a good degree of approximation.

In §4.3.3.1, I argue that all interventionism proposes is a representational individuation criterion which, when applied in the context of actual scientific practice, leads to the result that SO-CFs typically individuate *classes of systems*. That is, while interventions are enacted on particular systems, the interventionist counterfactuals which make up most of the scientific practice are about *classes of systems*, not about particular systems.

In §4.3.3.2, I contend that, in light of the aforementioned solution to the individuation challenge, *W-H*'s definition of hypothetical interventions in terms of "*hypothetical* values of *X* and *Y possessed by the very object o*, namely those values of *X* and *Y* possessed by *o* in the hypothetical situations where the value of *X* is changed by an intervention" (Woodward & Hitchcock 2003a, p.20) is not the best suited. Therefore, I suggest an alternative characterisation of hypothetical interventions.

Last, in §4.3.3.3 I analyse what the thesis that most SO-CFs are about classes of systems discloses about the interventionist's judgements concerning the explanatoriness of causal generalisations.

4.3.3.1 Science Considered: Models of Kinds

Recall that *W-H*'s interventionist theory of causal explanation aims to elucidate features of "actually existing science" (Woodward 2020, p.1908), such as how scientists reason about causation. I believe this motivation to track scientific methodology and reasoning invites the view that SO-CFs are typically underwritten by type-level causal generalisations which apply to *classes of systems*.

Undeniably, the object of scientific investigations is phenomena more generally and, only derivatively, individual events: scientists are interested in whether and to what extent smoking causes lung cancer, the curing of diseases via the intake of newly developed drugs, and the impact of meat and dairy consumption on the release of greenhouse effects. In short, scientists are interested in type-level causal claims. The focus on types is clear from the formal aspect of interventionism: interventionist counterfactuals are expressed in terms of *variables*, which ought to satisfy the formal requirements (*IV*) and are assigned values via the scientific theories describing the relevant processes. DISTIL' and REPLACE' are structurally indistinguishable: they are individuated by the same variable sets and relations between these. The focus on structure becomes even more manifest when one considers the *PROCESS IRRELEVANCE* of interventions (see §3.2). Thus, it is clear that all that interventionism offers is a *representational* individuation criterion for SO-CFs: the system pre- and post- intervention is the same provided there is

identity up to the variable level for the explanantia and explananda variables.[140] The variables that do not feature in the explanans (ICs which are not part of the explanation and BaCs) are held fixed at their actual values.[141] According to this individuation criterion, REPLACE' individuates the same system pre- and post-intervention.

One might question this individuation criterion by objecting that counterfactuals like DISTIL" individuate concrete systems. Indeed, the explananda and explanantia variables are more or less localised depending on the explanandum. One's causal interest may be focused on a particular system that exists at some place over some time – like the concrete sample of sea water in DISTILL" – or on all systems of a specific kind – like the generic 'sample of sea water' DISTIL' or REPLACE' refer to. This factoring in of agents' interests in causal explanations highlights the epistemic side of explanations: it is agents who decide what needs explaining in any given context. However, the explanation itself is ontic, for it cites features of the world. What makes true both DISTIL' and DISTIL", counterfactuals underwriting, respectively, a type-level and a token-level causal claim, is the same causal generalisation H_2O-CONDUCTIVITY. In other words, this generalisation makes true a class of counterfactuals, to which DISTIL', DISTIL", and REPLACE' all belong.

While one can enrich causal models as much as required to pick out the features of particular systems, "theoretical research typically involves the exploration of classes of models aimed at understanding classes of phenomena, not the study of individual observations or individual phenomena" (Weisberg 2013, p.114). Purely on descriptive grounds, most models are not detailed descriptions of specific target systems, but idealised and/or abstract descriptions which apply to classes of concrete systems. Hence, SO-CFs will typically individuate classes of systems. This resolution of the individuation challenge is in line with how other philosophers of science think about these counterfactuals. For instance, in reference to counterfactuals which appeal to what they view as abstract models, Jansson and Saatsi claim that SO-CFs should not be viewed as having "a particular, individual object as the explanatory target" (Jansson & Saatsi 2019, p.835), but rather as describing *generic* explananda. Generic explananda normally correspond to classes of targets, but may also be the result of an intersection of properties common to a series of objects or classes of objects (see Weisberg 2013, p.117). When the simple gravity pendulum law is used to explain the period of a simple gravity

[140] This observation already came up in chapter 2, where I contended that interventionism is neutral concerning the metaphysics of properties which causal models represent.

[141] This is the 'relativity to background conditions' proposed by Statham (2018) I discussed in §1.2.3.1.

pendulum, the SO-CFs we are typically interested in are those that describe, e.g. how the length of '*a* simple pendulum' (generic explanandum) and not '*this* pendulum' (concrete explanandum) affects its period. For explanations of this kind, the relevant counterfactuals are SO-CFs referring to generic explananda, or rather unspecified members from the class *simple pendula*. Similarly, in the context of explanations of what affects the conductivity of water, the counterfactuals which typically feature in scientific explanations describe e.g. how the quantity of ions of '*a* sample of water' and not '*this* sample of water' affects its electrical conductivity. To this end, DISTIL' and REPLACE' refer to any member of the class *sea water samples with volume* V_{SW}.[142]

4.3.3.2 Science Considered: Hypothetical Interventions

In the previous section, I argued that SO-CFs typically individuate classes of systems such as *simple pendula* or *samples of sea water with volume* V_{SW}. However, this still does not satisfactorily answer the individuation challenge. In this section, I contend that SO-CFs which individuate classes of systems describe interventions on non-specific members of a class of objects: *a* simple pendulum, *a* sample of sea water V_{SW}, etc. More specifically, I question the interventionist's view that hypothetical interventions are instantiations of hypothetical properties in actual objects, given that it is not sufficiently all-encompassing of all interventions. In particular, it can't handle hypothetical interventions on physically impossible objects such as ideal pendula or continuous water. Therefore, I suggest characterising hypothetical interventions as changes to the properties of concrete hypothetical objects (e.g. ideal pendula), which in time typically represent classes of actual objects (e.g. real pendula).[143]

Consider the sea water's electrical conductivity case study (see §4.3.2) anew. The best way to test the invariance of the generalisation H_2O-CONDUCTIVITY is by *actually* manipulating a concrete sample of water. I could measure the current in an actual tank of sea water V_{SW} with a multimeter, and then either distil it or replace it with an equal volume of distilled water V_{DW}, and measure the

[142] A caveat concerning the proposed individuation criterion is that it is merely a representational criterion and, as such, only ensures that counterfactuals are well-defined but not that they are explanatory. Undoubtedly, the explanatoriness of SO-CFs can't just be a matter of counterfactuals having the correct syntactical or logical form. Considering interventionism is a realist theory, we also need to say where the causal *oomph* comes from. A metaphysical criterion is needed for models to individuate systems in a realist sense, and a criterion for the aptness of models (see §1.2.3), albeit having a minimal or deflationary character, can fulfil this role. Nonetheless, this is only a tentative response, for more work on what makes interventionist causal models fittingly capture causal patterns is needed.

[143] Though they may of course represent particular systems, like the Moon.

electrical current of the water in the tank again to see whether this generalisation obtains. Nonetheless, interventions *can* but *need not* be actual. Actuality is interesting insofar as it provides human beings with direct evidence for the meaning of counterfactuals. Yet, as interventionism is an objectivist account of causation, interventionist counterfactuals have determinate meanings and truth values regardless of our epistemic success in uncovering causal relations. In fact, when reasoning counterfactually most of the times we don't actually go and check the truth or falsity of the entertained counterfactuals via experiments. In the sea water example, the chemistry teacher, who is familiar with the extant dependence relation between electrolytes and electric conductivity given the relevant background information, is able to assess the counterfactuals DISTIL'/DISTILL"/REPLACE' without intervening on any actual system.

At this point, one may wonder what object's properties' (represented by variables taking different values) are being changed in this example. If he is at the laboratory, in front of the water tank, reasoning counterfactually in this way may boil down to hypothetically attributing properties to the concrete water tank. But if he is back at the shore discussing the matter with his daughter, things get trickier, for there's no such system in sight. Then, the object of his counterfactual exercise is *any* such system. So what object is *instantiating* the relevant properties in hypothetical interventions like the one considered in the shore scenario?

In this case, the most immediate response would be to say that he is hypothetically ascribing properties to any such actual system. This is the option endorsed by interventionism: token-level relations are those that do or would obtain at the level of particular, spatiotemporally located individuals (Woodward 2003, p.40), and hypothetical interventions the "*hypothetical* values of *X* and *Y possessed by the very object o*" (Woodward & Hitchcock 2003, p.20). Note that this is none other than the direct view of models: supporters of the direct view defend the claim that models are special descriptions which directly describe actual systems differently than they actually are (see e.g. Toon 2012 or Levy 2015).

There is a main problem with this view, which is that – as the name indicates – it cannot accommodate cases of modelling not directly aimed at any actual object. It is not obvious that the SO-CFs regarding the H_2O-CONDUCTIVITY case study the chemistry teacher entertains in the shore scenario is a case of direct modelling, for the modelling is not directed at any specific sample of sea water. The point is hopefully clearer if one considers idealisations: it is by no means trivial that the SO-CFs which appeal to idealised models directly describe actual objects, for there are no such things like ideal pendula or ideal gases out there.

One alternative view is to consider that the hypothetical interventions SO-

CFs about generic explananda describe should be interpreted as instantiations of hypothetical properties on hypothetical objects, which in turn represent a particular actual object or classes thereof. This is none other than the indirect view of scientific models. Supporters of the indirect view of models (see, e.g. Thomasson 1999, Godfrey-Smith 2006, or Frigg 2010a) call these hypothetical objects 'model systems', and the physical systems they represent 'target systems' (Frigg 2010b, p.252).[144] They are divided between realists about these hypothetical objects, such as Thomasson, and anti-realists who offer a deflationary version thereof such as Peter Godfrey-Smith or Frigg. Realists are accused that their view is not ontologically parsimonious, because it postulates the existence of a hypothetical object for any instance of modelling and thereby poses the difficulty of comparing non-actual with actual objects. Anti-realists evade the parsimony objection, even though they ought to justify how to make sense of hypothetical property instantiation in hypothetical objects which are not real (see Toon 2012, p.58, Levy 2015, pp.789-790).

The challenges each camp faces will be discussed in more detail in chapter 5, but hopefully the brief exposition here suffices to make the point that the metaphysical conundra the indirect view faces are offset by its inclusivity: it can accommodate both SO-CFs such as DISTIL'/DISTILL"/REPLACE' and SO-CFs about ideal systems. This is, I believe, a considerable reason to prefer an indirect over a direct view of models. In the indirect view, SO-CFs may be understood as providing information about a concrete hypothetical object – a hypothetical sample of sea water, or an ideal pendulum – which in turn represents a particular actual object, like the Moon, or a class of actual objects, such as real samples of sea water or real pendula.

In short, the advantages of characterising the instantiation of hypothetical interventions via the indirect route are as follows. Firstly, this reformulation acknowledges the fact that modelling is indirect: we usually first toy with the model relations, and only then consider if – and, if so, how – these map onto physical systems. Causal dependencies like those encoded in the ideal gas law describe the behaviour of a hypothetical object, the ideal gas, which in turn represents a given class of physical systems. This is important because, as I will discuss in chapter 5, counterfactual reasoning is essentially model-based reasoning. Secondly, it accounts for the variable-relative definition **SO-CF** given in §4.2: SO-CFs are about classes of all those possible – actual and non-actual – objects individuated

[144] While they often represent target systems, they don't need too. However, I do exclude targetless models like the Game of Life (see Weisberg 2013, p.129) from the present discussion. While one may play with these targetless models, hypothetical changes in the values of their variables are not interventions, for they don't represent causal relationships.

by a model's variables.

4.3.3.3 Types of Generality & Explanatoriness

In §4.3.3.1, I showed how according to standard scientific methodology it is classes of systems models, and thereby SO-CFs, typically individuate. This conclusion puts some pressure on the dichotonic distinction the interventionist draws between notions of generality and the explanatoriness (or lack thereof) they warrant. If SO-CFs typically describe classes of objects, it is because the system-internal generalisations they support have a certain *scope*, i.e. they apply to a certain number of different objects of a certain kind.[145] In other words, SO-CFs presuppose a generality regarding number of objects.

This is clear when we compare the 'depth' of different explanations. Consider the following example: we want to explain the motion of a car.[146] We can use Newton's laws of motion to determine how, say, its velocity affects its displacement. The generalisation G:x=$v(t)$ will be invariant under a certain range of interventions on the velocity V. In this case, a relativistic correction to G – call it G_R – won't offer a deeper explanation of the car's motion. Here is why. Given that m is neither an explanandum nor an explanans variable, we keep it fixed at its actual value, i.e. the mass of the car. There is an upper limit to the velocity the car can reach, past which the engine would break down. We could definitely use G_R to describe the car's motion, though it wouldn't be invariant under a wider range of interventions than G is; it would just offer more accurate values of the velocity.[147] To get a sense of the explanatory potential of G_R one ought to consider an object that is able to reach relativistic speeds, such as a rocket. In this case, one would not dare to use G to describe the rocket's motion, for it would lead to highly inaccurate results. Let's call the range of interventions on the velocity of the rocket under which G_R is invariant V_R. The idea is that the explanation with G_R is *deeper* than the explanation with G because $V \subset V_R$; the former explanation answers more what-if-things-had-been-different questions regarding the dependence x=$v(t)$ than the latter.

But is it depth really the source of explanatory potential? Note that V and V_R are ranges of interventions on *different* objects. The higher explanatory potential of G_R with respect to G is only revealed once we focus on objects moving at relativistic speeds. While I subscribe to the interventionist's view of explanatoriness as a

[145] Recall that *breadth* of invariance is the relative version of scope (see §3.4).

[146] This example is inspired by Woodward & Hitchcock (2003, p.185).

[147] This is similar to the Hooke's law example: the range of admissible values for the displacement of a spring are those within which the spring does not break.

matter of degree of invariance, the interventionist ought to concede that comparing the depth of explanations indirectly presupposes scope, i.e. generality with respect to other objects. Judgements of explanatory depth are only meaningful provided that $V \subset V_R$; without presupposing that G_R applies to objects other than those to which G applies, namely, not just cars but also rockets, the comparison is pointless.

To conclude, the individuation challenge has hopefully shown that the stark distinction between, on the one hand, generalisations' explanatory depth and on the other hand, generalisations' scope, is too simplistic a characterisation of explanation. The generalisations supporting SO-CFs presuppose a certain scope, which plays a key role to their explanatory potential. This is the topic of the following section.

4.4 The Idealisation Challenge

Another area where spelling out the object of SO-CFs is crucial is in the context of explanations with idealised models. In the indirect view of models, idealised models describe ideal states of affairs, which in turn represent physical states of affairs. Or, rather, as I will argue in §4.4.1, they *mis*represent physical states of affairs. This misrepresentation is due to the fact that idealising assumptions violate invariance patterns. In the following, I look at what I call 'idealising counternomics,' invariance-violating counterfactuals we often encounter in model-based explanation whose antecedents are nomologically impossible antecedents in virtue of the idealisations presupposed by or explicitly described in them.

There is a widely-held intuition that idealising counternomics are somewhat explanatory of the behaviour of the physical objects the ideal models in question represent, yet due to their reference to impossible states of affairs their classification as SO-CFs is controversial. At the same time, I show they can't be straightforwardly classified as OO-CFs either. In §4.4.2, I introduce the 'idealisation challenge,' the challenge to determine what explanatory role, if any, interventionism reserves for idealising counternomics.

In §4.4.3, I contend that idealising counternomics are OO-CFs, though unlike other OO-CFs they have a distinct *explanation-enabling* function: the function to identify the objects to which causal generalisations apply and thereby *enable* causal explanations. I call the counterfactuals with an explanation-enabling function 'SO-enabling-counterfactuals' (SO-enabling-CFs for short) and argue that idealising counternomics are of this kind. In particular, idealising counternomics enable causal explanations with idealised models.

Finally, in §4.4.4 I propose an interventionist strategy to identify idealising counternomics and thereby single them out from the broader class of OO-CFs. Unfortunately, this strategy can only deal with a specific kind of idealisation whose removal does not destroy the explanation in question.

4.4.1 Idealisation in Science

Idealisations play a key role in scientific theories and models: they are *distortions*, *misrepresentations* of physical systems, introduced deliberately for different purposes, such as tractability or explanatory purposes. There is a great amount of literature which contains many sometimes conflicting characterisations of idealisation, as well as differing attempts at outlining the differences between idealisation and other types of representations such as abstraction.[148] As a starting point, I briefly summarise Weisberg's seminal paper 'Three Kinds of Idealization' (2007b), where he introduces a useful taxonomy for idealisations in science. Weisberg emphasises the role of idealisation as an activity or process rather than a product and, hence, as a property of the theory-world relationship (see Weisberg 2007b, p.640). As an activity, idealisation obeys different rationales, which help classify idealisations into three main kinds: *Galilean, minimalist* and *multiple-models*.

The 'Galilean' idealisation (McMullin 1985) is the most common one. Galilean idealisations are distortions of targets introduced in order to get traction on them: we come up with idealised models so that we can give a computational or mathematical characterisation of a system's behaviour. The justification for these distortions is thus tractability: they are introduced to simplify complicated differential equations into analytical ones. In Galilean idealisation there is an implicit assumption that, should there be advances in computational and mathematical techniques, these models could be made more specific via a successive removal of the distorting assumptions (Weisberg 2007b, pp.641-642). Newtonian mechanics is full of Galilean idealisations: massless objects (e.g. the string and bob in the simple pendulum model), point-like masses (e.g. the planets in Newton's model of universal gravitation), frictionless surfaces, etc.

Next, there is 'minimalist' idealisation. Minimalist idealisations are distortions of physical systems introduced for explanatory purposes: target systems are distorted by considering only the core causal features of the target phenomenon. This is why the resulting idealised model is called a *minimal model* of the phenomenon (Weisberg 2007b, p.642). One representative minimalist view of idealisa-

[148] One could also include approximation or fictionalisation as separate categories in this list, depending on how these are characterised.

tion is Strevens' *kairetic* account of causal explanation (Strevens 2008), for whom idealisation is the introduction of false yet non-difference-making causal factors into a causal explanation, such as the thermodynamic limit assumption. This type of idealisation is very similar to abstraction. This is why Weisberg also includes Batterman's (2002) account of asymptotic explanation and Cartwright's account of abstraction in this category.[149] However, there are nuances between minimalist idealisation and abstraction, which will be explored in §4.5.1.

Finally, there is 'multiple-models' idealisations, which I will briefly explain for completion. This is the practice of building multiple models which make different and often conflicting claims about the explanatory structure of the phenomenon in question (Weisberg 2007b, pp.645-646). One finds this construction of multiple models in fields which deal with complex phenomena, such as ecology, or meteorology. Here, these representations strive toward different goals and pursue different desiderata.

This classification will serve as the backbone against which I will tackle the idealisation challenge. However, I believe the focus on the motives behind the idealisation process is not the right angle from which to approach this classification. First, because in practice it is difficult to disentangle tractability from explanatory aims. Consider the ideal gas model; the assumption that gas molecules do not collide with each other seems motivated not only by the aim to get a simpler linear equation, but also by the aim to focus on the aspects of the gas which are causally relevant to explain its behaviour. Furthermore, and specific to *W-H*'s account of scientific explanation: Woodward's and Hitchcock's focus is models, not the intentions of modellers. In other words, while the product, the idealised model, results from the process of idealisation, it is eventually models which SO-CFs describe. For these reasons, in the context of the idealisation challenge I believe there is a better way to characterise this classification in terms of the *de-idealisation* of models. Galilean idealisations will be those which allow for de-idealisation without loss of explanatory power – reason why Elgin & Sober (2002, p.447) call these *harmless* – whereas minimalist idealisations are those for which de-idealisation is not possible because it is essential to *explain* or *explore* the phenomena in question.

Shech (2018) makes a further distinction between essentialists, which are those who advocate an explanatory indispensability role for idealisations, and those who are not essentialist about their *explanatory* but about their *exploratory* role, that is, those who think that idealisations may play an indispensible role in "the

[149]For more details on Cartwright's view on abstraction see Cartwright (1989) and Jones (2005).

exploration of scientific theories, intertheoretic relations, and phenomena" (ibid., p.11), and which thereby occupy a middle ground position between essentialists and dispensabilists about idealisation. Given the scope of this chapter, I won't distinguish between these two essentialist positions. To get an insight into the nuances of the debate, see Shech (2018).

4.4.2 What is the Idealisation Challenge? Idealising Counternomics

As I argued in §4.3, models in science often incorporate idealising assumptions; i.e. they are idealised. Consequently, in the context of scientific explanation there will be counterfactuals which appeal to idealised models.[150] These are counterfactuals such as the following:

> **IDEAL GAS**: If the pressure of a volume of ideal gas were to increase by an amount $\Delta P_{12} = P_2 - P_1$, its temperature would consequently increase by an amount $\Delta T_{12} \propto \Delta P_{12}$.

IDEAL GAS is a counterfactual associated with an idealisation, where the idealisation is that the ideal gas law, $PV{=}nRT$, holds. In the ideal gas law, P is the system's pressure [kPa], V the system's volume [m^3], n the number of moles [kg mol], T the temperature (K), and R the gas constant with the value 8.314 [kPa m^3/kg mol K]. The interventionist would deem IDEAL GAS true because the ideal gas law tells us exactly what the effect of a change in the value of P on the value of T would be.

However, ideal gases are physically impossible objects. The most likely usage of IDEAL GAS is in a context where the ideal gas law is used to *approximately* predict or explain the behaviour of a real gas in certain conditions. One such example is krypton, a monoatomic gas – i.e. composed of just one atom, and thereby lacking any covalent bonds – which at standard temperature and pressure conditions qualitatively behaves like an ideal gas. In such context, we might entertain the following counterfactual (Tan 2019, p.45):

> **KRYPTON**: If krypton were an ideal gas, then it would obey the ideal gas law $PV = nRT$ perfectly.

Likewise, consider the case of water, which can be approximated as a continuous and incompressible fluid, and as such describable by the Navier-Stokes equations of continuum mechanics. In this other context, we might utter the following counterfactual (Tan 2019, p.46):

[150] This tight connection between scientific models and counterfactuals will be explored in chapter 5.

> **WATER**: If water were a continuous, incompressible medium, then its behaviour would be described by the Navier-Stokes equations.[151]

What is distinctive about these two counterfactuals is that their antecedents describe physically impossible states of affairs, which are so in virtue of the idealisations to which they appeal. For instance, in the case of KRYPTON, for a gas to be ideal its particles must have null volume, which puts Newtonian and intermolecular forces on hold. The kind of impossibility it targets is physical because realising the antecedent requires that some invariant dependencies are broken or put on hold, resulting in ideal versions thereof.

In the case of WATER, it is physically impossible for water to be continuous, for water is made of discrete H_2O molecules. Water is also not incompressible, because a real fluid's motion does involve tangential forces when it comes to contact with another body; intermolecular attractions cause fluids to adhere to solid walls, causing shearing stresses (Morrisson 1999, pp.53-60).[152] The magnitude of the shearing stresses depends on a fluid's viscosity; since water's viscosity is very low, viscosity is neglected by setting the viscosity coefficient μ to zero. In cases of small viscosity, a viscous fluid's velocity relative to a solid boundary will be null. This idealising assumption is known as the 'no-slip condition,' and is a Dirichlet boundary condition (see chapter 3) which leads to important simplifications in the Navier-Stokes equations: in an incompressible fluid, its density ρ is constant in space and time, so all its derivatives are zero. This is equivalent to setting the divergence of the flow velocity u, $\nabla \cdot u$, to 0.

Both KRYPTON and WATER are examples of idealising counternomics, i.e. counterfactuals with antecedents that describe physically impossible scenarios as a result of idealising assumptions. On this point, Tan (2019, pp.45-46) argues that treating water as a continuous fluid is a *metaphysical* impossibility, and I believe some might join him in contesting the kind of modality I attribute to these impossibilities, on the grounds that these changes violate metaphysical principles such as the Kripkean necessity of identity and as such are not metaphysically well-defined.[153] I will not contemplate whether idealising counternomics are in fact

[151] The Navier-Stokes equations are set of five coupled differential equations that describe how the velocity, temperature, pressure, and density of a moving fluid are related. There is one for conservation of mass, one for conservation of energy, and three for conservation of momentum. They are based on the assumption that fluids are continuous. The assumption that fluids are incompressible results in more tractable Navier-Stokes equations.

[152] See Morrisson (1999) for a detailed account of Ludwig Prandtl phenomenological model of viscous fluids with thin boundary layers.

[153] Tan's argument is twofold. Treating water as continuous and incompressible is metaphysically impossible because either water is necessarily H_2O or it is metaphysically impossible for a continuous and incompressible fluid to play the functional role of water.

counterfactuals with metaphysically impossible antecedents; what is important is that these counterfactuals describe *at least* physically impossible states of affairs.[154] I here assume a nested model for possibility, so what is metaphysically impossible is also physically impossible. Whether these counterfactuals describe metaphysical impossibilities is a matter that will be relative to one's metaphysical views on necessity and identity.

The focus on idealising counternomics obeys the following reason: they are of interest because even though they look like OO-CFs, they aren't wholly non-explanatory. In fact, as I will argue in chapter 5, they are non-vacuously true. But what makes them so? The interventionist faces the 'idealisation challenge:' the challenge to account for the role of idealising counternomics in scientific explanations.

As a first observation, note that idealising counternomics are not archetypal SO-CFs: SO-CFs describe (physically) possible interventions, such as whether changes to the P of a gas sample have any effect on its T, or whether administering drug D to a patient alleviates their symptom S. These interventions test the invariance of system-internal generalisations: i.e. they give us information about the modal properties of a given – *same* – object. This is not what is at stake in the case of idealising counternomics, which describe how physical systems would behave if they were ideal. Consider KRYPTON: there is no possible intervention that transforms a real gas into an ideal one. This feature explains why idealising counternomics don't take the traditional form **SO-CF**. If anything, they seem to take the form **OO-CF** (see §4.2.1):

> **OO-CF**: If some object o^*, different from o and that does not possess property A, were to be an A, then it would be a B.

Table 4.1 below summarises how one would turn KRYPTON and WATER into OO-CFs.[155]

[154] See Bhogal (2020) for a contrary view.

[155] Note that the coarse-grainedness of property B in both examples is due to the fact that these are fairly schematic explanations. In more complex explanations, properties like B would likely be given a more fine-grained treatment.

OO-CF	KRYPTON	WATER
Object *o**	Krypton	Actual mass of water
Object *o*	Ideal gas	Ideal mass of water
Property A	Having a null interaction potential & molecular volume	Being continuous
Property B	Behaving according to the ideal gas law	Behaving according to the Navier-Stokes equations

Table 4.1: KRYPTON & WATER as OO-CFs.

Qua syntactic form, KRYPTON and WATER clearly fit the other-object camp. The problem is that granted that Woodward and Hitchcock's SO vs OO distinction is indeed exhaustive, classifying idealising counternomics as OO-CFs leaves them outside the realm of causal – and therefore, scientific – explanation. This is a problem because, unlike with archetypal OO-CFs, we are able to make sense of the meaning of idealising counternomics. As such, they must be explanatory to some extent, but it is a mystery how the interventionist account of causal explanation may account for these differences in explanatoriness among OO-CFs.

The sharp contrast between SO-CFs, and thereby causally explanatory counterfactuals, and OO-CFs, and thereby causally non-explanatory counterfactuals, fails to accommodate idealising counternomics, which do not fall clearly in either category. Consequently, the interventionist seems obliged to assign no role for them in scientific explanation. In §4.4.3, I provide one possible resolution to the idealisation challenge.

4.4.3 The Solution to the Idealisation Challenge

We have no clue what sort of intervention could possibly turn Bill Clinton, the government of Italy, or a neutron into an electric conductor, which is the sort of manipulation CONDUCTOR describes (see §4.2). Similarly, we have no coherent idea of what it would take for a whale to turn into a fish (see Williamson 2017, p.195). These are *archetypal* OO-CFs: they describe impossible interventions that involve a wild transformation of one physically possible object into another object of a radically different kind, to which it is not causally related.

KRYPTON or WATER also describe impossible interventions: it is impossible to, for instance, intervene on a real gas to make its particles' volume negligible or eliminate all intermolecular interactions. That said, this intervention is less far-fetched or radical than the ones described by archetypal OO-CFs – this is because

the objects involved in idealising counternomics are unusually close. Consider KRYPTON anew. Because there is no such thing as non-interacting and point-like gas particles, the ideal gas is a physically impossible object which is usually considered to fairly well *represent* physically possible objects like light real gases without strong intermolecular forces in certain conditions such as when they are not too cold or the pressure is not too high. KRYPTON thus describes what it would take for a physically possible object like krypton (*o**) to behave *as if* it were an ideal gas (*o*).

Idealising counternomics differ from archetypal OO-CFs in that they consider transformations of physically possible objects into impossible objects which represent them. In model talk, they describe what it would take for a *target system* to be exactly like the *model system* by which they are represented (see §4.3.3.2). This, in turn, is what accounts for their role in scientific explanations which, as I now explain, is to enable model-based causal explanations.

Both in chapter 3 and in §4.2.2, we saw how the Invariance View dispenses with the idea that explanatoriness reduces to exceptionlessness. But if laws don't apply universally, then we need an account of how to delimit the scope of causal generalisations. I call 'SO-enabling-CFs' those counterfactuals whose function is precisely to identify the objects to which causal generalisations apply. This is where *scope* meets *depth*: to use a causal generalisation to uncover an object's modal profile we first need to know whether the causal generalisation applies to the object in question. Since this is the sort of information SO-enabling-CFs provide, they might be seen as *enablers* of causal explanations (and, by extension, the SO-CFs supported by these). The best solution to the idealisation challenge is, I believe, to classify idealising couternomics as OO-CFs, given that they all have the form **OO-CF** ('If some object *o** were *o*, then...'), albeit with a distinct causal *explanation-enabling* function. More specifically, idealising counternomics are SO-enabling-CFs which enable causal explanations with idealised models.

This explanation-enabling function by which some counterfactuals are characterised resembles the sort of function Hüttemann (2021, Ch.1) reserves for system-external generalisations. Recall that he calls 'system-external' those generalisations which define the domain of applicability of laws: e.g. which objects *b* are covered by the ideal gas law, a presupposition for any explanation involving the ideal gas law to work (see §4.2.2). It is for this reason that system-external generalisations are attributed a "role justifying ground" in explanations, which means that they are *not* part of the explanans, but rather a "*presupposition* of the explanation" (Hüttemann 2021, p.19).

Another example of generalisations which, while not explanatory them-

selves, have an explanation-enabling role can be found in Strevens' kairetic account of causal explanation (Strevens 2008). In Strevens' account, system-external generalisations correspond to what he calls 'basing generalisations,' true generalisations which specify the intrinsic properties of the systems falling within a causal generalisation's scope (ibid., p.229). Strevens' account of scientific explanation accommodates the fact that not all generalisations play an explanation-enabling – what Strevens calls 'basing' – role. For instance: while the generalisation 'All ravens are black' is explanatory, 'All hexed ravens are black' is not, for the latter does not describe a relevant pattern of counterfactual dependence.[156]

The idealisation challenge and my solution to it are connected to well-known problems in the literature on scientific explanation. The question of how to identify SO-enabling-CFs from the broader class of OO-CFs is very closely related to the question of how to distinguish the system-external or basing generalisations from accidental generalisations. As both Hüttemann's and Strevens' book-length projects prove, this is no easy feat. That said, my focus is narrower than theirs: the idealising challenge concerns causal explanations with idealised models only. In §4.4.4, I develop an interventionist strategy to identify idealising counternomics from the broader class of non-explanatory OO-CFs, namely, a strategy to identify the counterfactuals which enable causal explanations with idealised models.

4.4.4 Finding SO-enabling-CFs Involving Idealisation

Why does KRYPTON enable explanations while the counterfactual 'NOELIA: If Noelia were an ideal gas, then she would obey the ideal gas law $PV = nRT$ perfectly,' doesn't?[157] The explanation-enabling role of KRYPTON can be justified via the fact that the ideal object it appeals to can be *approximated* by the physically possible object it represents: that is, an ideal gas can be approximated by krypton, but not by me.

Here I present an interventionist strategy to identify SO-enabling-CFs which enable causal explanations with idealised models. In a nutshell, the idea, applied to the counterfactuals above, is this: we can't intervene on real gases to turn them into ideal gases, but we can intervene on certain real gases in such a way that they approximate the behaviour of an ideal gas. So, while KRYPTON cannot be made true by interventions, there are well-defined SO-CFs in the vicinity of KRYPTON

[156] He proposes a very elaborate counterfactual test to distinguish explanatory generalisations from non-explanatory generalisations. I won't discuss it here, for that would divert us from the idealisation challenge. See Strevens (2008, Ch.7) for more details.

[157] Note that CONDUCTOR doesn't enable explanations either, though because the present strategy is restricted to counterfactuals which appeal to idealised models the suggested counterfactual poses a more relevant contrast.

that describe how interventions on samples of krypton affect their *closeness-to* ideal behaviour. Because there are such SO-CFs in the vicinity of KRYPTON, we can say that KRYPTON enables explanations of samples of krypton via the ideal gas model. In the following, I will generalise this interventionist strategy to identify explanatory-enabling counternomics appealing to idealisations.

To begin with, let me outline some definitions. First, I rely on the indirect view of modelling, whereby an idealised model directly represents an ideal *model system M*, which in turn denotes a physical *target system T* (see §4.3.3.2). For instance, in the case of KRYPTON, the model system is an *ideal gas*, which in turn represents the target system *krypton*. The argument unfolds as follows:

P1. SO-CFs describe *possible* interventions.
P2. OO-CFs describe *impossible* interventions.
P3. T is a possible system.
P4. M is an impossible system.
P5. Idealising counternomics describe interventions that would turn T into M, $I_{T>M}$.
P6. Idealising counternomics describe *impossible* interventions $I_{T>M}$. (from 3, 4, 5)
P7. Provided *quasi-gradual de-idealisation* $M \rightarrow T$ is possible, there are possible interventions *sufficiently close* to $I_{T>M}$.
P8. Provided *quasi-gradual de-idealisation* $M \rightarrow T$ is possible, there are SO-CFs describing interventions *sufficiently close* to $I_{T>M}$. (from 1, 7)
P9. Provided there are SO-CFs describing interventions *sufficiently close* to $I_{T>M}$, there are SO-CFs in the vicinity of idealising counternomics. (from 6, 8)
P10. Provided there are SO-CFs in the vicinity of idealising counternomics, M causally explains T. (from 9)

C1. Idealising counternomics are OO-CFs. (from 2, 6)
C2. Idealising counternomics enable M-based causal explanations of T. (from 10)

In the following, I shall call the SO-CFs describing states of affairs close to the ideal ones idealising counternomics describe 'counterpart counterfactuals.'

Note that the first part of the argument concerns the identification of idealising counternomics as OO-CF. P1 is none other than the definition of SO-CFs as interventionist counterfactuals; P2 identifies counterfactuals which describe impossible interventions as OO-CFs. P3 and P4 respectively indicate that, for

idealised models, T is a physically possible object, and M, qua idealised object, a physically impossible one. P5 is the definition of idealising counternomics as counterfactuals which describe interventions $I_{T>M}$.

Moving onto the actual argument, P6 describes the fact that, since M is a physically impossible object, the intervention $I_{T>M}$ is impossible. To illustrate it, consider KRYPTON. An intervention on krypton to turn it into an ideal gas would have to make the volume of the gas particles negligible compared with the total volume occupied by the gas, make the gas particles behave as rigid spheres colliding elastically with each other, and erase intermolecular forces between the gas particles. Since the laws of nature forbid the existence of such ideal objects, there are no such possible interventions. As a result, idealising counternomics describe physically impossible interventions. From this, provided the SO vs OO distinction is exhaustive, as per P2 we conclude that KRYPTON and, more generally, any idealising counternomics are OO-CFs (C1).

The second part of the argument concerns the role of idealising counternomics in causal explanations. P7 is one of the most central premise here, for it establishes the connection between the physically impossible and possible modal domains. This close connection is what enables the approximation of idealising counternomics by regular counterfactuals which serve as indicators of the explanation-enabling power of the original counternomics.

Given its centrality in the argument, let us carefully unpack P7. To begin with, let me note that by 'de-idealisation' I refer to the adding back of details to models (e.g. increasing the mass of particles, the strength of their interactions, etc.). This process assumes that there is, in principle, a continuum (or almost continuum) between the most specific model one could give of a target – i.e. a *full* representation of a target – and a given idealised model thereof, and M can be de-idealised without loss of explanatory (or exploratory) power (see §4.4.1). We can understand this adding back of details as uncovering an almost continuum of physically possible states, where perfect continuity is only broken by the initial state, M, which is not physically possible. Because of this 'modal' jump, I have chosen the term 'quasi-gradual' to refer to this de-idealisation process, which hopefully conveys a sense of identifiable continuity even if it does not describe de facto continuity. By de-idealising M we move to description of a state of affairs which is no longer physically impossible and, as such, whose behaviour can be investigated via interventions (P8). The qualifier 'sufficiently close' indicates the modal proximity of these physically possible interventions to the impossible interventions $I_{T>M}$.

For instance, in the case of KRYPTON, an infinitesimal increment on the

values of variables representing properties such as mass or volume already moves us to the space of physical possibility. The object in question is now a candidate for interventions (P8). Premise P9 links the modal proximity of physically possible and impossible interventions to the modal proximity of the counterfactuals that describe these. Provided a quasi-gradual de-idealisation $M \rightarrow T$ is possible, there will be counterpart SO-CFs in the vicinity of idealising counternomics.

To illustrate the inference of P8 from P7, let us examine the case of KRYPTON again. One could attempt it to give it the syntactic form of a SO-CF as follows:

> **KRYPTON**$_{int}$: If an intervention were to change the density D of a volume of krypton in a container from $D = d_1$ to $D \rightarrow d_0$, the gas would obey the ideal gas law $PV = nRT$ perfectly.

In the expression above, $D = \frac{n}{V}$.[158] The value d_1 is such that the coefficients a and b in, say, the van der Waals equation of state (see eq. (4.4) below) are significant, and d_0 a "value" for which they are perfectly negligible.[159] Therefore, $d_0 < d_1$. Coefficients a and b account, respectively, for the forces among molecules and the molecules' volumes, and are characteristic of every gas. For instance, for krypton they take the values a_{Kr}=5.125 $atm\ dm^6\ mol^{-2}$ and b_{Kr}=1.06·10^{-2} $dm^3\ mol^{-1}$ (Atkins 2018, p.899).[160]

$$\left(p + a\left(\frac{n}{V}\right)^2\right)\left(V - nb\right) = nRT \tag{4.4}$$

We will also assume that the value of n is large enough so that the particles of the gas move in random motion, and that we are far from the vapour-liquid equilibrium line so that the gas can be described classically at all times. More precisely, we assume that the gas is within the following regime (Baierlein 1971, Ch.9):

$$\frac{n}{V}\lambda^3 >> 1 \tag{4.5}$$

In eq. (4.5), λ is the thermal de Broglie wavelength $\lambda = \left(\frac{mk_BT}{2\pi\hbar^2}\right)^{-\frac{1}{2}}$.

Since what we are interested in is whether changing krypton's density makes it perfectly obey the ideal gas law, we can introduce a binary coarse-grained variable I for whether the volume of gas perfectly obeys the ideal gas law, where

[158] In fact, n/V is the density except for a factor 1/M. The density is $\rho = m/V = n{\cdot}M/V$.

[159] Arguably, wee could also define d_1 as a value for which the gas approximates but does not perfectly obey the ideal gas law.

[160] This corresponds to table 1C.3 in the Resource section.

$I = 1$ if it does and $I = 0$ if it doesn't, to study $I = I(D)$.[161]

Note that while KRYPTON$_{int}$ now has the appearance of a SO-CF, it is not, for the post-intervention state ($D \rightarrow d_0$) does not correspond to any point in the physical modal space: there is no physical object that *perfectly* obeys the ideal gas law, hence the limit. Since such intervention is impossible, the content of KRYPTON$_{int}$ can't be evaluated, so at best KRYPTON$_{int}$ is trivially true. Nonetheless, KRYPTON$_{int}$ can be approximated by the following counterpart SO-CF:

> **KRYPTON$^*_{int}$**: If an intervention were to change the density D of a volume of krypton in a container from $D = d_1$ to $D \approx d_0$, the gas would obey the ideal gas law $PV = nRT$ *approximately*.

Unlike KRYPTON$_{int}$, KRYPTON$^*_{int}$ is a well-defined SO-CF because $D \approx d_0$ is a physically possible value. As such, the intervention described by KRYPTON$^*_{int}$ is well-defined: there are possible interventions on the density of the gas that affect how approximately ideal its behaviour is. Since it is still the case that $d_0 < d_1$, $I(d_0) > I(d_1)$. However, note that KRYPTON$^*_{int}$ requires that we now treat I as a continuous variable representing to what degree a gas approximates the ideal gas law, and thereby a variable that takes values in the interval $(0, 1) \in \mathbb{R}$. Via this example, I hope to have illustrated what it means that there are counterpart SO-CFs in the vicinity of idealising counternomics when quasi-gradual de-idealisation $M \rightarrow T$ is possible.

Back to the argument, P10 is another fundamental premise which establishes that whenever idealising counternomics have counterpart counterfactuals, this signals the explanation-enabling role of the former in explanations involving idealising assumptions. In other words, should quasi-gradual de-idealisation be possible, this highlights the aptness of the causal representation of T by M. For instance, gradual de-idealisation from M does bring us back to a full representation of T=*krypton*, for if we intervene on a sample of krypton to decrease its density, its behaviour is better approximated by the ideal gas law. From P10, we conclude that idealising counternomics enable explanations of the targets in question by the idealised models under consideration (C2). Note that, together with C1, this is the expected outcome: idealising counternomics are OO-CFs which enable causal explanations with idealised models.

[161] Again, to be precise, not the density but the molar density $\frac{n}{V}$.

4.4.4.1 Analysis

In view of the idealisation challenge posed by idealising counternomics, which do not straightforwardly fit the binary classification of counterfactuals regarding their (non-)explanatoriness suggested by interventionists, in §4.4.3 I offered a solution which classifies them as SO-enabling-CFs, namely, OO-CFs with a distinctive explanation-enabling function.

Nonetheless, how do we discriminate the SO-enabling-CFs from the broader class of OO-CFs? In §4.4.4, I suggested that explanatory-enabling idealising counternomics have counterpart counterfactuals which result from a quasi-gradual de-idealisation process $M \rightarrow T$. When this sort of de-idealisation is possible, we are entitled to describe T using M. Consider the general counterfactual 'If b had been an ideal gas, it would have behaved in accordance with the equation PV=nRT' (Hüttemann 2021, p.16).[162] What the proposed interventionist strategy to justify the explanation-enabling function of certain counterfactuals does is provide a means to justify why whenever b=*krypton* (but not when, say, b=*Noelia*) we are justified in using the ideal gas law to describe the behaviour of the object in question, in this case krypton.

Under the current proposal, KRYPTON is not part of any explanation which appeals to the ideal gas law, but it offers an explanation of *why* we can use the ideal gas model to represent krypton's behaviour but not me. In other words, KRYPTON backs explanations which appeal to the ideal gas model, and as such it accounts for the epistemic utility of the ideal gas model. This chimes with our initial intuition that KRYPTON is explanatory in some sense or other – unlike NOELIA, which is just a regular non-explanatory OO-CF.

To sum up, idealising counternomics constitute a distinctive subclass of OO-CFs: they are SO-enabling-CFs. In particular, they enable idealised model-based causal explanations by identifying the objects to which the causal generalisations at the core of idealised models apply. In §4.4.4, I proposed an interventionist strategy that identifies the origin of this explanation-enabling role in the possibility of de-idealisation from the model to the target. However, as I show in §4.4.4.2, this possibility of de-idealisation significantly limits the scope of the strategy.

4.4.4.2 Limitations

The interventionist strategy to identify SO-enabling-CFs which enable explanations with ideal causal generalisations applies only to idealised models which can be

[162] In the original, the ideal gas law is written as pV=vRT, equivalent to the formulation $PV = nRT$, to which I have stuck for consistency across the piece.

quasi-gradually de-idealised, i.e. it relies on *approximation*. Unfortunately, many idealising assumptions cannot be thus removed.

Consider the counterfactual WATER (see §4.4.2), which describes water as if it were continuous, or the idealising assumption that treats a hydrogen atom as if it were a Bohr atom. These are black-and-white idealising assumptions, in the sense that they don't lean themselves to de-idealisation because they are essential to explain or explore the phenomenon in question, be it the spectrum of the hydrogen atom or water's irrotational flow.

Another interesting subset of idealising assumptions are *infinite idealisations*. For instance, the thermodynamic limit of statistical mechanical systems is the assumption that systems are infinitely large, and is a necessary assumption to model them via probability theory. The thermodynamic limit is also essential to model genuine physical discontinuities such as phase transitions and critical phenomena. Infinite idealisations are also ubiquitous in special sciences such as biology; as in the following example from population genetics:

> **HWE**: If a diploid population X were in Hardy-Weinberg equilibrium, then its behaviour would follow the Hardy-Weinberg equation $p^2 + 2pq + q^2 = 1$.[163]

Where p is the dominant allele frequency, and q the recessive allele frequency in a given population. HWE incorporates all sorts of assumptions, from the assumption that mating is random to assumptions that ensure that the population is not affected by disruptive factors such as mutations, natural selection, or drift.[164] While mutation may be idealised away by setting its probability to zero, since genetic drift is a stochastic process and hence "built into the mechanics of meiosis itself" (Strevens 2019, p.1718), it cannot be idealised away in the same way. The way population geneticists get rid of genetic drift is by setting population size N (a parameter in these stochastic Hardy-Weinberg models) to infinity $N \to \infty$, an assumption which ensures that the effect of drift is negligible.

Furthermore, among infinite idealisation we should distinguish between those like the thermodynamic-type or HWE-type limits, which exhibit controllable asymptotic behaviour, from those like infinite idealisations at critical points, which exhibit uncontrollable asymptotic behaviour. The former are cases of 'well-behaved' mathematical limits, i.e. cases where the behaviour at the limit can be approximated by the limiting behaviour. On the contrary, in the latter case such

[163] This equation is the binomial $(p + q)^2$, which can be generalised to $(p + q)^n$, where n is the ploidy, i.e. the number of complete sets of chromosomes in a cell.

[164] See 'Hardy-Weinberg Equilibrium'.

an approximation is not possible, for the behaviour at the limit and the limiting behaviour are qualitatively different: they constitute *genuine* physical discontinuities (Batterman 2005, p.236).[165] Given that the interventionist strategy to identify SO-enabling-CFs presupposes a well-behaved approximation to the ideal state of affairs, genuine discontinuities fall outside its scope.

One question that remains open is what to make of controllable approximations to infinite idealisations, for they look like the diametrically opposed case to infinitesimal idealisations like those implied by KRYPTON. I have made the case that infinitesimal idealisations can be quasi-gradually removed without it having devastating consequences for the explanatory potential of the model in question, but infinite idealisations cannot. To answer this question, some further investigation is needed to clarify whether the possibility of de-idealisation is linked to underlying similarities – and, if so, of what kinds – between the models and the phenomena they represent.

Before concluding, I would like to preclude any remarks or objections along the following lines. In my analysis of the idealisation challenge, I did not provide a careful analysis of the modal status of the impossibilities idealising counternomics describe. In §4.4.2, I stressed that some philosophers have argued that idealising assumptions describe metaphysically impossible model systems: this is the case of Tan (2019), and Kimpton-Nye (2020, p.524) also seems sympathetic to this idea. Even if we set this inquire into metaphysical possibility aside, one might also argue that the status of idealising counternomics as *counternomics* is disputable. For instance, the idealising assumptions that a system is infinite or that agents are perfectly rational, some might say, don't presuppose any law violations but only that contingent facts of our universe had been different than they are.[166] According to the view on physical possibility presented in this thesis, it is not clear these counterfactuals are not law-violating. While WPP and VWPP are flexible with respect to ICs (see §3.5.3), both notions of possibility rule out interventions which are not permitted by the actual patterns of invariance. On this point, the sort of ICs required by the above mentioned idealising assumptions don't seem compatible with actual invariant relationships, be it lower- or higher-order ones.[167]

To sum up: there is a (certainly not small) subset of counterfactuals involving idealisations which the interventionist strategy presented in §4.4.4 cannot handle.

[165] This example will be more carefully examined in the context of abstraction.

[166] In conversation with Prof Alyssa Ney, she pointed out that the view that infinite systems are not law-violating is a rather usual one in the philosophy of physics.

[167] That said, should one disagree with my analysis, another plausible interpretation would be to treat these idealising counterfactuals as *counterphysicals*: they are not counternomics because they don't violate any laws but, but since ∞ is not a value, they still misrepresent physical systems. Note that this alternative interpretation would not preclude the idealisation challenge.

The previous result highlights the limitations of *W-H*'s interventionist account of causal explanation when it comes to discriminating between explanatory-enabling counterfactuals and non-explanatory counterfactuals. Whether *W-H*'s account can accommodate the explanatory-enabling role of these idealising counternomics in some other way is of course an open question, but I hope to have shown that it is far from easy to see how.

Many proposals to justify the explanatory role of idealised models have been developed. Bokulich, for instance, proposed to extend Woodward's view of explanatoriness as counterfactual dependence along interventionist lines beyond causation via her 'structural model explanation' account of explanation. In a structural model explanation, a model system explains a target system provided the former exhibits similar patterns of structural counterfactual dependence to the latter: the Bohr atom represents the hydrogen atom because there is an isomporphism between the counterfactual structure of the Bohr's model and the spectral phenomena (Bokulich 2011, p.43). A similar approach is taken by Psillos, who argues that "[idealised models] can provide causally relevant information about worldly systems because on their basis certain counterfactual conditionals can be settled" (Psillos 2011, p.16). First of all, it is not clear in which sense these are interventionist proposals. Second, and more importantly, these are views on how idealised models – which are misrepresentations – can explain, which do not focus on the explanatory roles of the associated counterfactuals. To fill the aforementioned gap, in chapter 5 I develop a fiction-based view of counterfactuals which not only provides a clear link between scientific models and counterfactuals but is also able to provide truth-conditions for counterfactuals regardless of the type of idealising assumptions they make – something *W-H*'s interventionist account of causal explanation can't offer.

4.5 No Abstraction Challenge

Models in science not only contain idealisations, but they also contain abstractions. In §4.5.1, I introduce representation via abstraction, and explain why despite the similarities between idealisation and abstraction there is no 'abstraction challenge' analogous to the idealisation challenge. This is because abstract models don't incorporate false assumptions and, as such, there is no special reason for doubt about how abstract models represent their targets.

Nonetheless, there are certain views of abstraction under which it is not straightforward that no abstraction challenge arises. One such view is the 'irrele-

vance' approach to non-causal explanation put forward by Batterman (2000, 2002, 2010)), Batterman & Rice (2014), and Pincock (2012, 2015), under which the resulting explanations, by moving away from the microphysical details, provide no same-object information. I introduce such a view in §4.5.2 but, as I argue in §4.5.3, it poses no threat to *W-H*'s account of causal explanation: Woodward explicitly rejects the connection between abstraction and non-causal explanation.

Next, I explain how the interventionist account of causal explanation has been modified in the light of the existence of non-causal explanatory dependencies, leading to what has come to be known as 'change-relating (CR-)monism.' In §4.5.4, I show how in 'Some Varieties of Non-Causal Explanation' Woodward explores the plausibility of CR-monism. There, he hints at a second view of non-causal explanation: in the context of renormalisation group (RG) explanation, he thinks these may be non-causal because they involve "...transformations *in a more abstract space*" (Woodward 2018b, p.124).[168] I argue that this assertion does not jeopardise his rejection of the irrelevancy view. There are two types of RG explanation: inter-system RG explanation (Sullivan 2019) and what I call 'alternative RG explanation'. As I explain in §4.5.5, only the former kind of explanation is causal and thereby involves SO-CFs. That said, RG explanation highlights the rich variety of counterfactuals that may feature in such an explanation: in particular, there are *stability-exhibiting* counterfactuals which seem to provide information about the same object. To conclude, I examine what stability-exhibiting disclose about the SO vs OO counterfactuals distinction.

4.5.1 Abstraction vs Idealisation

No analogous challenge to the idealisation challenge arises for abstract models. In order to see why, let me start by discussing the respects in which idealisation and abstraction differ.

Traditionally, the literature on scientific models highlights a main difference between idealisation and abstraction which concerns their **connection to truth**. Abstraction/abstractness refers to a truthful description's degree of detail, whereas idealisation concerns the deliberate introduction of 'falsities' in a physical system's representation (see Jones 2005, Godfrey-Smith 2009, or Levy 2021).[169] In other words, idealisations are *mis*representations of target systems because we lie about some aspects of the systems in question, whereas abstractions are *detail-poor*

[168] Italics mine.

[169] Levy (2021) introduces a process-product distinction; 'abstraction' is the process of 'leaving out' or 'stripping away' detail, whereas 'abstractness' is used to refer to the level of detail of the resulting representation.

representations because we neglect to mention certain aspects of the systems in question (see Jones 2005, p.174).

Regarding **purposes of modelling**, abstractions are – like minimalist idealisations – justified on explanatory grounds: the omission of detail in representations of physical systems respects a desire to move away from features that are unnecessary or irrelevant for the explananda.[170] What distinguishes abstractions from minimalist idealisations (see §4.4.1) is that the sort of omission involved in abstraction is "omission *without* misrepresentation" (Jones 2005, p.175). To illustrate the contrast, Jones discusses the case of modelling a fluid's flow as irrotational when it is not (ibid.). There is a sense in which one could interpret this assumption as omitting certain aspects of the fluid. However, since this omission involves a misrepresentation of the fluid as e.g. incompressible, this is not a case of abstraction. Another such example is the simple pendulum model. The dependence of its period T on its length l and the gravitational acceleration g does not depend on the microphysical details about the string, or nomological facts about gravity or friction. Since these details are not relevant to the explanation of the pendulum's T, Jansson & Saatsi (2019, p.822) have deemed these abstractions. However, I would argue that the simple pendulum model is an idealisation, for it misrepresents the bob as point-like and the string as massless and inextensible.[171]

Abstraction is pervasive in scientific explanations. Consider vitamin B12, also known as cobalamin.[172] Vitamin B12 is a water soluble vitamin involved in the formation of red blood cells and in the maintenance of the central nervous system. It has many forms, which differ with respect to the molecule they have attached to the cobalt ion; e.g. if it is a methyl group it is methylcobalamin, if it is a cyanide molecule it is cyanocobalamin.[173] Suppose we want to explain why some person is tired, the cause being a deficit of vitamin B12. Here certain structural features such as which molecule is attached to the cobalt ion won't make a difference to the vitamin's function (it is not a difference-maker). Since details about which molecule is attached to the cobalt ion are irrelevant for the metabolic explanation in question, one can abstract away from these. The resulting model for vitamin B12 is an abstract model precisely because it omits these details.

[170] There is disagreement as to whether abstraction concerns omission of *relevant* or *irrelevant* features of a system, where by 'relevant' I mean relevant to those aspects of the behaviour of the system which we seek to predict or explain. For instance, Levy (2021) argues that idealisations misrepresent *relevant features* of the modelled system whereas abstractions move away from *irrelevant features* of the system while keeping the relevant features intact.

[171] These two examples show how in practice the distinction between idealisation and abstraction is subtle; in fact, scientists are oblivious to the distinction and call models containing idealisations and/or abstractions 'ideal.'

[172] I owe this example to Francesca Bellazzi.

[173] There are others: hydroxocobalamin and adenosylcobalamin.

The fact that abstractions are detail-poor representations instead of misrepresentations is key to see why no analogous challenge to the idealisation challenge arises for the case of abstract models. Unlike idealised models, abstract models don't include false assumptions about their targets, so there is a straightforward sense in which they represent them. Counterfactuals with antecedents like 'If methylcobalamin were cobalamin, then...' don't show up in scientific explanations, for there is no doubt that cobalamin and methylcobalamin are not two distinct objects. Hence, abstraction does not pose any threats to *W-H*'s account of causal explanation.

4.5.2 An Objection: Abstraction as Distinctive of Non-Causal Explanation

There is a popular view of non-causal explanation, which has been defended in the context of mathematical explanations in science, under which an explanation is non-causal by virtue of being abstract. This is what Reutlinger & Andersen (2016) call the 'abstractness assumption'; Sullivan (2019) calls this view of abstraction the 'irrelevancy approach.' The irrelevancy approach has been most prominently supported by Batterman (2000, 2002, 2010), Batterman & Rice (2014), and Pincock (2012, 2015).

In the irrelevancy approach, the abstract nature of an explanation is inversely related to its causal content. As a matter of example, consider Pincock's (2015) abstract explanation of soap films. Soap films and bubbles follow Plateau's laws, a certain set of regularities such as the fact that they are smooth surfaces, or than when these smooth surfaces meet at points they always do so at equal angles, a consequence of the fact that soap films try to minimise their surface area. J. E. Taylor (1976) realised that the structure of soap films can be explained by the mathematical properties of almost minimal sets, i.e. 2D surfaces embedded in 3D Euclidean space $\mathbb{R}$ (Pincock 2015, p.861).

The explanation of soap films via almost minimal sets is an abstract explanation, Pincock argues, because it does not provide *fixed-object information*.[174] Therefore, it can't be causal: "[t]he value of the Taylor case lies instead in the other-object information that it provides" (Pincock 2015, p.868). Taylor's proof, he says, "... tells us, then, in what respects the class of soap films are similar to one another and also different from all those things that are not soap films. This does not tell us how any particular physical system will change under an intervention. So, there is no fixed-object information in the explanation" (ibid.). The abstract

[174] 'Same-object' information in Woodward's and Hitchcock's terminology.

explanation of soap films resembles another well-known non-causal explanation: RG explanation. Since this is a much more discussed example, let me now focus on RG explanation.

Consider an Ising model, a model used to represent the phase transitions physical systems undergo, such as the transition from ferromagnet to paramagnet a magnetic system experiments, in the absence of an external magnetic field, when it approaches the critical or Curie Temperature T_C and loses its bulk magnetisation. When $T \to T_C$, thermal agitation acts against the spontaneous alignment of the ferromagnet's dipole moments and increased temperature weakens the magnetic ordering. As a result, $M \to 0$: the magnetic alignment can no longer be sustained and the material becomes a paramagnet. In the Ising model, the polarisations of the metal's atomic dipoles are represented by variables that take the values ± 1 depending on whether the magnetic moments of atomic spins point up/down.

As I will shortly explain, in order to characterise the behaviour of a physical system at the critical point what we are interested in is the system's long-range behaviour. Therefore, we abstract away from the system's microphysical details. This long-range behaviour is captured by the *order parameters*, measures of the degree of order across the system's boundaries, such as magnets' net magnetisation M, which captures the bulk properties of physical systems undergoing phase transitions. The relation between the magnet's temperature T and M is characterised by the scaling law (Holovatch 2006, p.239)[175]:

$$M \propto |T_C - T|^{\beta} \tag{4.6}$$

The critical exponent β varies across magnetic systems. However, certain classes of magnets (uniaxial magnets like FeF_2 or MnF_2) share the same critical exponents. Furthermore, they coincide with those of simple fluids (Xe, SF_6, CO_2), among other types of systems like gases or superconductors. Hence, the Ising model does not only characterise magnets, but universality classes of systems with the same critical exponents (ibid.). Order parameters such as M approaching 0 are an indicator of a system's degree of disorder. The correlation length ξ, which characterises the range of fluctuations within a system (i.e. whether particles interact only with their neighbours or also particles which are further away) is a function of $T_C - T$:

$$\xi \propto |T_C - T|^{-\nu} \tag{4.7}$$

In eq. (4.7) ν is – like β – another critical exponent. When $T \to T_C$, $\xi \to \infty$. This

[175]Note that here T and M do no longer refer, respectively, to *target system* and *model system*, but to *temperature* and *magnetisation*.

divergence of ξ means that the system is so correlated that we cannot *see* anything. That's when RG theory comes into play. RG theory is basically an iterative scaling or 'coarse-graining' procedure from Mean Field Theory that decreases the degrees of freedom of the Hamiltonian of the physical system(s) in question, a coarse-grained procedure which is repeated until a 'fixed point' is achieved. In other words, it is a procedure that allows us to *zoom out* from the microphysical details.

Recall that the Hamiltonian is the function that represents the total energy of a system. The Hamiltonian we obtain once the coarse-graining procedure stops – let us call it 'fixed-point Hamiltonian' – characterises the different classes of physical systems with the same critical exponents, because they all *flow* to the same fixed points (Reutlinger 2016, pp.141-142). These classes of physical systems share the same behaviour, and therefore they constitute a universality class. Just like the soap film explanation, one important aspect of RG explanation is that it tells us in which respects the members of a universality class are similar to one another and different from non-members of the universality class. It is because they abstract away from microphysical details that, in the irrelevancy view, both soap film explanation and RG explanation are non-causal.

4.5.3 A Reply: Interventionism Beyond Causal Explanation

The irrelevancy approach to non-causal explanation has been challenged on the grounds that many explanations in the special sciences ignore lower-level details and are nonetheless *causal* (Cartwright 1989; Strevens 2008; Woodward 2003). In the metabolic explanation of why some person is tired (see §4.5.1), B12 may be cited as a cause, yet this explanation ignores details about the chemical structure.[176] *W-H*'s account of causal explanation definitely makes room for abstraction. In chapter 1, we saw how Woodward provides some heuristic criteria for variable choice; proportionality was put forth as a way to strike the right balance between sparsity and specificity. Proportional causal explanations are generally (though not necessarily) coarse-grained with respect to microphysical details. Moreover, in chapter 3 we saw how generalisations are causally explanatory provided they are invariant under ICs and BCs, including at least some interventions. Therefore, generalisations are explanatory even when they describe higher-level patterns of invariance. For the interventionist, abstraction is clearly not constitutive of non-causal explanatory power. But then, what constitutes non-causal explanatory power, if anything does?

[176]For challenges to the irrelevancy approach, see for example Reutlinger (2014), Reutlinger & Andersen (2016), or Sullivan (2019).

While interventionism was initially conceived as an account of causal explanation, with explanatoriness exclusively limited active counterfactual dependence, i.e. counterfactual dependence under interventions (so causation, see §4.2.2), in the last decade there has been a growing support for what Khalifa et al. (2020) have dubbed 'change-relating monism' (CR-monism): the view that counterfactual dependence underlies not just causal but also non-causal explanation.[177] What CR-monists propose is a unificatory counterfactual account of explanation, which broadly construes explanation as the provision of information about patterns of counterfactual dependence between explanans and explanandum, or in the catchy phrase in vogue in debates about scientific explanation, as answering 'what-if-things-had-been-different' questions (also known as w-questions) (Woodward 2003, p.221). According to CR-monism, non-causal explanation is the provision of counterfactual information where the counterfactual dependence in question is, e.g. mathematical or conceptual.[178] For instance, the dependence between Mother being unable to distribute 23 strawberries evenly among her three children without cutting any and 23 not being divisible by 3 is mathematical (Lange 2013, pp.488-489), as is the dependence between 13-cycle or 17-cycle cicadas and, respectively, the fact that 13 and 17 are prime numbers.[179] Woodward himself has entertained CR-monism in 'Some Varieties of Non-Causal Explanation,' where the dependence between Xantippe's widowhood and Socrates' death is discussed as a case of an explanatory conceptual dependence (Woodward 2018b, p.121).[180]

4.5.4 An Own Goal? Two Types of RG Explanation

While the examples of non-causal explanations discussed in §4.5.3 are underwritten by non-physical dependencies, it is not clear what type of dependence underwrites other presumed instances of non-causal explanation such as the explanation of the dependence between the stability of the orbits on spatial dimensionality or critical phenomena on spatial dimensionality. With respect to the explanation of the stability of the orbits, Woodward says that there are no physically possible interventions on the dimensionality of space, and all we can do is "construct(s) a model in which the dimensionality of the system is different and then calculate(s) the consequences" (Woodward 2018b, p.124). Regarding

[177] See Bokulich (2011), Saatsi & Pexton (2013), Rice (2015), and Reutlinger (2016), as cited in Woodward (2018b, pp.123-124).

[178] Note that this view is monistic with respect to what explanation amounts to, but pluralistic with respect to the kinds of dependencies that are explanatory.

[179] See 'Homopteran' for more information about periodical cicadas.

[180] This example can also be understood as mediated by a metaphysical dependence (see Emmerson 2021).

the RG explanation, he argues that "the dependence of critical phenomena on variables like spatial dimensionality does not describe physical transformations of the systems being analyzed, but rather transformations *in a more abstract space*" (ibid.).[181]

Via the above examples, Woodward contemplates the possibility of there being another sort of non-causal explanation, which is non-causal not because the dependence between explanans and explanandum is mathematical or conceptual, but rather because the manipulations entertained are physically impossible and, therefore, not interventions.[182] Does this second breed of non-causal explanations make Woodward fall back into the irrelevancy approach to non-causal explanation? In other words, does RG explanation – clearly a scientific explanation – challenge interventionism's core commitment to abstraction? I don't think so.

In the following, I first review Sullivan's criticisms of non-causalism about RG explanation to show how RG explanation is compatible with abstraction. In 'Universality Caused,' not only does she attack the abstractness assumption, but also deems misguided the assumption – most clearly defended by Reutlinger (2016) – that the explanatory power of universality is rooted in abstraction or the "abstract space of models" (Sullivan 2019, p.17).[183] It is in the context of her critique of Reutlinger's non-causal argument that the key to answering the question posed above lies: RG theory involves a type of explanation that's causal, and another type that's non-causal.[184]

Let us start by briefly sketching Sullivan's critique of the irrelevancy approach. As operators that represent the energy of the interactions between the components (or degrees of freedom) of systems, Hamiltonians clearly have physical content. When we apply RG theory to a physical system's Hamiltonian, this process washes away the microphysical interactions. The reduced Hamiltonian does not misrepresent any physical process, it only represents, in mathematical language, the causal behaviour of the physical process at the critical point in some less microphysical detail. In other words, the reduced Hamiltonians "... retain a

[181] Italics mine.

[182] Or lower-order physically impossible, according to the Extended Invariance View I introduced in chapter 3. I called these sorts of changes 'higher-order interventions'. In Woodward (2018b), he does not clarify what sort of dependence relation mediates this explanation; all he says is that it "... essentially involves assumptions that, however their status is understood, are not a priori mathematical truths," and continues by saying that "[i]t is hard to understand these assumptions except as empirical claims, even if it is unclear what empirical evidence might support them (ibid., p.126).

[183] Sullivan criticises a third assumption endorsed by some non-causalists, which I shall not discuss here: the assumption that there is no time-asymmetry of event types in RG explanations.

[184] Well, in fact there are two types of causal explanations that may be featured in the context of RG theory, but as it will become clear later in the section only the causal status of one of these is contested.

causal-tether mapping of the interactions between the degrees of freedom, spins, the effects of external fields, and so on" (Sullivan 2019, p.22). The explanation we end up with is a *single-system* RG explanation, which provides causal information about a particular physical system (and, I would add, any member of the class represented by the same order parameters). Nonetheless, supporters of the irrelevancy approach would agree with this verdict. The sort of explanation these non-causalists deem non-causal is what Sullivan calls *inter-system* RG explanation (ibid., p.20): why do different physical systems (or rather, types thereof) exhibit the same long-range behaviour near the critical point? Because they are members of a shared universality class: by reiteratedly coarse-graining their Hamiltonians via RG theory, they all flow to the same fixed-point, and as such are represented by the same fixed-point Hamiltonian. Since RG "flows do not correspond to any physical change of the system, but rather capture equivalence classes of models that share the same long-distance physics" (Saatsi & Reutlinger 2018, p.470), inter-system RG explanation is an even more abstract yet still causal explanation of the behaviour of a class of systems which belong to the same universality class. In Sullivan's words, "... there is nothing in principle special about an abstraction that crosses over several systems that precludes a causal interpretation" (Sullivan 2019, p.23).[185]

Next, Sullivan tackles the view that inter-system explanations draw their explanatory power from the abstract space of models (Saatsi & Reutlinger 2018). The assumption, she believes, is erroneous because once again it obliterates the fact that each of these models has a causal-tether. She invites us to consider the following counterfactual, which Reutlinger discusses:

> **ALT-RG:** There is a physical possible Hamiltonian H^* such that if (1) a physical system had the original Hamiltonian H^* (instead of its actual Hamiltonian H), (2) H^* were subject to repeated RG transformations, and (3) we determined the resulting flow of the Hamiltonians to a fixed

[185] There is certainly people who would object that the fact that an explanation cites the explanandum's causes does not mean that it works by virtue of describing these (Lange 2016, p.20). Lange (2013, 2016) or Andersen (2018) defend the view that if the explanandum-explanans relationship holds in virtue of mathematically necessary facts, the resulting explanation is a distinctively mathematical explanation, and therefore non-causal. Note that this view is distinct from Reutlinger's, who thinks (contingent) facts about the applicability of mathematical operations should also count as explanatory non-causal mathematical facts despite not being distinctively mathematical (Reutlinger 2014, p.1166). As an example of a distinctively mathematical explanation, Andersen discusses the case of the Lotka-Volterra model, which *holds of* systems in virtue of mathematical necessity: "[t]his is a kind of situating of a model in the network of causal relations in the world that gives explanatory leverage on the system as an explanandum, but in a way that is not causal, even though it is because of the causal structure of the system that the model holds of it" (Andersen 2018, p.495). I personally do not find this narrow view of causal explanation compelling, for as the previous quote shows, even their supporters acknowledge that it is the causal structure of systems what ensures that abstract causal models represent their systems.

> point, then a system with original H^* would be in a different universality class than a system with original Hamiltonian H. (Reutlinger 2016, p.743)

I will not analyse the intricacies of the present discussion, because what I find interesting about it is what it reveals: counterfactuals like ALT-RG are in fact tangentially related to inter-system RG explanation. Here is why.

The sort of explanation Reutlinger discusses is not an inter-system explanation, i.e. an explanation of the universal behaviour exhibited by different physical systems during a phase transition, where the explanans is the fact that all these systems all belong to the same universality class, and as such they all exhibit the same causal behaviour near criticality. A counterfactual like ALT-RG involves a w-question about what would happen to a physical system were it to belong to a different universality class than the one it currently belongs to, i.e. a different type of object, hence described by a different model. ALT-RG is a clear OO-CF, and hence non-explanatory by the interventionist's lights, for it falls outside of the scope of *W-H*'s account of causal explanation. If explanatory at all, ALT-RG would be *non-causally* explanatory, but not in virtue of it involving "...transformations *in a more abstract space*" (Woodward 2018b, p.124).

While changes that can be made sense in the context of a given universality class can be captured via SO-CFs, changes which consider how a system would behave were one to intervene on its spatial dimensionality are unintelligible to the interventionist. Therefore, counterfactuals like ALT-RG are at best tangentially related to counterfactuals one might encounter in the context of inter-system explanation; unlike the former, the latter might be given a causal interpretation.

4.5.5 Same-Object Counterfactuals and Abstract Models

Counterfactuals like ALT-RG (see §4.5.4) are OO-CFs because they describe how an object would behave were they a different object. As such, they are non-explanatory by interventionist standards. This is the expected result: the counterfactuals the interventionist deems explanatory are those supported by system-internal generalisations.

Note first that the result mentioned above is not analogous to the idealisation challenge: idealising counternomics consider how a physical system would behave were they *ideal*. We cannot turn krypton into an ideal gas because the ideal gas is a physically impossible object and, as such, the intervention in question is physically impossible. In contrast, ALT-RG considers a transformation of a physical object

into another physical object: the intervention in question is not impossible because it involves any ideal object but because it is so radical that it is not well-defined.

Inter-system RG explanations are peculiar abstract explanations because they draw on highly abstract models, which represent the common features of all the systems that fall under it. However, I hope to have shown that the fact that universality classes encompass heterogeneous classes of physical systems is no challenge to the causal nature of the resulting explanations. Both reduced and fixed-point Hamiltonians are system-internal generalisations which describe the (more or less heterogeneous) classes of systems the model in question applies to.[186] This aligns with the *solution* to the idealisation challenge: in scientific contexts, models typically describe classes of physical systems (see §4.3.3). Therefore, the objects of SO-CFs are typically classes of physical systems. Note that this is not typically but *always* the case for counterfactuals which appeal to abstract models: they are well-defined SO-CFs; as argued in §4.5.1, there is thus no abstraction challenge.

Nonetheless, given that CR-monism has been endorsed by many as a natural extension of *W-H*'s account of causal explanation beyond causation, let me first examine where CR-monism leaves the distinction between SO-CFs and OO-CFs on which their account relies. CR-monism rests on a separation of the two key elements of causal explanations – interventions and counterfactual dependence – and identifies the latter as key to explanatory power. The result is a pluralism about explanatory dependencies, all of which are nonetheless explanatory because they support change-relating counterfactuals (again, reason why CR is an explanatory monistic view). In the strawberries example, what explains the fact that Mother cannot equally divide 23 strawberries into 3 is the mathematical fact that 23 is not divisible by 3: what is doing the explanatory work here is a mathematical explanans-explanandum dependence. In explanatory contexts where the relevant dependencies are mathematical or conceptual, the laws vs domains for laws distinction (and with it the distinction between system-internal and system-external generalisations) is nonsensical. By abandoning the laws vs domains for laws distinction, CR-monism leaves us with the following picture: spectra of explanatory dependencies captured by change-relating counterfactuals. What the limits of such spectra are is a question that exceeds the scope of the present project, but should one want to endorse an explanatory monism of this kind, one might think these limits have to do with there not being sufficient or there being too much 'explanatory distance' between explananda and explanantia (see E. Taylor 2021).

[186] Note that expressions for the order parameters of second-order phase transitions such as eq. (4.6) or eq. (4.7) may also be seen as system-internal generalisations.

Regarding CR-monism, one last question that deserves further scrutiny is whether change-relating counterfactuals exhaust the class of explanatory counterfactuals. This is a question that has been discussed by Khalifa et al. (2020) in the context of RG explanation. They consider 'stability-exhibiting counterfactuals': counterfactuals that describe how microstructural changes to an object affect their macroscopic behaviour. In the context of a ferromagnetic-paramagnetic phase transition, one might, for instance, consider a change that "take(s) us from the Ising model to the Heinseberg model [different models of magnetic behaviour], or from Ni to Fe, or from a nearest neighbour interaction to a next nearest neighbour interaction" (Kadanoff 1971, p.105).[187] Note that this type of change is less radical than that described by ALT-RG, for it doesn't affect the system's behaviour near criticality: this is why the resulting counterfactual is a *stability-exhibiting* counterfactual.

Stability-exhibiting counterfactuals in RG explanation highlight two issues. The first is that the idea of explanatoriness as the provision of counterfactual information about the *same object* applies not only to change-relating counterfactuals but also to stability-exhibiting counterfactuals: note that they are SO-CFs with respect to the given universality class. From this, we could conclude that causal change-relating counterfactuals are a subset of the broader class of SO-CF which also comprises stability-exhibiting counterfactuals. This brings me to the second issue: whether – and, if so, in virtue of what – stability-exhibiting counterfactuals indeed constitute a distinct class of counterfactuals. Khalifa et al., who are pluralists about explanation, think they have a characteristic explanatory function of finding unity in diversity (Khalifa et al. 2020, pp.13-19). I made a similar proposal in chapter 3, where I argued that counterfactuals which consider how changes to some laws affect other laws – which may be seen as another class of stability-exhibiting counterfactuals – exhibit a distinctive constraining role upon explananda. By acknowledging a pluralism of explanatory functions, whether these physical dependencies are causally or non-causally explanatory will depend on their functions.[188] What the relevant explanatory functions are (e.g. providing counterfactual information exploitable for interventions, answering w-questions, finding unity in diversity, or constraining explananda) and whether – and, if so, how – these interact are questions for future research.

[187] As cited in Khalifa et al. (2020, p.15).

[188] The view I have in mind is one that endorses pluralism about explanatory functions, but monism about explanatory dependence: I take it that the interventionist project aims to provide information about an object's *physical* dependencies. One might think there are physical dependencies, whose explanatory function is set by context, which plays a partial role (at least) in determining whether the salient explanation is causal or non-causal.

4.6 Conclusion

In this chapter, I placed a magnifying glass over the same-object vs other-object distinction for counterfactuals central to *W-H*'s interventionist account of causal explanation. Interventionist counterfactuals are SO-CFs because they describe the causal dependencies internal to a system, and this is what makes them explanatory. Counterfactuals supported by external generalisations are OO-CFs, and hence non-explanatory. I claimed that this distinction faces two challenges: the individuation challenge, and the idealisation challenge. I also argued that there is no analogous challenge to the idealisation challenge for abstraction.

The individuation challenge is the challenge of defining what the subject-matters of these counterfactuals are. I argued that despite the fact that interventions are defined as tokens or particular instantiations of properties, it is typically not particular systems the interventionist is interested in. The aim of science is to explain phenomena, not individual events. This repeatability is the reason why the interventionist is focused on type-level causal claims, and why interventionism uses the formalism of causal modelling in the first place. For these reasons, I suggested that SO-CFs typically individuate *classes of physical systems* – this shows how interventionist explanation is not, after all, disconnected from questions of scope. In light of this solution to the individuation challenge, I argued that hypothetical interventions should be seen as hypothetical instantiations of properties on the hypothetical systems models describe, a characterisation which accommodates the modelling of non-actual targets.

The idealisation challenge is the challenge to account for the role of idealising counternomics if one takes *W-H*'s account of causal explanation as the right account of scientific explanation. According to such account, they are OO-CFs. However, that would render them – against our intuitions – non-explanatory. One solution to this challenge is that the interventionist more carefully accounts for the differing explanatory potential of OO-CFs. To this end, I argued that idealising counternomics are SO-enabling-CFs: OO-CFs with a distinct explanation-enabling function to enable causal explanations with idealised model.

Furthermore, I proposed an interventionist strategy to distinguish idealising counternomics from regular OO-CFs. Unfortunately, this interventionist strategy to discriminate explanation-enabling counterfactuals in the context of idealisations from those that are not is severely limited, because it is only applicable to idealised models which can be quasi-gradually de-idealised, such as the ideal gas model. This highlights how *W-H*'s account of causal explanation doesn't provide the

tools to evaluate a significant class of counterfactuals in science. Considering how pervasive idealisation is, this is a serious result.

Fortunately, I contended that abstraction poses no analogous challenge to the idealisation challenge. Abstraction does not involve false assumptions, so there won't be counternomics involving abstract models. Characterised as detail-poor representation, abstraction is perfectly compatible with *W-H*'s account of causal explanation: models in the special sciences are coarse-grained, because they track higher-level invariance relationships. Since this means abstraction is not constitutive of non-causal explanation, I explored how interventionism has been liberalised to account for non-causal explanatoriness via CR-monism.

Besides CR-monism, we have seen that Woodward (2018b) hinted at another strand of non-causal explanations which seem to suggest that abstraction is constitutive of non-causal explanation. Via the example of RG explanation, I showed that there is no such threat: inter-system RG explanations are unproblematically causal. The non-causal explanations he has in mind there do not draw their explanatory power from their abstractness. Abstract models do uncover the causal structure of their target systems; whether these target systems form a more or less homogeneous class does not erase the causal tether to the world from these models.

What this concrete example of of abstract explanation revealed is that counterfactuals come in many different varieties. For instance, in the context of RG explanation we find stability-exhibiting counterfactuals, which despite not being change-relating counterfactuals do provide information about the same object. This suggests that the label 'SO' might apply to a broader class of dependencies with different explanatory functions. What explanatory functions these may have, as well as how they may interact, are focal points of future research.

To conclude, the narrow picture regarding the explanatoriness of counterfactuals that interventionism paints does not reflect the diversity of counterfactuals in science. This is what prompts the unificatory proposal of the next chapter: a proposal based on fiction which gives meaning to both SO-CFs and SO-enabling-CFs and thereby captures the diverse counterfactual landscape of scientific explanations.

Chapter 5

Possibility via Fiction: Scientific Counterfactuals as Make-Believe

5.1 Introduction

My goal in this chapter is to positively motivate a practice-based analysis of counterfactual conditionals in science: what I call the *make-believe view of scientific counterfactuals*.[189] This is a fiction-based account, which exploits the connection between counterfactuals and fiction already explored by Kim & Maslen (2006), Kimpton-Nye (2020), McLoone (2019), and Wilson (2021). However, it differs from these accounts in some respects: the motivation for fictionalism about counterfactuals, its scope, and the morals drawn.

Part of my motivation is naturalistic; I want to give an account of counterfactual reasoning that reflects how scientists use counterfactuals in practice. In §5.2 I identify two key features of such practice-based characterisation of counterfactuals and counterfactual reasoning in science: scientific counterfactual reasoning is none other than model-based reasoning (§5.2.1), and counterfactuals in science are non-vacuous (§5.2.2). This leads to two desiderata for a practice-based account of scientific counterfactuals: it ought to capture their connection to models, and it ought to capture their non-vacuity. I shall call the former *CONNECTION-TO-MODELS*, and the latter *NON-VACUITY*.

Furthermore, an account of counterfactuals in science which takes scientific practice at face value is one that results from a pluralist enterprise. Despite the solid institutional walls between disciplines, the philosophical study of the meaning of counterfactuals shouldn't be disentangled from the cognitive study of counterfactual reasoning. Although this empirical turn in the philosophical analysis of counterfactuals is very recent, it is slowly gaining momentum. Woodward's latest book *Causation with a Human Face*, for instance, revolves around the central idea that "*causation* and *causal cognition* need to be understood together" (2021, p.1).

[189] This chapter has been recently published under the same name in a special issue of *Synthese* entitled 'Modal Modelling in Science: Modal Epistemology Meets Philosophy of Science.' See Iranzo-Ribera (2022).

There, he advocates a multidisciplinary approach to causation: "when it comes to causation, the armchair philosopher who appeals to intuitions, the experimental philosopher, and the empirical psychologist are in the same business, at least in many respects" (ibid., p.21). All three groups "should be understood as advancing empirical claims about patterns of causal judgement. The intuitions of the armchair philosopher have, in this respect, no special authority" (ibid.). In the same spirit, while philosophers and cognitive psychologists pursue different programs regarding counterfactuals, there is a big area of overlap across these which a comprehensive study of counterfactuals – and especially of counterfactuals *in science* – shouldn't overlook. This is why in §5.3 I review both the standard philosophical account of counterfactuals, namely the Lewis-Stalnaker similarity analysis (§5.3.1), as well as what I consider cognitive psychology's best theory of counterfactual reasoning – the mental models view – developed mainly by Johnson-Laird (1983); Johnson-Laird & Byrne (1991), Byrne & Johnson-Laird (2009) (§5.3.2). I consider each with respect to the desiderata of *CONNECTION-TO-MODELS* and *NON-VACUITY*. §5.3.3 brings these analyses together and identifies a connecting theme between the two desiderata: the (scientific) imagination.

§5.4 contextualises, presents, and gives examples of the *make-believe view of scientific counterfactuals*, which is my fiction-based approach to counterfactuals and counterfactual reasoning in science. By linking counterfactuals' antecedents to models, which are props in games of scientific make-believe, a clear connection can be established with *CONNECTION-TO-MODELS*. And by making counterfactual evaluation a matter of which fictional truths are propositions that are to be imagined, metaphysics doesn't get in the way of semantics, so *NON-VACUITY* is respected. §5.5 concludes.

5.2 Counterfactuals in Science

A counterfactual is a subjunctive conditional whose antecedent is assumed to be false, either because it *did not* occur or because it *could not* occur. We might call the former 'regular counterfactuals:' those counterfactuals with antecedents which describe unactualised possibilities, that is, states of affairs which were once possible but did not take place. "If the coffee milk had been preheated it would not have curdled" is one such example. In the second group we find what philosophers typically call 'counternomics' and 'counterpossibles.' What these have in common is that they all have impossible antecedents, that is, their antecedents describe states of affairs which could not have occurred. The name 'counternomics' – or

'counterlegals' – is reserved for those counterfactuals with nomologically impossible antecedents, that is, antecedents which describe states of affairs which the actual laws of nature don't permit. 'Counterpossibles' are counterfactuals with logically, mathematically, or metaphysically impossible antecedents. Hence, they can be respectively called 'counterlogicals,' 'countermathematicals,' or 'countermetaphysicals.'

The focus of this chapter is what I call 'scientific counterfactuals,' namely, counterfactuals that feature in scientific practice, that is, counterfactuals that are involved in explanations or predictions about the natural world. In the following sections, I identify their relevant features.

5.2.1 Counterfactuals and Scientific Models

Scientific reasoning is full of counterfactuals. Below there is a sample list of counterfactual claims we oftentimes encounter:

> **SEA LEVEL:** 'If concentrations of greenhouse gases had stabilised in the last decade, the North Sea level would have continued to rise.'
>
> **CORAL REEFS:** 'If atmospheric CO_2 had remained at preindustrial levels, there would be less bleaching of the world's coral reefs.' (List 2022, p.194)
>
> **IDEAL PENDULUM:** 'If pendulum X (an actual pendulum) were a simple pendulum, then for small swings its period T would only depend on its length l and the gravitational acceleration g.' (adapted from Godfrey-Smith 2009, p.168)
>
> **H-BONDS:** 'If water had not had intermolecular hydrogen bonding, then it would have been a gas at room temperature.' (adapted from Tan 2017, p.959)
>
> **BOHR ATOM:** 'If atoms were Bohr atoms, then an electron's angular momentum L in the ground state would have been observed at $L = \hbar$.' (Tan 2019, p.48)
>
> **SPACE TIME:** 'If space-time had had a dimension ≥ 4, the planetary orbits would have been unstable.' (adapted from Woodward 2003, p.220, Woodward 2018b, p.123)

Note that SEA LEVEL and CORAL REEFS are regular counterfactuals, while the rest are all either counternomics or counterpossibles. But what do all of these counterfactuals have in common? First, counterfactuals in science appeal to scientific models. That is, their antecedents make reference to scientific models – they are a conjunction of all of a model's assumptions – and their consequents are the predictions that follow from these. A minimal notion of 'model' is here adopted: models are small sets of propositions interpreted by some person to be about something. The 'predictions' captured by the consequents are those propositions entailed by the propositions that make up the model's description together with relevant information that is extraneous to the model in question. A more detailed characterisation of this connection between models and counterfactual reasoning will be given in §5.4.2.1.

The idea that the content of a model can be captured via a counterfactual conditional, where the antecedent is a conjunction of all of the model's assumptions and the consequent a conjunction of all the predictions that follow from these, is not new (see Plutynski 2006 Sugden 2009, Sober 2011, Godfrey-Smith 2020, and McLoone 2020). Examining which assumptions go into models brings up another typical feature of counterfactuals in science: many are counternomics or countermetaphysicals.[190] This is not surprising considering the fact that models usually contain idealisations, abstractions, approximations, analogies, etc. Idealising away from problematic features is necessary in order to make phenomena mathematically and computationally tractable, and abstracting away from explanatory irrelevant features is needed for explanatory success, which often requires striking a balance between generality and depth (see Strevens 2008). For instance, the simple pendulum model provides us with an equation which can be solved analytically under the assumption that the oscillations of the bob are small. The model incorporates other assumptions, all of which go against the laws of nature: the string is massless, the bob is a point mass, and there is no air resistance. IDEAL PENDULUM is thus a counternomic. So is BOHR ATOM: actual atoms are not described by Bohr's model of the atom, which – contra observation – renders the hydrogen atom unstable. Some would say that it is metaphysically impossible for water to have been bonded differently, or for space-time to have had a different dimension than the actual; on this basis both H-BONDS and SPACE TIME might be deemed counterpossibles.[191]

[190] One caveat: because counterlogicals and countermathematicals don't feature in scientific explanations and models, they fall outside of the current proposal. Hereafter whenever I refer to counterpossibles I will be referring exclusively to countermetaphysicals.

[191] Note that counternomics turn out to be countermetaphysicals for those who consider physical structure or laws of nature metaphysically necessary. This point will be discussed in detail in

5.2.2 Counterfactuals and Non-vacuous Truth

Some of the most influential philosophical views on counterfactuals endorse vacuism about counterpossibles, the view that counterpossibles are vacuously true, that is, true in virtue of the fact that since their antecedents are metaphysically impossible they are *necessarily* false. Here one notably finds the semantic analyses by Stalnaker (1968) and D. K. Lewis (1973), as well as Williamson (2007)'s. For future reference let us call vacuism about counterpossibles the 'vacuity thesis,' following McLoone (2020). Despite its popularity within philosophy, the vacuity thesis is at odds with scientific practice. Even when antecedents describe impossibilities, scientists attribute the corresponding counterfactuals non-trivial truth values: they judge these to convey information on the basis of their truth or falsity. In addition to recent philosophical support (Jenny 2018, Tan 2019, and McLoone 2020), this 'non-vacuity thesis' has also received support from the first experimental study on counterpossible reasoning in science: McLoone et al. (2023) presented a group of self-described biologists with two counterpossible formulations, and asked them to judge whether these were true/false and why. Said judgements were found to be independent of the counterfactual antecedents' (im)possibility.[192] Since these results are important evidence for a naturalistic account of counterfactual reasoning, I will briefly recall the experiment they conducted. In said experiment, participants were presented with the logistic equation of population growth, and told that it would be used to model a population of real rabbits. The logistic equation is:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \tag{5.1}$$

Equation (5.1) is a differential equation that captures the evolution of population size N considering its intrinsic growth rate r and the habitat's carrying capacity K. Rabbits are discrete, but eq. (5.1) assumes them to be continuous. In view of this features of the logistic equation, the participants were asked to look at the following two counterpossibles (from McLoone 2020, p.12161), which for future reference I have labelled 'SURVIVAL' and 'EXTINCTION':

> **SURVIVAL:** If some population of rabbits satisfied the assumptions of the logistic equation, then the size of the population (N) would

§5.3.1.

[192] This study counters Dohrn's objection that the claims of philosophers like Jenny, Tan, and McLoone are not to be taken seriously "because the reconstruction of scientific practice by philosophers is just another piece of philosophical theory and should be treated accordingly." Dohrn (2021, pp.118-119).

eventually be equal to the carrying capacity (K).

EXTINCTION: If some population of rabbits satisfied the assumptions of the logistic equation, then the population would eventually go extinct.

Participants were first briefed about the differences between nomological and metaphysical impossibility, and asked whether the assumption that rabbits come in non-integer values was metaphysically impossible. After that, they had to assign truth-values to these two counterfactuals and justify their answers. The results show that a significant majority of participants judged SURVIVAL to be true and EXTINCTION to be false, and relied on the mathematical relation between the antecedent and the consequent to do so regardless of what they thought about the (im)possibility of their shared antecedent. It is precisely the fact that the (impossible) modal status of the states of affairs described by these counterfactuals' antecedents is left out from counterfactual evaluation that ensures that all the obtained truths are non-vacuous. Clearly more experimental studies on the non-vacuity of counterfactuals with impossible antecedents are needed, but even this first study puts extra pressure on the vacuity thesis.

In short, in this section I have explored the scientific counterfactual space and introduced some assumptions regarding its characteristics: counterfactuals in science are connected to models and they are non-vacuous, especially those with impossible antecedents, which comprise a big subset thereof.

5.3 The Toolbox for Counterfactual Analysis

5.3.1 A Philosopher's Take: The Similarity Analysis

The meaning of counterfactual conditionals is often described using the similarity analyses developed almost simultaneously by Stalnaker (1968) and D. K. Lewis (1973).[193] Their accounts construe counterfactual conditionals as factual, categorical statements, which are analysed via possible-worlds semantics, a framework

[193] Even though the similarity approach has been traditionally attributed to these two philosophers, it was developed contemporaneously by a few people (see Starr 2021, section 2.3). Note that the Lewis-Stalnaker possible-worlds-based semantic analysis is not the only philosophical analysis of counterfactuals on the table. Another well-known analysis is Edgington's suppositional theory of conditionals (2003; 2008), a Ramsey-inspired view which builds upon Ernest Adams' logic of conditionals (Adams 1965). Since the counterpossibles literature has engaged only with the former – for under the latter the vacuity thesis does not obtain – I'll here focus on the former as well.

which models different modal notions (necessity, possibility, and counterfactuality) on the basis of truth at possible worlds, where possible worlds are usually understood as a convenient way of speaking about how the actual world could have been.[194]

'If I had struck this match, it would have lit.' Intuitively, this counterfactual is true. However, had I struck this match in a room without oxygen, on a wet surface, or not energetically enough – so that there was not enough friction between the match head and the surface – this would not be so. Stalnaker and Lewis agreed that the context is here important: in order to assess the meaning of a counterfactual A $\Box\!\!\rightarrow$ C we should look at just the relevant possible world (Stalnaker) or set of possible worlds (Lewis), which are those which are *closest* or *most similar* to the world at which the counterfactual under scrutiny is evaluated. Formally, A $\Box\!\!\rightarrow$ C is true iff the counterfactual's consequent (C) is true at every closest possible world at which the antecedent (A) is true.

What the notion of similarity precisely amounts to has been the focus of debates about the suitability of the similarity analysis for counterfactual evaluation. In the following I focus on Lewis' account, for he proposed a system to weigh similarities between worlds against their differences in order to arrive at a notion of overall comparative similarity between those worlds (see D. Lewis 1979, p.472). As a matter of example, the first of these weights tells us to avoid big, widespread, diverse violations of law. The second recommends that one maximises the spatiotemporal region of perfect match of particular fact. And so on. Lewis' similarity analysis has, with good reason, pervaded the philosophical psyche: even if sketchy, it provides a recipe to evaluate counterfactual possibilities in terms of *distance* to the actual world. However, it falls short of resources for a naturalised view of scientific counterfactuals like the one I am after. Here's why.

Regarding *CONNECTION-TO-MODELS*, models are *not* like possible worlds. Possible worlds are complete, which means that in possible-worlds-based semantics all propositions are determinate with respect to truth. However, models are not complete: intuitively, models only enable the assignment of determinate truth-values to the relevant propositions, namely, those that fall within the scope of the model. Those that fall outside of a model's scope are indeterminate, for the model does not include the facts that would make these propositions true or false (Currie 1990; Salis & Frigg 2020; Walton 1990).[195] Hence, possible-worlds-based semantics for counterfactuals do not naturally apply to scientific models.

[194] This is usually the case since Kripke (1980), with only modal realists like Lewis himself envisioning possible worlds in more ontologically substantive terms.

[195] Walton calls these propositions 'silly questions' (Walton 1990, pp.174-182).

In relation to *NON-VACUITY*, the Lewis-Stalnaker analysis of counterfactuals leads to the vacuity thesis: if a counterfactual's antecedent (A) is *metaphysically* impossible, there is no possible world which satisfies A – there is no A-world – and so no closest A-world. The counterfactual is then trivially true. Why is the vacuity thesis problematic? First of all, trivially true counterfactuals are uninformative. In the Lewis-Stalnaker semantics counterpossibles turn out true in virtue of the fact that their antecedents are impossible, and so any consequent is logically implied. Therefore, the consequents of counterpossibles add nothing to their truth-value or meaning (Brogaard & Salerno 2013, p.642). For instance, 'If wizards were green, they would photosynthesise' is true because there are no wizards, and not because "it describes a substantive fact about what the wizards are like or which properties they have" (Tan 2019, p.34).[196] Second, and more importantly, the vacuity thesis contravenes scientific practice, which as seen in §5.2 renders some counterpossibles non-trivially true and others false.[197]

One way to handle the vacuity of counterpossibles would be to reject those metaphysical views under which the vacuity thesis obtains as implausible or fringe. However, this would be an ad hoc move. Furthermore, the low prevalence of the vacuity thesis is challengeable, for the vacuity thesis obtains under a spectrum of modal views. These include, on the one hand, philosophers who think that these counterfactuals' antecedents violate the essential nature of non-fundamental kinds. Tan (2019, p.46), for instance, argues that it is a matter of metaphysical necessity that water is H_2O and so it is composed of discrete molecules. As a result, any scenario where water's microstructure is different than its actual one is metaphysically impossible. On the other hand, the vacuity thesis concerns those philosophers for whom the counterfactuals in §5.2 come out as metaphysically impossible due to the necessity of the laws of nature. Here we find dispositional essentialists, for whom the laws are metaphysically necessary contingent on the instantiation of the relevant natural properties or kinds, as well as modal necessitarians, for whom the laws of the actual world are the laws of all the possible worlds. Due to the fact that realising the antecedents of the counterfactuals in §5.2 would involve the violation of one or more laws of nature, these are counternomics. Given that the laws of nature are (conditionally or unconditionally) metaphysically necessary,

[196] Dohrn (2021, p.719) questions the link between informativeness and non-vacuity. He argues that informativeness is tied to counterfactuals being embedded into the right argumentative context, following Williamson (2007, p.173) and Jenkins (2010, p.258). However, so far it looks like the view that vacuity should be preserved is minoritarian.

[197] While I have here focused on the similarity analysis, the criticisms of the vacuity thesis are applicable to other factualist views of counterfactuals. See McLoone (2020) for a critical engagement with Williamson's defence of the vacuity thesis.

these turn out to be counterpossibles as well.[198]

Recently, other responses to the vacuity thesis have been offered, which either extend Stalnaker-Lewis semantics with impossible worlds (see McLoone 2020), or repurpose possible worlds using modal counterpart theory (see Hicks 2022). While these meet the demands of *NON-VACUITY*, McLoone (2020, pp.18-20) himself acknowledges his strategy is not precisely parsimonious, which is something which could also be said of Hicks' account. Whether and, if so, how they meet *CONNECTION-TO-MODELS* would have to be analysed. I believe a third overlooked strategy, which meets both desiderata in a more straightforward way, is up for grabs.

5.3.2 A Cognitive Psychologist's Take: The Mental Models Theory

In cognitive science, the term 'possibility' is often used in a broader sense than it is used in philosophy. For instance, for Nichols & Stich (2003, p.25) 'possibility' designates any description which doesn't include any obvious logical contradiction. What they call the 'Possible World Box' (PWB) is a cognitive workspace which basically represents what the world would be like given some assumptions whose truth-value is unknown to us. The PWB is one of the cognitive mechanisms that underlie the capacity for pretence, which has been argued to be implicated in counterfactual reasoning (Goldman 1992).

Counterfactuals not only have been and continue to be a controversial topic for philosophers, but also for cognitive scientists. Research in cognitive psychology has given rise to different theories about the cognitive process(es) that underlie the production and evaluation of counterfactuals. Alongside the view that counterfactual reasoning is a matter of calculating conditional probabilities, one also finds the view that human reasoning proceeds by means of simulating 'mental models.' Nowadays we call 'mental models' accounts those based on this idea (see e.g. Johnson-Laird 1983, Johnson-Laird & Byrne 1991, and Byrne & Johnson-Laird 2009). In §5.3.3 I argue that they offer valuable insights with respect to counterfactual reasoning in science.

[198] Opinions are divided with respect to whether and how necessitarian accounts of laws can provide a realist semantics for counternomics. Handfield (2004) has defended that the dispositional essentialist can evade the vacuity thesis with the help of 2D semantics: e.g. while it is metaphysically necessary that water is a discrete fluid, it is epistemically or conceptually possible that it is continuous. Kimpton-Nye (2020) agrees with Handfield, but argues that his strategy can unfortunately not help the modal necessitarian. Tan (2019, p.46) believes that differently-structured water – continuous and incompressible water – could not fill the functional and theoretical roles that actual water fills.

According to the proponents of mental models accounts, people understand conditionals by mentally simulating (im)possibilities (Kahneman et al. 1982). Counterfactual reasoning is thus none other than thinking about possibilities in the loose sense of the aforementioned PWB (see Byrne 2016, p.342), represented as small mental models of the situation. In everyday counterfactual reasoning people usually represent dual possibilities, reality and an alternative to it (see Byrne 2005, 2007). This idea that when evaluating counterfactuals what people do is imagine situations in which these counterfactuals' antecedents hold and then run mental simulations forward to see what happens in these imagined situations – leaving much else unchanged – is broadly in line with Lewis' account. In fact, in *The Rational Imagination* Byrne (2005) shows that people tend to imagine worlds with the same natural laws, a result which is captured by Lewis' first criterion of similarity of worlds (see §5.3.1).[199] The idea that mental models are *imagined* counterfactual situations fits nicely with *CONNECTION-TO-MODELS*: in science, models are often understood as imagined systems – imagined economies, populations, gases, etc.[200]

What about *NON-VACUITY*? In mental models theory, whether a counterfactual is true is a matter of whether a certain consequent follows in our mental simulations of the situation in question. For more details on the theory's account of the meanings of conditionals see Espino et al. (2020, pp.1265-1266). What is of main interest here is the fact that this theory does not rely on the notion of conditional probability, and so the content of counterfactuals, some of which may have impossible antecedents, can be represented just as well as the content of indicative conditionals.[201] In fact, in recent work Byrne (2022) concludes that mental models can be straightforwardly extended to counterpossibles: "[p]eople simulate impossible conjectures as if they were possible, relying on knowledge of reality to constrain their interpretation of them as true or false" (p.19). This is because modal judgements about antecedents, i.e. judgements about how probable or similar to actuality a possibility is, do not preclude their mental simulation and evalu-

[199] Byrne also shows that people tend to imagine alternative scenarios to more recent events rather than earlier events, as well as to controllable rather than uncontrollable events.

[200] Following the semantic view of theories models might also be understood as mathematical structures to which an interpretation is then given, but this is not the sense of 'model' with which I am concerned here; models *as* imagined systems are already representations in themselves.

[201] One of the problems of using conditional probability to assess conditionals is that if an antecedent A is impossible it has null probability $Pr(A) = 0$, so the conditional probability $Pr(B/A) = P(A \cup B)/Pr(A)$ is undefined.
Despite this advantage over probabilistic accounts, J. B. T. Evans & Over (2004, p.124) consider the fact that mental models accommodate counterfactuals as well as indicative conditionals a bug rather than a feature; they argue that mental models have too little representational content, as they cannot represent whether a possibility is very close or very distant to actuality.

ation of the resulting counterfactuals.[202] This result is in line with the research outputs of the experiments undertaken by McLoone et al. (2023), who as seen in §5.2 have given some evidence that reasoners' judgements of the truth-values of counterfactuals do not bear on their judgements about the (im)possibility of the counterfactuals' antecedents.

In short, cognitive research in the mental models theory of counterfactual reasoning highlights the central role of the imagination, which is often acknowledged by philosophers of science working on modelling. In the mental simulation of the scenarios described in counterfactuals' antecedents, their modal status plays no role, so the vacuity thesis can be sidestepped.

5.3.3 Assembling the Pieces: Scientific Models and the Imagination

In §5.1 I identified two desiderata for a naturalised view of counterfactuals and counterfactual reasoning in science. The first, *CONNECTION-TO-MODELS*, tells us that it ought to respect the key role counterfactuals play in scientific modelling. The second, *NON-VACUITY*, is self-explanatory: most counterfactuals are either non-trivially true, or false.

With respect to *CONNECTION-TO-MODELS* both armchair and experimental philosophers of science agree on the connection between counterfactual reasoning in science and models, which have a pivotal and widespread use in science. In §5.3.1 I argued that models are *not* like possible worlds, for models are incomplete. How should we then conceptualise them? Looking at cognitive science and actual scientific practice, models look more like *imagined* scenarios which the antecedents of counterfactuals like those in §5.2 invite us to consider. Philosophers working on modelling in science have long emphasised the involvement of the imagination in modelling; models have been described as e.g. imagined objects (Frigg 2010a) or imagined images of the world (Morgan 2004). Byrne (2005) has argued, contra folk belief, that the imagination is a *rational* cognitive faculty which allows reasoners to generally correctly evaluate the connections between counterfactuals' antecedents and their consequents.[203] This idea that "many people are rational, in the sense

[202] Closeness to actuality does however constrain the evaluation of their truth-values. In Byrne (2022) she shows that these evaluations differ depending on whether consequents are same/opposite to reality, or their correspondence to reality is unknown.

[203] Otherwise, Byrne (2005) argues, its usefulness and reliability would be somewhat mysterious. Her argument for the rationality of the imagination is more elaborate than that, but unfortunately I don't have the space to discuss it here. It rests on three separate assumptions: i) the imagination is rational, ii) human reasoning depends on the imagination of possibilities, and iii) counterfactual thoughts rely on thinking about possibilities.

that they make normatively appropriate causal inferences and judgements much of the time" has also been discussed by Woodward (2021, p.5). Having identified the imagination as key in counterfactual reasoning, we now need to fine-tune what type of imaginative process counterfactual reasoning in science is. Byrne (2005) leaves the characterisation of imagination quite open for day-to-day counterfactual reasoning: perception, or something like our comprehension of discourse, might be involved in mental modelling. Given that our focus here is on counterfactuals, which are expressed as propositions, I believe a *propositional* notion of imagination is the obvious candidate: we want to be able to speak of the content of counterfactuals as true or false. This choice is reinforced by the fact that counterfactuals in science, which often trade in impossible hypothetical scenarios, may be difficult to visualise. In this case, only a propositional type of imagination is able to render these intelligible.

Regarding *NON-VACUITY*, in §5.2 we have seen that in recent years some theorising against the philosophical orthodoxy of the vacuity thesis has been put forth by scientifically-minded philosophers. The idea that all counterfactuals with impossible antecedents may have non-trivial truth-values has now received empirical support from the research conducted by McLoone et al. (2023). It is precisely because of the non-triviality of their antecedents that informative consequents can be inferred from these. This non-vacuity results from abandoning the unfruitful idea that what makes a certain counterfactual true are some objective relations between matters of fact at some possible world which stands in some similarity relation to the actual world.[204] Rather, what makes a counterfactual true is a matter of whether a counterfactual's consequent (C) follows from its antecedent (A). This basic mechanism, which underlies the mental models account of conditionals (§5.3.2), can equally accommodate all types of counterfactuals: regular counterfactuals, counternomics, and countermetaphysicals. The task at hand is to make precise what 'follow' means here; I will do so in §5.4.2.1.

To sum up, a good account of counterfactuals and counterfactual cognition in science ought to pay attention to the connection with modelling practice, which emphasises the key role of the imagination – understood propositionally – in counterfactual evaluation. Furthermore, it ought to accommodate the empirical results which show that counterfactuals are evaluated on the basis of whether their consequents follow from their antecedents, an evaluation which is silent to the modal status of the states of affairs described by their antecedents. Luckily, there is one such account available, which I call 'the make-believe view of scientific

[204] Recall that in Lewis' framework, 'actual' is indexical; which world is considered as actual is designated by the context in which a counterfactual is uttered.

counterfactuals.' This is a fiction-based account of counterfactuals, which closes the gap between traditional philosophical theories of counterfactuals and scientific practice, and reclaims counterfactuals as a rightful object of study for philosophers of science.

5.4 Fictionalism About Scientific Counterfactuals

5.4.1 Contextualising Fiction

On the one hand, counterfactuals feature prominently in scientific models. On the other, counterfactual reasoning is an imaginative activity. When put together, these two observations suggest a view of models as "works of fiction" (Cartwright 1983, p.153). The idea that the notion of *fiction* might be relevant to the practice of model-building and scientific representation with models is not new. It was explored in a series of essays edited by Suárez (2009), as well as by those who advocate any version of what is nowadays known as *the fiction view of models*. In the following I extend this view of scientific models to counterfactuals in science, and thereby propose a fiction view of counterfactual reasoning in science, the make-believe view of scientific counterfactuals. While I am not the first to propose a fictionalist view of counterfactuals (see e.g. Kim & Maslen 2006, Kimpton-Nye 2020, McLoone 2019, and Wilson 2021), my account offers new insights with respect to the motivation, scope, and implications of fictionalism about counterfactuals in science. First, I introduce both the notion of fiction it uses (§5.4.1.1), as well as the fiction view of models it draws on (§5.4.1.2). Then, I develop the make-believe view of scientific counterfactuals and discuss in which respects it improves on the aforementioned literature (§5.4.2).

5.4.1.1 Fiction as Imagination: Make-believe

Prima facie, the word 'fiction' usually connotes *falsity*. The word has been ordinarily used to refer to a description which involves falsity or inaccuracy, and which is known to involve it by those who employ it. If the fiction is a proposition, it is understood as a deviation from truth, if it is an object, it signals a deviation from existence. The fact that fictions in this sense are introduced 'deliberately' or 'purposefully' is what distinguishes them from mistakes.[205] This view of fiction as counterfeit, forgery, or fake has received many names, such as

[205]Teller (2009) argues that while they are ontologically on a par, they are epistemically different categories.

the *truth-conditional* sense of fiction (Suárez 2009), or fiction as *infidelity* (Frigg & Nguyen 2020, pp.105-113).

However, this is not the only meaning of the word 'fiction.' Fiction also connotes *imaginings*, regardless of the accuracy of their content. This second sense of fiction, called *functional* (Suárez 2009), or fiction as *imagination*, is the one I henceforth use. In particular, I borrow Kendall Walton's (1990) understanding of fiction in his 'pretense theory,' the systematic study of what he calls 'games of make-believe.'

Make-believe is "the use of (external) props in imaginative activities" (Walton 1990, p.67). The notion of prop must be understood loosely; anything which prompts one to imagine, ranging from stumps to paintings or texts, might qualify as a prop. Now, what sorts of imaginings are mandated by props is constrained by 'principles of generation,' the rules of what is to be imagined. These are largely ambiguous, and range from tacit knowledge and belief to logical rules of inference or mathematical knowledge. "[W]hat principles of generation there are depends on which ones people accept in various contexts" (Walton 1990, p.38), so make-believe is essentially a social activity.

As previously noted, props prescribe an attitude of "imaginative engagement" (Frigg 2010a, p.110): when we say 'It is fictional that p' what we mean is that 'It is to be imagined that p.' What makes it fictional is "the attitude that the reader is expected to adopt towards it" (ibid.). I follow Currie (1990, p.20), McLoone (2019), and Salis & Frigg (2020) in characterising this attitude of imaginative engagement as *propositional*.[206] In the game of make-believe triggered by Kafka's *The Metamorphosis*, for instance, the reader is mandated to imagine p = *that Gregor Samsa finds himself transformed into a gigantic insect.* Propositions like p have the property of being 'fictional.' If a proposition is fictional with respect to a certain game of make-believe, we say it is a 'fictional truth' in that game of make-believe. Hence, p is fictionally true (f-true) in the game of make-believe prescribed by Kafka's *The Metamorphosis*. Truths directly generated by the prop

[206] Walton (1990) argued that the sort of imagination required for make-believe is imagining *de se*, where it is described as a form of imagining where one imagines herself "*doing* or *experiencing* something (or *being* a certain way)," i.e. where one is part of the content of what is imagined (p.29). While imaginings de se and objectual imaginings are compatible with make-believe, given that the focus of this chapter is scientific counterfactual reasoning and representation, it is make-believe's construal as essentially a propositional attitude that makes the application interesting, for it allows us to speak of truth/falsity within games of make-believe. This is not to say that propositional and imagistic imagination may not come together, but as Salis & Frigg (2020) argue, it is the propositional imagination which characterises scientific models and, by extension, as I'll argue, counterfactuals. Murphy (2020) has attacked this monist characterisation of the scientific imagination by mainly pointing out that thought experiments appeal to different varieties of the imagination. Her criticism thus does not at face value affect the claim that the variety of the imagination which characterises models and counterfactuals is propositional.

are called 'primary,' those indirectly generated by the prop and some rules of inference are called 'implied' (Walton 1990, pp.140-144). Importantly, imagining p does not commit one to its truth, even though make-believe is not incompatible with it. Therefore, truth-in-fiction is *not* a species of truth, even though fictional truths might ground truths.[207]

5.4.1.2 The Fiction View of Models

Scientists often reason about imaginary biological populations, neural networks, economies, or pendula, which are ordinarily used as stand-ins for the chunks of the world we investigate. The 'fiction view of models' is an umbrella term for those views of models which compare these imaginary objects to those of literary fictions, an analogy originally formulated by Godfrey-Smith in his paper 'The strategy of model-based science' (2006). I find this analogy helpful, for the reasons I now expose.

The main contemporary versions of the fiction view of models recognise the central role of the imagination in modelling practices, and build on Walton's notion of make-believe as a social imaginative activity (Salis 2020, p.2). Its mechanics are the following: model descriptions are seen as props, which prescribe imaginings about either model systems or directly about physical systems, i.e. target systems'. The former view is called the 'indirect fiction view of models' because models represent their target systems indirectly, via the mediation of model systems (see e.g. Godfrey-Smith 2009, or Frigg 2010a; Frigg & Nguyen 2020).[208] A supporter of this view will e.g. defend that the ideal gas model prescribes imaginings about a model system, the ideal gas, which is taken to represent a target system, a real gas under normal conditions of pressure and temperature. The latter view is analogously called the 'direct fiction view of models' because the model description directly prescribes imaginings about a target system without the mediation of a model system (see e.g. Toon 2012 or Levy 2015). Back to the gas example, a supporter of the direct view will defend that the model description of the ideal gas prescribes imaginings of a real gas described somewhat differently from the way it really is.

Both fiction accounts of models face difficulties. The main hindrance of the indirect fiction view is the problem of transfictional propositions, namely, the problem of cashing out model-target comparisons on deflated views of models (see e.g. Godfrey-Smith 2009, 2020 or Levy 2015 for expositions of the objection).

[207] See §5.4.2 for an explanation of how fictional truth might map onto truth.

[208] Note that Godfrey-Smith holds an indirect fiction view of models but rejects make-believe; he criticises Frigg's account on the basis that comparisons between the uninstantiated properties of model systems and the real properties of target systems are problematic (Godfrey-Smith 2009, p.113, 2020, p.161)

In fact, the direct fiction view was proposed as a reaction to the problem of transfictional statements: by eliminating the intermediate object which mediates the representation of the target system said problem does not arise. However, the direct fiction view of models faces a challenge of a different sort: it cannot accommodate targetless models – such as multi-sex populations (Weisberg 2013, ch.7), the ϕ^4 in quantum field theory (Frigg 2010a), or Schelling's famous segregation model (Schelling 1971) – nor what Weisberg calls 'generalised' models (models of general phenomena, such as parasitism) and 'hypothetical' models (models of nonexistent targets, such as xDNA). In Godfrey-Smith's words, direct fiction views "have to massage the phenomena quite a lot" (2009, p.159). Toon's (2012) response to these cases is to treat them as exceptions: model descriptions prescribe imaginings of target systems unless the models have no target, in which case they prescribe imaginings about fictional systems. However, this latter option inherits the problem of transfictional statements for targetless models. Levy's (2015) response to the objection is to dismiss these cases as either only apparently targetless or as bits of mathematics. This poses an immediate question: under the assumption that the target will eventually be clarified, what is the model *about* until then? Consider the following counterfactual: 'If neutralinos existed, they would only interact with the intermediate vector bosons.' We don't know whether these hypothetical particles of supersymmetry models will eventually be found, but we nonetheless treat models of these as entities in their own right. The main difficulty of the direct fiction view is the fact that it makes modelling parasitic on what is already known to exist. By making the definition of 'model' independent from modelers' epistemic states, the indirect fiction view comes out as a more inclusive alternative.[209]

The aforementioned inclusivity results from the workings of the indirect fiction view: the indirect view cashes out modelling as a two-step process, where in the first step one investigates the model, and in the second one investigates whether the model bears on the world (and if so, what the relevant representational relation is). This two-step fashion which characterises the indirect fiction view explains why the two notions of fiction – fiction as falsity and fiction as imagination – need not be mutually exclusive: fictions might still represent beyond themselves, but crucially, it is in virtue of their imaginative nature and not

[209] Another viable inclusive alternative is to reconceptualise the indirect view of models into the 'new fiction view of models' (Salis 2020), a view which in order to account for the relationship between the properties attributed to imaginary objects by the model and real objects abandons the notion of a model system and appeals in favour of the notion of a 'mental file'. Mental files are cognitive structures used by anti-realists about intentionality for the storage of information about an object, which may or may not exist.

this representational power that they are fictions. Translated to truth, this is the idea that when modelling one first *pretends* the truth of the claims implied by the model, and only afterwards considers whether these fictional truths map onto truths via one's preferred theory of scientific representation, such as Frigg's and Nguyen's DEKI account of scientific representation (2016; 2020). Asking about the (target)representational power of a model is then a question that is bracketed when investigating a model. I believe the characterisation of modelling as a two-step activity offers a good and unifying reconstruction of counterfactual reasoning in science.[210] When reasoning with counterpossibles, one often first asks whether a consequent follows from the antecedent in question before examining how the content of the imaginary exercise in question relates to the world. Given that there don't seem to be any relevant cognitive differences between regular counterfactual reasoning and counterpossible reasoning (see §5.3.2), it is only natural to characterise regular counterfactuals in the same way. This independence from knowledge of actual states of affairs afforded by the indirect fiction view of models thus enables a uniform treatment of counterfactuals, which the direct fiction view can't offer. This is the main reason why I favour an indirect fiction view of models.

One key feature of the fiction view of models is that it brings imagination to the foreground. The 'imaginary' is a category which signals merely possible objects which have a constructed nature, so it captures the fact that even if models often describe mind-independent physical patterns, models are themselves *created* by scientists. Hence, and unlike abstract mathematical objects, models are candidates for concreteness, namely, "candidates for entering into causal relations with real objects." (Godfrey-Smith 2020, p.159). Furthermore, the type of imagination involved in make-believe is not "an unbridled flow of free thoughts" (Frigg & Nguyen 2020, p.120). One worry here is that the central role of the imagination in modelling can be stressed without reference to fiction.[211] Weisberg (2013), for instance, has an account that recognises the role of the imagination in modelling, yet at a high price: imaginary systems are treated as integral to the practice of modelling, but are not to be identified with the models themselves. Imaginary systems are thus part of the folk ontology of modelling, and this is because he conceives of the imagination as primarily *imagistic*. But how can folk ontology be essential? I believe that this can be easily remedied by turning to a characterisation of the key imaginative engagement with models as essentially *propositional* instead

[210] As Levy (2015) argues, while positing and analysing a model, on the one hand, and assessing its relation to the actual world, on the other, might not come apart in actual practice, they are logically distinct operations. The indirect fiction view recognises the logical distinctness of these two operations.

[211] I would like to thank an anonymous reviewer for raising this concern.

of imagistic – but this is none other than what the fiction view of models proposes.

In the fiction view of models, model descriptions prescribe games of make-believe regimented by props and principles of generation, which in science are mathematical knowledge, laws of nature, and logical rules of inference. In short, reality-oriented principles and inferential tools. In make-believe there is no general recipe of what constitutes a principle of generation beyond the expectation that something like Walton's 'reality principle' (Walton 1990, pp.144-150) – a principle which used in the context of modelling would instruct us to import actual theoretical information unless otherwise specified by the model's description – or the 'mutual belief principle' (ibid., pp.150-161) – under which the implied truths generated are the result of beliefs held by the scientific community – obtains.[212] While principles of generation might arguably be perceived as vague, and so epistemically idle, this is only so because of their highly contextual form.

In the context of scientific inquiry, it is not only the content of games of make-believe that is constrained – or rather, guided – but also their scope: make-believe within the fiction view of models excludes modelling for merely recreational purposes (Godfrey-Smith 2020, p.160); scientific make-believe has epistemic aims such as explanation, justification, the drive of empirical discovery, or the enhancement of understanding.

5.4.2 The Make-Believe View of Scientific Counterfactuals

5.4.2.1 Exposition

In the following, I expose my fiction-based view of counterfactuals in science, which draws on the fiction view of models in important respects. Before that, it is nonetheless paramount to dispel one potential concern straight away.[213] In 'The Scientific Imagination' Salis & Frigg (2020) present counterfactual reasoning as a type of propositional imagination distinct from make-believe. According to them, the difference lies in the fact that even though counterfactual reasoning and make-believe share a minimal core, they differ in the additional elements which characterise them. These differences are, however, less clear-cut than they are presented as. The two distinctive elements of counterfactual reasoning, they argue, are *selectivity* and *reality orientation* (Salis & Frigg 2020, p.32-34). Since these elements capture aspects of the Lewisian similarity relation (see §5.3.1), I believe

[212] The operation of these principles is supported by evidence from mental models experiments (Byrne 2005), which as seen in §5.3.2 shows that when devising counterfactual scenarios people tend to imagine worlds with the same natural laws.

[213] Thanks to an anonymous reviewer for the suggestion to incorporate the following clarification.

they are none other than Walton's principles of generation for make-believe, i.e. the reality principle and the mutual belief principle.[214] At the same time, the distinctive feature of make-believe is that it is a *social activity* with a *normative aspect* (Salis & Frigg 2020, p.35-36). Counterfactual reasoning in science is certainly social, for scientific knowledge is socially embedded: scientific communities set the limits of which games of make-believe are permitted. The normative aspect of make-believe goes via its connection to imagined models, as it will hopefully become clear shortly. The additional elements thus turn out to be less distinct than intended. As a result, counterfactual reasoning and make-believe are not mutually exclusive categories.

Having shown how make-believe makes room for counterfactual reasoning, let me now introduce the make-believe view of scientific counterfactuals. This view characterises counterfactuals in science via make-believe in the way suggested by Kimpton-Nye in his paper 'Necessary Laws and the Problem of Counterlegals': a counterfactual's antecedent (A) serves as a prop in a game of make-believe. If the counterfactual's consequent (C) is prescribed for imagination given A and the pertinent principles of generation, then the counterfactual in question is true within the fiction, or f-true (Kimpton-Nye 2020, pp.529-530). These are the semantics of the make-believe view of scientific counterfactuals. My view will build upon Kimpton-Nye's account by fine-tuning its details so that the account fulfills the desiderata of *CONNECTION-TO-MODELS* and *NON-VACUITY*, which I identified as key to a naturalised account of counterfactuals in science.

Let us first look at *CONNECTION-TO-MODELS*. Throughout this chapter I have argued that there is a tight connection between counterfactuals and models in science; consequently, the make-believe view of scientific counterfactuals builds upon the fiction view of models. Scientific counterfactuals are connected to scientific models via their antecedents, which appeal to more or less full-blown scientific models. Recall the counterfactuals IDEAL PENDULUM and BOHR ATOM from §5.2: the antecedent of IDEAL PENDULUM, 'If pendulum *X* (an actual pendulum) were a simple pendulum...,' alludes to the simple pendulum model. Similarly, BOHR ATOM's antecedent, 'If atoms were Bohr atoms...,' makes explicit reference to Bohr's model of the hydrogen atom. Other counterfactuals will have antecedents which will not be connected to well-established full-blown scientific models, but under the minimal characterisation of 'model' here adopted,

[214] *Selectivity* describes the fact that when reasoning counterfactually one selects a particular manner in which things could have been different. To some extent, one could argue that the mutual belief principle targets the explanatory relevance that selectivity aims to track. *Reality orientation* is actually none other than the reality principle: from all the possible models a counterfactual's antecedent is compatible with, we select the one that is the closest to reality.

where models are sets of fictional propositions that are interpreted by someone to be about something, their widespread acceptance and use is not required. This needs to be so in order to accommodate impossible counterfactuals, which encode states of affairs whose modelling might be in a more embryonic state. This is the case for counterfactuals such as H-BONDS and SPACE TIME, which respectively describe an alternative bonding structure for water and a different dimensionality for space-time.

One might worry that *CONNECTION-TO-MODELS* is unnecessarily strong; while granting the close relationship between scientific counterfactuals and models, a potential objection to the present account is that there might be cases where counterfactual reasoning in science is not model-based.[215] Indeed, in some of the available accounts of scientific modelling the connection between theoretical activities and modelling in science is looser than the one I suggest. Weisberg (2007a), for instance, distinguishes modelling, which he characterises as *indirect* representation, from what he calls 'abstract direct representation' (ADR). In his account, the periodic table is *not* an instance of modelling because chemical properties of chemical elements such as valency or atomic weight are *directly* represented. But what does it mean to directly represent a target system? Weisberg differentiates theoretical activities on the basis of the *actions and intentions of theorists* (ibid., p.222) rather than the *products of theorising* (ibid., p.228). The key distinction aims to be whether it is the primordial aim of the theorist to "analyze and represent the properties of a real-world phenomenon, suitably abstracted, in the first instance" (ibid., p.222). If so, we have an instance of ADR, otherwise said representational object is a model. On Weisberg's account, all of the counterfactuals introduced in §5.2.1, except for perhaps H-BONDS and SPACE TIME, describe instances of ADR rather than modelling. But these, such as BOHR ATOM, certainly appeal to models, scientists would complain. My main reason for rejecting Weisberg's taxonomy of theoretical activities is that the make-believe view of scientific counterfactuals aims to be an account which takes scientific practice at face-value, and it is precisely the products of theorising, namely, those non-imagistic representations of real-world systems containing idealisations, abstractions, and/or approximations, which practitioners usually call models. My proposed characterisation of models accounts for this feature of practice, and as explained in §5.4.1.2 I consider this inclusivity a blessing rather than a curse. [216]

[215] Thanks to a reviewer for raising this concern, and pointing at the view I subsequently discuss.

[216] Furthermore, I worry that the focus on modelers' intentions doesn't sit well with targetless systems. Consider the case of aether models: presumably an example of ADR that fell out of steam once special relativity was developed? A focus on the products of theorising seems to better stand the test of time.

Having hopefully dispelled the concern that *CONNECTION-TO-MODELS* is unnecessarily strong, let us now assess how the present view fares with respect to *NON-VACUITY*. The output of non-vacuous truths results from the conceiving of counterfactual evaluation as a matter of whether a counterfactual's consequent follows from its antecedent and a set of principles of generation. The principles of generation involved in scientific counterfactual reasoning are those involved in make-believing with scientific models (see §5.4.1.2), namely, laws and other physical principles, mathematical knowledge, and rules of logical inference. What is distinctive about counterfactual reasoning in science versus everyday counterfactual reasoning is that the former presupposes the figure of a competent reasoner, namely, a member of the relevant scientific community. This competent user finds herself in a particularly good epistemic position, or at least the most optimal one, to identify the relevant principles of generation which a certain instance of counterfactual reasoning requires, and thus minimises the risk of faulty inferences.

The make-believe view of scientific counterfactuals also inherits the constraints on the scope of application of the fiction view of models. That is, make-believing in scientific contexts mostly has epistemic functions; other purposes such as recreation, art, or planning are excluded. Whether, and if so, how much the imagination contributes to such epistemic outcomes is a focus of debate in the literature on the (scientific) imagination.[217] Notwithstanding the lack of consensus regarding the precise contribution of the imagination on the epistemic character of its outputs, we can nonetheless discuss the roles these play. The examples provided in §5.2.1 can help identify some epistemic uses of counterfactuals in science: to convey explanations (a function of all the counterfactuals there listed), to drive empirical discovery (as is the case for H-BONDS), to justify the reliance on our current scientific theories (as is the case for BOHR ATOM), or to enhance tractability (such as in the case of idealisations like IDEAL PENDULUM). This list is not exhaustive, but only illustrative.

Regarding *NON-VACUITY*, reliance on make-believe has the further advantage that counterfactual evaluation is not defied by impossibilities. The make-believe view of scientific counterfactuals puts modal judgements about the content of propositions on hold: one need not believe or know that a certain counterfactual is true in order to engage in a game of make-believe triggered by its antecedent and the pertinent rules of generation, so the purported impossibility of a counterfactual

[217] For an overview of the debate about the epistemic uses of the imagination, see Kind & Kung (2016) and Badura & Kind (2021). For more specific discussion of the epistemic contribution of the scientific imagination, see Levy & Godfrey-Smith (2020). Myers (2021), Friend (2020), and Godfrey-Smith (2020), for instance, agree that the imagination cannot be used as an ultimate source of justification for knowledge.

antecedent makes no difference for the purposes of counterfactual reasoning and evaluation. This is in line with the theorisation by McLoone (2019, 2020), the results of experimental philosophical experiments presented by McLoone et al. (2023), and the results of cognitive psychological experiments (see §5.3.2).

One very reasonable worry about fiction-based accounts of counterfactuals is that while truth-in-fiction might suffice to evaluate a counterfactual's fictional meaning, it cannot help evaluate counterfactual claims extraneous to the fiction. To some extent, this is true: what happens in the fiction stays in the fiction. However, the make-believe view of scientific counterfactuals allows one to export what one has learnt about the fictional system outside of the game of make-believe via a suitable theory of scientific representation for models. It is when representation of real-world systems by imagined systems is successful that we can say that fictional truths about imagined systems ground truths about real-world systems. This is where *CONNECTION-TO-MODELS* and *NON-VACUITY* converge: the make-believe view of scientific counterfactuals shows that the question of the semantics of counterfactuals in science is none other than the question of scientific representation in disguise. This way the make-believe view of scientific counterfactuals closes the gap between standard philosophical theories of counterfactuals and scientific practice, and reclaims scientific counterfactuals as a rightful object of study for philosophers of science.

5.4.2.2 Extension on Previous Views

Fictionalist views of counterfactuals have been previously proposed or entertained by, for example, Kim & Maslen (2006), Kimpton-Nye (2020), McLoone (2019) and Wilson (2021). These views have a common kernel: counterfactuals' antecedents are props in games of make-believe. The make-believe view of scientific counterfactuals shares this kernel, but diverges from the previous views in a number of different respects.

Regarding *motivation*, while *NON-VACUITY* undoubtedly plays a central role in fictionalist proposals for counterfactuals, fictionalism is here positively considered a genuine alternative to other extant accounts of counterfactuals. In this respect, it differs from the proposals by Kimpton-Nye (2020) or Wilson (2021). They both suggest it as a last resource for the modal necessitarian, for whom counterlegals turn out to be counterpossibles, in order to evade the vacuity thesis the Lewis-Stalnaker semantics push the view into. Furthermore, Kimpton-Nye's view has somewhat bizarre consequences: the picture painted is one of coexistence between two different semantics for counterlegals depending on one's metaphysical

view on laws – the modal necessitarian may make use of make-believe semantics, while the rest of metaphysicians may resort to possible-worlds semantics as usual. Unfortunately, a semantic dualism of a similar kind also affects the modal necessitarian's own toolbox, who while encouraged to make use of fiction when faced with counterpossibles, still makes use of possible-worlds semantics when faced with regular counterfactuals.[218] This is a strange place to draw the line between fiction and non-fiction, especially considering that, as highlighted in §5.4.1.2, it might sometimes be complicated to judge whether a counterfactual is a counternomic.[219]

Regarding *scope*, the make-believe view of scientific counterfactuals also differs from the views by Kimpton-Nye (2020) and Wilson (2021) in that it does not apply exclusively to counterpossibles. While I have partly motivated my view, via *NON-VACUITY*, as a strategy that is able to accommodate both counternomics and countermetaphysicals in science, I have also argued it is in principle applicable to all scientific counterfactuals. Recall that the notion of fiction that make-believe exploits is *not* fiction understood as falsity but rather fiction as imagination, and that evaluation via make-believe is independent of one's beliefs about the (im)possibility of the state of affairs described by a counterfactual's antecedent (McLoone et al. 2023, Byrne 2022). These are observations which challenge the existence of any constitutive relationship between make-believe and nomological or metaphysical impossibilities. With respect to scope, the make-believe of scientific counterfactuals is a narrower version of Kim & Maslen (2006)'s make-believe view of counterfactuals simpliciter. However, their proposal does not dwell on anything like *CONNECTION-TO-MODELS*.[220] McLoone (2019) does make this link to this desideratum, by saying that "a modeler engages in a game of counterfactual make-believe when considering a model" (p.670). However, he proposes an inverse relation between the two (namely, from models to counterfactuals), and

[218] The reason for said encouragement is that "[i]f we were to embrace thoroughgoing fictionalism in the interest of avoiding semantic dualism, [modal necessitarianism] would lose its substance because it would be unclear in virtue of what counterlegal suppositions are different from counterfactual suppositions consistent with the laws, in which case, the claim that the laws are strongly necessary loses its bite" (Kimpton-Nye 2020, p.531). While I find this argument compelling, this semantic dualism might as well be read as an invitation to abandon modal necessitarianism.

[219] What I have in mind here are counterfactuals for which we have a high credence that they have an almost zero chance of being physically possible, such as those about supersymmetric particles (neutralinos, squarks, sleptons, etc.). Thanks to Mike T. Hicks for raising this point.

[220] Arguably, given my minimal characterisation of models, the two views could be made to coexist. However, there is definitely more structure in scientific counterfactual reasoning than in day-to-day counterfactual reasoning, so it would have to be worked out where the differences lie, whether in the nature or in the constraints on make-believe.

his proposed fictionalism extends only to formal models.[221]

Finally, regarding *morals*, the make-believe view of scientific counterfactuals highlights a new angle from which to evaluate counterfactuals in science: via the prism of scientific representation. I don't claim it is a better view of counterfactuals in science – scientific representation is yet another thorny subject – but it is definitely one that is better connected to how actual scientific practice works. After all, it is model-world similarities, and not similarity relations across possible worlds, what scientists talk about.

5.4.2.3 Case Studies

In what follows, I show how the make-believe view of scientific counterfactuals applies to a couple of counterfactuals from §5.2. Let us start by looking at IDEAL PENDULUM:

> **IDEAL PENDULUM:** "If pendulum X (an actual pendulum) were a simple pendulum, then for small swings its period T would only depend on its length l and the gravitational acceleration g." (Godfrey-Smith 2009, p.167)

In this case, the prop is the counterfactual's antecedent. Given that it clearly refers to the ideal model of the simple pendulum, we can consider the description of said model to be (the relevant) part of the prop. Now, what does said description include? In the first place, the pendulum's equation of motion, the differential equation:

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \frac{g}{l} sin\theta = 0 \tag{5.2}$$

Note that eq. (5.2) in its own does not constitute a model; it only becomes a simple pendulum model when its variables are given an interpretation, i.e. when these

[221] McLoone clearly thinks that the consequents of mathematical counterfactuals are entailed by their antecedents. However, if what is doing the work in the inference is a deductive inference from the antecedent, it is unclear what role make-believe plays. Furthermore, there is nothing like purely mathematical counterfactuals in science. Consider the example of the well-known Lotka-Volterra predator-prey model from ecology: "If a pair of populations had features F, then it would have/do G" Godfrey-Smith (2020, p.168). This is not an instance of mathematical necessity, for information about initial conditions regarding the type and number of prey, predator, as well as the parameters of growth and death rates need to be specified in order to evaluate whether the consequent follows. Even those who view models as essentially abstract mathematical objects, like Weisberg (2013), add an interpretative element to the definition of models.

are construed as representing properties of a real-world system. In this case, l is the length of the string, θ the angular displacement from the vertical to the pendulum, and g is the local acceleration of gravity acting on the bob.[222]

Other things that are part of the model's description and hence the prop are the assumptions made when constructing said model: the string is massless, the bob is a point mass, there is no air resistance, etc. Finally, one might include a visualisation of said model which facilitates comprehension of the model's features, such as the following:

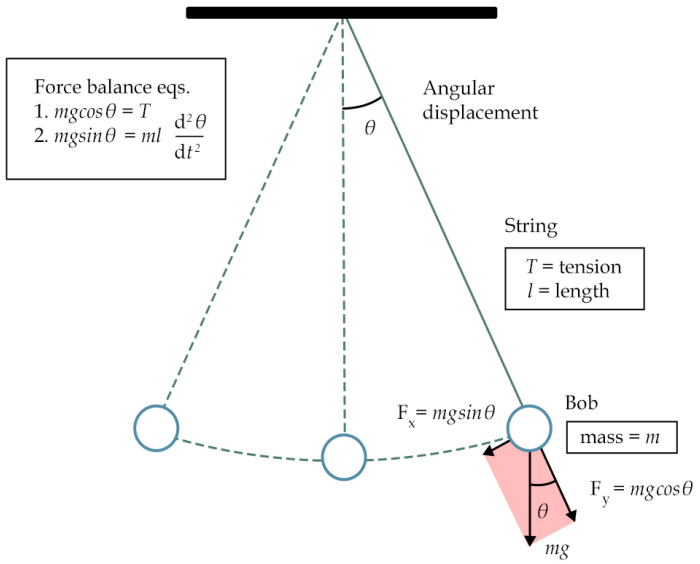


Figure 5.1: Simple pendulum representation

In short, the relevant part of the prop is the description of the simple pendulum model, which includes: the equation of motion, the information of what property of real-world systems each variable stands in for, the assumptions that went into the model, and potentially a visualisation like fig. 5.1.

Next, there is the question of what the principles of generation for this game of make-believe are. One principle of generation clearly at play in this case is the mathematical knowledge of the Taylor series that for small θs (no more than 15°), $sin\theta \approx \theta$ – what is called the 'small angle approximation.'[223] In this regime eq. (5.2) becomes an analytically solvable second-order differential equation analogous to the one for the simple harmonic oscillator. Furthermore, by adding in the small angle approximation, another principle of generation becomes manifest: the physical knowledge that the pendulum's period, T, will then not depend on the

[222] See Salis (2020) for a detailed exposition of this example.

[223] Lima & Arun (2006) report an even lower value for the angular displacement. If an error less than 0.1% is desired, the displacement cannot be more than 7°.

angular displacement, as eq. (5.3) shows:

$$T \approx 2\pi \sqrt{\frac{l}{g}} \tag{5.3}$$

Equation (5.3), together with the interpretation of what physical properties the variables stand in for, constitutes another principle of generation at play in this example. Now, when plugging in the simple pendulum model description together with the aforementioned principles of generation, one is prescribed to imagine that '... then for small swings its period T would only depend on its length l and the gravitational acceleration g.' Hence, IDEAL PENDULUM is f-true. Note that here I am assuming the figure of a competent reasoner, who has the epistemic resources – i.e. understands the prop and has access to the right principles of generation – to successfully pretend the imaginings triggered by IDEAL PENDULUM.

Having examined in detail an application of the fiction view of scientific counterfactuals to a counterfactual which clearly invokes a familiar model, I would now like to turn to a less crystal-clear example. Consider the following counterfactual:

> **H-BONDS:** "If water had not had intermolecular hydrogen bonding, then it would have been a gas at room temperature." (adapted from Tan 2017, p.959).

One might argue that H-BONDS looks less intuitively connected to a model than IDEAL PENDULUM. This is because it doesn't appeal to a well-established scientific model, yet that is not to say there isn't one. The evaluation of impossible counterfactuals will often not only require the analysis of a model but also the construction of one (e.g. by interpolating from the data points in fig. 5.2 and providing an interpretation of these); this is not strange, since these possibilities are more remote than those expressed by regular counterfactuals, so it should not surprise us that the models that represent them haven't yet been explored.

Having said that, let us examine H-BONDS. Here the prop is the description of water not bonded via hydrogen bonds. An informed reasoner knows that water's normal boiling point is 100°, so it is liquid at room temperature. However, the other molecules of the same group of hydrides, H_2S H_2Se and H_2Te (group 16), are gaseous at room temperature. This is because they are bonded by van der Waals bonds, which are weaker than the dipole-dipole interactions of the hydrogen bonds. Their boiling points, represented by the blue line in fig. 5.2, obey

a nearly linear relationship with respect to molecular mass, only broken by H_2O's behaviour.[224] Extrapolating from this relationship one would expect H_2O's boiling point to be about -70 or -80ºC (Tan 2017, p.959).

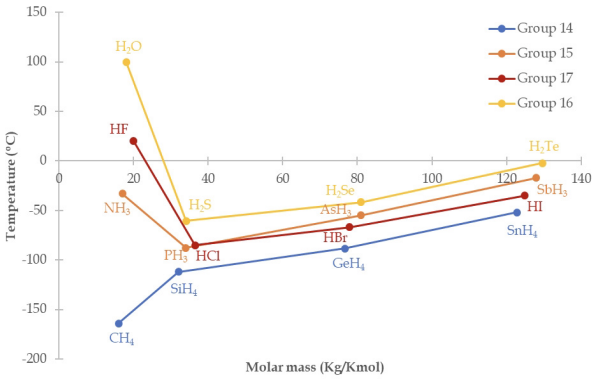


Figure 5.2: Relation of boiling points (ºC) versus molar mass (Kg/Kmol) for molecules in hydrides' groups 14, 15, 16, and 17.

The almost-linear relationship between molar mass and boiling point for hydrides, as well as the knowledge that all molecules of group 16 except for H_2O are bonded via van der Waals bonds, are the principles of generation in action in this game of make-believe. Note that here the periodic table, which Weisberg classifies as an instance of ADR (see §5.4.2.1), is part of the principles of generation in the game of make-believe induced by this counterfactual's antecedent. Additionally, given that H-BONDS' antecedent doesn't specify an alternative bonding structure for H_2O, we may add in a reality-oriented principle à la Walton which ensures that its alternative bonding is that of its most similar molecules, namely, those in the same group. Via these implied truths, the model is generated.

Under the fiction of H-BONDS' antecedent, it follows that '... then [water] would have been a gas at room temperature.' H-BONDS is therefore f-true. Does this counterfactual tell us anything about real water? This f-truth is informative beyond the content of the game of make-believe provided one takes the model output by said counterfactual to represent actual water. Here the answer will depend on one's account of scientific representation, as well as one's metaphysical commitments. Note that at this point modal questions are brought back into the picture: would anyone argue that it is essential to water that it is bonded in the way it actually is, because this is what gives it its life-sustaining properties, the modal modelling of the sort involved when reasoning about this counterfactual wouldn't offer any connections between truth-in-fiction and truth. Unfortunately,

[224]Note that the same almost-linear pattern is observed for the other three groups.

this discussion is beyond the scope of this chapter. Regarding modality, the important point to emphasise is that questions of this sort can be bracketed when evaluating scientific counterfactuals, and this is a feature which make-believe successfully captures.

5.5 Conclusion

In this chapter, I vindicated the *make-believe view of scientific counterfactuals*. This is a fiction-based view of counterfactuals which has been defended as an account of counterfactuals and counterfactual reasoning in science which takes scientific practice at face value. On the basis that modelling is the keystone of scientific reasoning, I established *CONNECTION-TO-MODELS* as a desideratum for any view which aims to respect scientific practice. The make-believe view of scientific counterfactuals builds upon indirect fiction views of models in the following way: counterfactuals' antecedents appeal to more or less well-established scientific models, which are props in games of make-believe which may or may not invite the truth of the respective consequents. In this way, games of counterfactual make-believe in science are none other than rational reconstructions of model-building and model-based reasoning processes.

What else can be learnt by reflecting upon actual scientific practice? Undeniably, many of the assumptions that go into scientific models describe nomologically or metaphysically impossible states of affairs. In this respect, recent theorising and experimenting from scientifically-minded philosophers and cognitive scientists has defied the vacuity thesis, a well-known result in the Lewis-Stalnaker similarity analysis of counterfactuals. Hence, I argued that naturalising the study of counterfactuals also requires paying attention to the desideratum of *NON-VACUITY*. One way to handle nomological and metaphysical impossibilities in counterfactuals is by treating the states of affairs which counterfactuals describe as imaginary. The central role of the imagination in counterfactual reasoning has been highlighted both by cognitive science researchers and philosophers of science. The make-believe view of scientific counterfactuals respects this intuition, because it characterises fiction as the attitude of imaginative engagement Walton called 'make-believe.'

Furthermore, the make-believe view of scientific counterfactuals characterises the nature of the imagination as propositional and the resulting semantics for counterfactuals in the following way: a counterfactual is fictionally-true iff its consequent is prescribed to be imagined by its antecedent (a model description,

and hence satisfying *CONNECTION-TO-MODELS*) and by the pertinent principles of generation, which in science are logical rules of inference, mathematical knowledge, and other field-specific theoretical knowledge. The modal status of the state of affairs described by a counterfactual antecedent is not an ingredient of make-believe, so the vacuity thesis can thus be evaded. One positive feature of the present account is precisely this independence of counterfactual evaluation from any particular metaphysics.

Such modal questions, temporarily bracketed during games of make-believe triggered by counterfactuals' antecedents and the appropriate principles of generation, are foregrounded when one asks whether – and if so, how – a certain model represents the world. Whether the imaginings triggered by games of counterfactual make-believe are true or false outside the fiction is a question our preferred theory of scientific representation shall address. The close connection of scientific counterfactuals with the question of scientific representation via models is precisely the feature which the present account has aimed to highlight, a feature which has been to date significantly overlooked in the study of the meanings of counterfactuals in science. Via make-believe, I urge philosophers of science to reclaim their position in the present debate.

Conclusions

Is there any causal relationship between the relative distance between the Earth and the Moon and the magnitude of the Earth's tides? Via this case study, I have introduced the Possibility Question (PQ): what notion of possibility underlies interventions? This is a central yet underexplored question in James Woodward's interventionist theory of causation. Since the PQ is the point of confluence of many issues regarding modality and explanation discussed, I will use it as a backdrop against which to tie together the various conclusions reached throughout this thesis and, based on these, suggest future lines of research.

The most attractive feature of interventionist theories of causation is that they provide tools to determine whether causal claims and the associated counterfactuals are meaningful. 'If I were to drop an apple, then it would fall' is supported by any instance of an apple falling when dropped. Similarly, an intervention introducing a euro in my pocket will confirm that 'All coins in my pocket are dimes' is not a true causal generalisation. However, Woodward's account does not constrain such manipulations to actual interventions; these need only be hypothetical (i.e. possible but non-actual). Thus, the PQ seeks to find the notion of possibility which underwrites interventions.

In chapter 1, I contested Woodward's argument that the relevant notion of possibility is not physical. Firstly, the argument is unsound: physical possibility gets conflated with the possibility of ideal interventions and reduced to actualisability. Furthermore, since physical possibility is essentially determined by the laws of nature, the possibility of interventions is relative to the laws relevant to the explanatory context in question. Therefore, the inference that there is no physically possible intervention that would double the relative distance between the Moon and the Earth from the fact that there is no actual (direct, surgical, arrow-breaking, and statistically independent) process that brings the Moon to its new position is too quick. Secondly, Woodward's resulting suggestion that the relevant notion of possibility for interventions is logical or conceptual possibility is in tension with the causal realism presupposed by interventionism. Causal realism amounts to worldly dependence, not dependence between concepts, so conceptual possibilities might be causal and non-causal. Moreover, answering the PQ via conceptual possibility totally undermines the rationale for using interventions and makes it unclear how one should elucidate the meaning of interventionist counterfactuals. Despite a palpable scepticism towards the prospects of conceptual possibility as

a candidate possibility notion for interventions, in chapter 2 I examined the best possible reconstruction of an objective version of conceptual possibility.

One of the most promising ways of developing the claim that causal relations are mind-independent and that we might find out about these via conceptual possibilities is a rationalist strategy that links conceivability with an objective alethic modal notion. David Chalmers' weak modal rationalism (WMR) provides such a strong link between conceivability and possibility: the former entails the latter. The feasibility of the inference from (ideal 1-) conceivability to 2-possibility depends on the metaphysical details about physical properties with which one fills Woodward's interventionism. To this end, I broadly classified views about the individuation of physical properties into two camps – functionalist and non-functionalist views – and showed that the inference from 1-conceivability to 2-possibility only obtains in functionalist views. Nonetheless, this rationalist route might be contested on the grounds that objectivity is attained via reason and not a posteriori, which does not fare well with interventionism's empiricist motivations.

The rationalist route to the PQ demands too many concessions, the reason why in chapter 3 I settled for what I think is the best answer to the PQ: interventions are physical possibilities. The key lies in Woodward's Invariance View of laws, according to which laws are relationships which are invariant under BaCs and ICs including interventions. The world's vast network of invariance patterns defines the space of physical possibility. Since interventions need not be actual, I argued that the PQ is best answered in terms of Weak Physical Possibility (WPP), according to which interventions are physically possible iff they are consistent with the actual invariance patterns and some range of possible ICs consistent with the actual invariance patterns including the actual ICs. Beyond the original PQ, I suggested an even weaker notion of physical possibility than WPP to characterise interventions on the content of laws and thereby answer the resulting generalised PQ these interventions trigger: interventions are physically possible in the sense of Very Weak Physical Possibility (VWPP) iff they are consistent with actual *higher-order invariance* patterns and some range of ICs^n consistent with these patterns including the actual ICs^n.

Since invariance can generally be empirically uncovered, the physical possibility space, qua modal space delineated by invariance patterns, has an empirical basis. That said, in chapter 3 we saw how invariance can also be hypothetically tested, provided there is a significant accrual of information about invariance on which we can rely for these hypothetical tests. This epistemic point links to Woodward's suggestion that possibilities unravelled via models are conceptual: in the Moon-tides example, since there is no actual intervention that would double

the relative distance between the Earth and the Moon but a Newtonian model would permit us to calculate such intervention, this might suggest it is a conceptual possibility. I don't think so: when models are representations of the world, the possibilities they encode are not conceptual: they are epistemic possibilities which latch onto worldly (and, therefore, objective) possibilities. In this respect, scientific models differ from metaphysical or mathematical models: they rely partly on a posteriori knowledge. Provided empirically-tested invariance is part of the models, the possibilities encoded by these models are physical. Confirming my assessment of Woodward's argument in chapter 1, the Moon-tides intervention is physically possible in the sense of WPP because it is consistent with the actual law of Newtonian gravitation and a possible but non-actual value for the relative distance between the Moon and the Earth consistent with said law. The possibility in question is objective and causal, whether it qualifies as an intervention – i.e. it strictly satisfies (IV) – will depend not only on one's (puritan or liberal) view of interventionism but also on whether alternative strategies to secure the explanatoriness of this case study are available.

Invariant dependencies, which are worldly, make counterfactual and causal claims true. In chapter 3, I argued that interventionism's causal realism could be vindicated with a *primitivism about invariance*: there are brute facts about invariance which make counterfactuals true. This primitivism about invariance has many attractive features. First, it is metaphysically very minimal: it does not require that one provides grounds for laws and does not appeal to metaphysical modality, a notion that is not part of scientific characterisations of laws. Therefore, it does not interfere with the *functional* motivation of the Invariance View to account for how the concept of 'law' is used in practice. Second, it would shield the theory against one of the most recurrent criticisms it has been met with: that it provides no truth-makers for causal claims. Third, the result is still a non-reductive view of laws (or, rather, invariance).

The primitivism about invariance mentioned above also secures Woodward's interventionism as an objectivist account of causal explanation. While the focus of the first part of the thesis was on the modal metaphysics and epistemology underlying interventionism, the second part examined its prospects as a theory of scientific explanation. As we saw in both chapter 3 and chapter 4, if only those counterfactuals which could be made true via interventions are explanatory, this poses significant challenges for counternomics. Since counternomics describe physically impossible interventions, they turn out to be non-explanatory – or, more precisely, causally non-explanatory. However, this verdict goes against the theory's motivation to provide a functional account of causation, laws, and counterfactuals.

For instance, scientists seem to assign non-trivial truth-values to counterfactuals describing what would happen were the laws different or were physical situations idealised. In both chapter 3 and chapter 4, I suggested potential ways in which the interventionist account of causal explanation might be extended in order to assign truth-values to these counternomics and account for their explanatory functions.

In chapter 3, I developed the Extended Invariance View, which builds upon Woodward's invariance-based view of laws: the world is structured by invariant relationships which, depending on their degree of invariance, are more or less fundamental. This allows us to understand changes to the content of Newton's law of universal gravitation as invariant under higher-order invariant relationships such as Newton's 2^{nd} law. Thus, changes to a law's content are *higher-order interventions* (or non-I-changes, depending on one's view of interventionism) because they are only physically possible with respect to invariant relationships with a higher order than the one whose content is being modified – it is also for this reason that they are physically possible in the sense of VWPP. This view has some limitations, most importantly: just like the Invariance View, it requires that the worldly infrastructure of causation is such that counterfactuals admit a characterisation in terms of invariant relationships, on the one hand, and the ICs under which they are invariant, on the other.

In chapter 4, I argued that idealising counternomics have an explanation-enabling function: while they are not causally explanatory, their role is to enable causal explanations with idealised models. The explanation-enabling function of idealising counternomics can be identified via an *approximation* strategy: we cannot intervene on a real object to turn it into an ideal object, but we can intervene on a real object to test whether its behaviour approximates that of the ideal object. Whenever this is the case, the idealised models – and, consequently, the associated idealising counternomic – enable causal explanations of the system under study. Unfortunately, this strategy is highly restrictive, for it only applies to idealised models which can be quasi-gradually de-idealised. Since it is not clear whether interventionism can elucidate the meaning of all counternomics, in chapter 5 I turned to Walton's notion of fiction as make-believe to provide a unified account of counterfactuals in science.

The make-believe account of scientific counterfactuals is a fiction-based view of counterfactuals. It takes scientific practice at face value because i) it acknowledges the tight connection between scientific models and scientific counterfactuals, and ii) it takes seriously the empirical data that supports the view that counterfactuals with impossible antecedents have non-trivial truth-values. More specifically, it builds upon the indirect fiction view of models as follows: counterfactuals'

antecedents appeal to more or less well-established scientific models, which are props in games of make-believe which may (not) invite the truth of their consequents. Make-believe offers unique resources for counterfactual reasoning because it denotes a propositional attitude of imaginary engagement and, as such, can accommodate the result of non-vacuity via fictional truth (f-truth): a counterfactual is f-true iff its antecedent and the rules of the game prescribe the imagining of its consequent. Whether and, if so, how f-truths map onto the world is something one's preferred theory of representation must answer. The make-believe account of scientific counterfactuals is certainly an epistemically demanding view, for the successful evaluation of counterfactuals depends on the knowledge and competent use of the props and principles of generation. However, so is Woodward's, for as we saw in chapter 3, the evaluation of hypothetical interventions sets similarly high standards for required background causal knowledge. Where the make-believe view of scientific counterfactuals does significantly better than the interventionist is in unificatory potential: it can clarify the meaning of all sorts of impossible counterfactuals, provided they serve some epistemic aim.

All in all, chapter 5 brings together some of the themes underlying this thesis: modal metaphysics (nomological and metaphysical (im)possibilities), modal epistemology (conceivability, experimentation, and the imagination as guides to (im)possibility), scientific explanation (causal and non-causal), and scientific representation (via scientific models). More generally, it also links to the overarching advocating for a scientifically-respectful methodology (scientific methodology and reasoning are central to interventionism). On this note, I would like to conclude by highlighting the questions that deserve further scrutiny or need further development:

- *The applicability question:* where do causal notions apply? This is a question I raised alongside the PQ, but was beyond the scope of this thesis. If causal concepts apply to the Moon-tides case study, then the criteria for interventions (*IV*) need to be amended. More generally, it looks like the *SURGICALITY* and *ARROW-BREAKINGNESS* criteria are barely ever satisfied at the physical level, so it would be interesting to explore whether a more realistic notion of 'cause' that does not require to treat (*IV*) as regulative ideals is available and, if so, what the pay-offs of this notion would be, e.g. would that leave us with a too flexible and, therefore, not epistemically useful concept of cause?

- *The scope of interventions:* another interesting question that arose in the context of the Extended Invariance View is whether the notion of an intervention can be successfully implemented in non-causal explanation. Concerning this

point, it would be good to more carefully weigh the two characterisations of changes to the content of laws options considered in chapter 3, namely, whether changes to the content of lower-order invariant relationships which are possible according to higher-order invariant relationships ought to be treated as interventions or not. Treating them as higher-order interventions requires liberalising the notion of an intervention, and treating them as explanatory non-interventions requires liberalising the notion of explanation along the lines of CR-monism. What makes explanations of higher-order invariance, just like RG explanations, different from traditional examples of non-causal explanation discussed in the context of CR-monism is that these explanations relate physical – and not mathematical or conceptual – facts. Therefore, one question that deserves further scrutiny is whether the resulting dependencies are physical. This brings together research topics such as the (physical/mathematical) nature of kinematical constraints or the (causal/non-causal) role of highly abstract models in scientific explanations.

- *A theory of causal aptness:* causal realism can only be vindicated provided that models represent their targets. The question of what makes causal models apt was briefly introduced in chapter 1 and has permeated many subsequent discussions in this thesis. The interventionist theory of causation only provides heuristics for causal aptness, yet no full-blown theory of causal representation is developed. To this end, I believe the interventionist ought to say more about when a causal model represents a target. As seen in chapter 4, this is a particularly pressing question for idealised models.

- *The epistemic uses of the imagination:* one question I haven't had the space to discuss is what the precise contribution of make-believe to the epistemic character of its outputs is. Of course, addressing this question would require a more precise characterisation of the props (the models' descriptions) and the principles of generation. More generally, this question should be framed within debates on the role of the imagination in science.

- *The explanatory functions of scientific counterfactuals:* scientific counterfactual reasoning has epistemic aims, among which explanation is central. In chapter 4, RG explanation presented itself as a very informative case study, for it brought about the diversity of explanations in which abstract models and the associated counterfactuals may feature. Exporting the discussion beyond abstraction, a rich future research project might involve the provision of a taxonomy of explanatory functions of scientific counterfactuals.

Bibliography

Adams, E. (1965). The logic of conditionals. *Inquiry*, *8*(1-4), 166–197.

Airy, G. B. (1830). On certain conditions under which a perpetual motion is possible. *Transactions of the Cambridge Philosophical Society*(3), 369–372.

Albert, D. (2000). *Time and chance*. Cambridge, Massachussets: Harvard University Press.

Albert, D. (2015). *After physics*. Cambridge: Harvard University Press.

Andersen, H. (2013). When to Expect Violations of Causal Faithfulness and Why It Matters. *Philosophy of Science*, *80*(5), 672–683.

Andersen, H. (2018). Complements, Not Competitors: Causal and Mathematical Explanations. *The British Journal for the Philosophy of Science*, *69*(2), 485–508.

Andreas, H., & Casini, L. (2019). Hypothetical Interventions and Belief Changes. *Foundations of Science*, *24*(4), 681–704.

Armstrong, D. (1989). *A combinatorial theory of possibility*. Cambridge: Cambridge University Press.

Atkins, P. W. (2018). *Atkins' physical chemistry* (11th ed.). Oxford: Oxford University Press. (first edition: 1978)

Badura, C., & Kind, A. (Eds.). (2021). *Epistemic uses of imagination*. New York, NY: Routledge.

Baierlein, R. (1971). *Atoms and information theory: An introduction to statistical mechanics*. San Francisco: W. H. Freeman.

Batterman, R. W. (2000). Multiple Realizability and Universality. *The British Journal for the Philosophy of Science*, *51*(1), 115–145.

Batterman, R. W. (2002). Asymptotics and the Role of Minimal Models. *The British Journal for the Philosophy of Science*, *53*(1), 21–38.

Batterman, R. W. (2005). Critical phenomena and breaking drops: Infinite idealizations in physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *36*(2), 225–244.

Batterman, R. W. (2010). On the Explanatory Role of Mathematics in Empirical Science. *The British Journal for the Philosophy of Science*, *61*(1), 1–25.

Batterman, R. W., & Rice, C. C. (2014). Minimal Model Explanations. *Philosophy of Science*, *81*(3), 349–376.

Baumgartner, M., & Gebharter, A. (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science*, *67*(3), 731–756.

Bhogal, H. (2020). Nomothetic Explanation and Humeanism about Laws of Nature. In *Oxford Studies in Metaphysics Volume 12* (pp. 164–202). Oxford University Press.

Bird, A. (2007). *Nature's Metaphysics*. New York: Oxford University Press.

Block, N., & Stalnaker, R. (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *The Philosophical Review*, *108*(1), 1–46.

Bokulich, A. (2011). How scientific models can explain. *Synthese*, *180*(1), 33–45.

Brogaard, B., & Salerno, J. (2013). Remarks on counterpossibles. *Synthese*, *190*(4), 639–660.

Bursten, J. R. S. (2021). The Function of Boundary Conditions in the Physical Sciences. *Philosophy of Science*, *88*(2), 234–257.

Byrne, R. M. J. (2005). *The rational imagination: how people create alternatives to reality*. Cambridge: MIT.

Byrne, R. M. J. (2007). Précis of *The Rational Imagination: How People Create Alternatives to Reality*. *Behavioral and Brain Sciences*, *30*(5-6), 439–453.

Byrne, R. M. J. (2016). Counterfactual Thinking. In *The Routledge Handbook of Philosophy of Imagination* (pp. 339–352). London [England]; New York: Routledge.

Byrne, R. M. J. (2022). How people think about the truth of impossibilities. *Unpublished manuscript*.

Byrne, R. M. J., & Johnson-Laird, P. (2009). 'If' and the problems of conditional reasoning. *Trends in Cognitive Sciences*, *13*(7), 282–287.

Carnap, R. (1947). *Meaning and necessity: a study in semantics and modal logic*. Illinois: University of Chicago Press.

Cartwright, N. (1979). Causal Laws and Effective Strategies. *Noûs*, *13*(4), 419.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: New York: Clarendon Press; Oxford University Press.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Clarendon.

Cartwright, N. (1999). Causal Diversity and the Markov Condition. *Synthese*, *121*(1/2), 3–27.

Cartwright, N. (2002). Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science*, *53*(3), 411–453.

Chalmers, D. J. (1996). *The conscious mind: in search of a fundamental theory*. New York: Oxford University Press.

Chalmers, D. J. (2002). Does conceivability entail possibility? In J. Hawthorne & T. S. Gendler (Eds.), *Conceivability and Possibility* (pp. 145–200). Oxford, New York: Oxford University Press, Clarendon Press.

Chalmers, D. J. (2004). Epistemic Two-Dimensional Semantics. *Philosophical Studies*, *118*(1/2), 153–226.

Chalmers, D. J. (2010). *The character of consciousness*. Oxford: Oxford University Press.

Chalmers, D. J. (2012). *Constructing the world*. Oxford: Oxford University Press.

Chalmers, D. J. (2016). Panpsychism and Panprotopsychism. In G. Bruntrup & L. Jaskolla (Eds.), *Panpsychism* (pp. 19–47). Oxford: Oxford University Press.

Chalmers, D. J. (2021). Finding Space in a Non-Spatial World. In *Philosophy Beyond Spacetime* (pp. 154–181). Oxford University Press.

Chisholm, R. M. (1955). Law Statements and Counterfactual Inference. *Analysis*, *15*(5), 97–105.

Clarke-Doane, J. (2021). Metaphysical and absolute possibility. *Synthese*, *198*(S8), 1861–1872.

Costa, F., & Shrapnel, S. (2016). Quantum causal modelling. *New Journal of Physics*, *18*(063032), 1-21.

Coulomb, C. (1785). Second mémoire sur l'électricité et le magnétisme. *Histoire de l'Académie Royale des Sciences*, 578–611.

Craver, C. F. (2007a). Constitutive Explanatory Relevance. *Journal for Philosophical Research*, *32*, 3–20.

Craver, C. F. (2007b). *Explaining the Brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Curiel, E. (2016). *Kinematics, Dynamics, and the Structure of Physical Theory.* Retrieved from `http://philsci-archive.pitt.edu/11959/`

Currie, G. (1990). *The nature of fiction*. Cambridge [England]; New York: Cambridge University Press.

Davidson, D. (1980). Mental Events. In *Essays on Actions and Events* (pp. 207–227). Oxford: Clarendon Press. (original date: 1970)

Davies, M., & Humberstone, L. (1980). Two notions of necessity. *Philosophical Studies*, *38*(1), 1–30.

DeLong, D. M. (2021). *homopteran | Natural history.* Retrieved 2022-09-26, from `https://www.britannica.com/animal/homopteran`

Dohrn, D. (2021). The Counterpossibles of Science vs. the Science of Counterpossibles. *The British Journal for the Philosophy of Science*, 716–769.

Dowe, P. (1992). Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory. *Philosophy of Science*(59), 195–216.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Edgington, D. (2003). What if? Questions About Conditionals. *Mind and Language*, *18*(4), 380–401.

Edgington, D. (2008). Presidential Address: Counterfactuals. *Proceedings of the Aristotelian Society (Hardback)*, *108*(1 part 1), 1–21.

Education, N. (2014). *Hardy-Weinberg equilibrium.* Retrieved 2022-07-10, from `https://www.nature.com/scitable/definition/hardy-weinberg-equilibrium-122/`

Ehrenfest, P. (1917). In what way does it become manifest in the fundamental laws of physics that space has three dimensions? *Proceedings of the Amsterdam Academy*, *20*, 200–209.

Elgin, M., & Sober, E. (2002). Cartwright on explanation and idealization. *Erkenntnis*, *57*(3), 441–450.

Emmerson, N. (2021). A Defence of Manipulationist Noncausal Explanation: The Case for Intervention Liberalism. *Erkenntnis*.

Esfeld, M. (2012). *Physique et métaphysique: une introduction à la philosophie de la nature* (1st ed.). Lausanne: Presses polytechniques et universitaires romandes.

Espino, O., Byrne, R. M. J., & Johnson-Laird, P. (2020). Possibilities and the parallel meanings of factual and counterfactual conditionals. *Memory & Cognition*, *48*(7), 1263–1280.

Evans, J. B. T., & Over, D. E. (2004). *If*. Oxford; New York: Oxford University Press.

Evans, P. W. (2018). Quantum Causal Models, Faithfulness, and Retrocausality. *The British Journal for the Philosophy of Science*(3), 745–774.

Ferrari, R., & Joyce, T. (2004). *Introduction to Observational Physical Oceanography, course notes.* Retrieved from `https://ocw.mit.edu/courses/12-808-introduction-to-observational-physical-oceanography-fall-2004/resources/course_notes_15n/`

Fine, K. (2002). Varieties of Necessity. In T. S. Gendler & J. Hawthorne (Eds.), *Conceivability and Possibility* (pp. 253–281). Oxford, New York: Oxford University Press, Clarendon Press.

Franklin-Hall, L. R. (2016). High-Level Explanation and the Interventionist's 'Variables Problem'. *The British Journal for the Philosophy of Science*, *67*(2), 553–577.

French, S., & McKenzie, K. (2012). Thinking Outside the Toolbox: Towards a More productive Engagement Between Metaphysics and Philosophy of Physics. *European Journal of Analytic Philosophy*, *8*, 42–59.

Friend, S. (2020). The Fictional Character of Scientific Models. In A. Levy & P. Godfrey-Smith (Eds.), *The Scientific Imagination* (pp. 102–127). New York: Oxford University Press.

Frigg, R. (2010a). Fiction in Science. In J. Woods (Ed.), *Fictions and models: new essays* (pp. 247–287). Philosophia Verlag.

Frigg, R. (2010b). Models and fiction. *Synthese*, *172*(2), 251–268. (2010a)

Frigg, R., & Nguyen, J. (2016). The Fiction View of Models Reloaded. *The Monist*, *99*(3), 225–242.

Frigg, R., & Nguyen, J. (2020). *Modelling Nature: An Opinionated Introduction to Scientific Representation*. Springer.

Frisch, M. (2014). *Causal reasoning in physics*. Cambridge: Cambridge University Press.

Frisch, M. (2020). *Causation in Physics* (Fall 2020 ed.). Metaphysics Research Lab, Stanford University. Retrieved from `https://plato.stanford.edu/archives/fall2020/entries/causation-physics/`

Frisch, R. ([1938] 1995). Autonomy of Economic Relations. In D. Hendry & M. S. Morgan (Eds.), *The Foundations of Econometric Analysis.* Cambridge, U.K.: Cambridge University Press.

Gebharter, A. (2017). Causal Nets and Woodwardian Interventionism. In *Causal Nets, Interventionism, and Mechanisms* (pp. 71–114). Cham: Springer International Publishing.

Gendler, T., & Hawthorne, J. (Eds.). (2002). *Conceivability and possibility*. Oxford, New York: Oxford University Press, Clarendon Press.

Gendler, T. S. (1998). Galileo and the Indispensability of Scientific Thought Experiment. *The British Journal for the Philosophy of Science*, *49*(3), 397–424.

Giere, R. N. (1999). *Science without laws*. Chicago: University of Chicago Press.

Glymour, C. N. (2006). Markov Properties and Quantum Experiments. In W. Demopoulos & I. Pitowsky (Eds.), *Physical Theory and its Interpretation* (pp. 117–126). Springer.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology & Philosophy*, *21*(5), 725–740.

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies*, *143*(1), 101–116.

Godfrey-Smith, P. (2020). Models, Fictions, and Conditionals. In A. Levy & P. Godfrey-Smith (Eds.), *The Scientific Imagination* (pp. 154–177). New York: Oxford University Press.

Goldman, A. I. (1992). Empathy, Mind, and Morals. *Proceedings and Addresses of the American Philosophical Association*, *66*(3), 17.

Gottlob, F. (1892). On Sinn and Bedeutung. In M. Beany (Ed.), *The Frege Reader* (pp. 151–171). Oxford: Blackwell (1997).

Groen, P., Claudia, C., Tricker, R. A., & Gordon, A. L. (2022). *Tide | physics.* Retrieved from `https://www.britannica.com/science/tide`

Guay, A., & Pradeu, T. (2020). Right out of the box: how to situate metaphysics of science in relation to other metaphysical approaches. *Synthese*, *197*(5), 1847–1866.

Haavelmo, T. (1944). The Probability Approach in Econometrics. *Econometrica*, *12*, iii–vi+1–115.

Handfield, T. (2004). Counterlegals and Necessary Laws. *The Philosophical Quarterly*, *54*(216), 402–419.

Hausman, D. M. (1998). *Causal Asymmetries*. Cambridge: Cambridge U.P.

Hausman, D. M. (2005). Causal Relata: Tokens, Types, or Variables? *Erkenntnis*, *63*(1), 33–54.

Hausman, D. M., & Woodward, J. (1999). Independence, Invariance and the Causal Markov Condition. *The British Journal for the Philosophy of Science*, *50*(4), 521–583.

Hawthorne, J. (2001). Causal Structuralism. *Nous*, *35*(s15), 361–378.

Heil, J. (2003). *From an Ontological Point of View* (1st ed.). Oxford: Oxford University Press.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175.

Hicks, M. T. (2022). Counterpossibles and counterparts. *The Journal of Philosophy*, *119*(10), 542-574.

Hicks, M. T., & Wilson, A. (forthcoming). How Chance Explains. *Noûs*.

Hirvonen, I., Koskinen, R., & Pättiniemi, I. (2021). Modal inferences in science: a tale of two epistemologies. *Synthese*, *199*(5-6), 13823–13843.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy*(98), 273–299.

Hitchcock, C. (2012). Events and times: a case study in means-ends metaphysics. *Philosophical Studies*, *160*(1), 79–96.

Hitchcock, C. (2021). *Probabilistic Causation* (Spring 2021 ed.). Metaphysics Research Lab, Stanford University. Retrieved from `https://plato.stanford.edu/archives/spr2021/entries/causation-probabilistic/`

Hitchcock, C., & Woodward, J. (2003). Explanatory Generalizations, Part II: Plumbing Explanatory Depth. *Nous*, *37*(2), 181–199.

Holovatch, Y. (2006). Introduction to renormalization. *Condensed Matter Physics*, *9*(2(46)), 237–262.

Hüttemann, A. (2021). *A minimal metaphysics for scientific practice*. Cambridge, United Kingdom: Cambridge University Press.

Iranzo-Ribera, N. (2015). *Study of the spontaneous activity in neuronal cultures* (Bachelor's thesis, University of Barcelona). Retrieved from `http://hdl.handle.net/2445/67336`

Iranzo-Ribera, N. (2020). *Quantum Queerness: Can we causally explain bipartite quantum entangled systems?* Retrieved from `https://www.nonlinearlight.com/quantum-queerness-can-we-causally-explain-bipartite-quantum-entangled-systems/`

Iranzo-Ribera, N. (2022). Scientific counterfactuals as make-believe. *Synthese*, *200*(6), 473.

Ismael, J., & Schaffer, J. (2020). Quantum holism: nonseparability as common ground. *Synthese*, *197*(10), 4131–4160.

Jansson, L. (2014). Causal Theories of Explanation and the Challenge of Explanatory Disagreement. *Philosophy of Science*, *81*(3), 332–348.

Jansson, L. (2018). When are Structural Equation Models Apt? Causation versus Grounding. In A. Reutlinger & J. Saatsi (Eds.), (Vol. 1, pp. 250–266). Oxford: Oxford University Press.

Jansson, L., & Saatsi, J. (2019). Explanatory Abstractions. *The British Journal for the Philosophy of Science*, *70*(3), 817–844.

Jenkins, C. S. (2010). Concepts, Experience and Modal Knowledge. *Philosophical Perspectives*, *24*(1), 255–279.

Jenny, M. (2018). Counterpossibles in Science: The Case of Relative Computability: Counterpossibles in Science: The Case of Relative Computability. *Noûs*, *52*(3), 530–560.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference and consciousness*. Cambridge [Cambridgeshire], New York: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK ; Hillsdale, USA: L. Erlbaum Associates.

Jones, M. R. (2005). Idealization and Abstraction: A Framework. In M. R. Jones & N. Cartwright (Eds.), *Idealization XII: Correcting the Model: Idealization and Abstraction in the Sciences* (pp. 173–217). New York: Editions Rodopi.

Kadanoff, L. (1971). Critical Behaviour: Universality and Scaling. In M. Green (Ed.), (pp. 100–117). New York: Academic Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982). The Simulation Heuristic. In *Judgement under Uncertainty: Heuristics and Biases* (pp. 201–208). New York: Cambridge University Press.

Khalifa, K., Doble, G., & Millson, J. (2020). Counterfactuals and Explanatory Pluralism. *The British Journal for the Philosophy of Science*, *71*(4), 1439–1460.

Kiiveri, H., & Speed, T. P. (1982). Structural Analysis of Multivariate Data: A Review. *Sociological Methodology*, *13*, 209.

Kim, S., & Maslen, C. (2006). Counterfactuals as short stories. *Philosophical studies*, *129*(1), 81-117.

Kimpton-Nye, S. (2020). Necessary Laws and the Problem of Counterlegals. *Philosophy of Science*, *87*(3), 518–535.

Kind, A., & Kung, P. (Eds.). (2016). *Knowledge through imagination* (1st ed.). Oxford University Press.

Knox, E. (2016). Abstraction and its Limits: Finding Space For Novel Explanation: Abstraction and its Limits: Finding Space For Novel Explanation. *Noûs*, *50*(1), 41–60.

Knox, E. (2019). Physical relativity from a functionalist perspective. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *67*, 118–124.

Krickel, B. (2018). Saving the mutual manipulability account of constitutive relevance. *Studies in History and Philosophy of Science Part A*, *68*, 58–67.

Kripke, S. A. (1980). *Naming and Necessity*. Cambridge, Massachussets: Harvard University Press.

Ladyman, J., & Ross, D. (2007). *Every thing must go: metaphysics naturalized*. Oxford, New York: Oxford University Press.

Lange, M. (2000). *Natural laws in scientific practice*. Oxford New York: Oxford University Press.

Lange, M. (2009). *Laws and lawmakers: science, metaphysics, and the laws of nature*. Oxford; New York: Oxford University Press.

Lange, M. (2013). What Makes a Scientific Explanation Distinctively Mathematical? *The British Journal for the Philosophy of Science*, *64*(3), 485–511.

Lange, M. (2016). *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. New York: Oxford University Press.

Levy, A. (2015). Modeling without models. *Philosophical Studies*, *172*(3), 781–798.

Levy, A. (2021). Idealization and abstraction: refining the distinction. *Synthese*, *198*(S24), 5855–5872.

Levy, A., & Godfrey-Smith, P. (Eds.). (2020). *The Scientific Imagination* (1st ed.). Oxford University Press.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Noûs*, *13*(4), 455–476.

Lewis, D. K. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. K. (1983, December). New work for a theory of universals. *Australasian Journal of Philosophy*, *61*(4), 343–377.

Lewis, D. K. (1986a). Causation. In *Philosophical Papers* (Vol. 2, pp. 159–213). Oxford: Oxford University Press.

Lewis, D. K. (1986b). Events. In *Philosophical papers, Vol. II* (Vol. 2, pp. 241–269). Oxford: Oxford University Press.

Lewis, D. K. (1986c). *Philosophical papers, Vol. II*. Oxford: Oxford University Press.

Lewis, D. K. (2000). Causation as Influence. *Journal of Philosophy*(97), 182–197.

Lewis, D. K. (2009). Ramseyan Humility. In Braddon-Mitchell & R. Nola (Eds.), *Conceptual Analysis and Philosophical Naturalism* (pp. 203–222). Cambridge, Massachussets: MIT Press.

Lima, F. M. S., & Arun, P. (2006). An accurate formula for the period of a simple pendulum oscillating beyond the small angle regime. *American Journal of Physics*, *74*(10), 892–895.

List, C. (2019). Levels: Descriptive, Explanatory, and Ontological. *Noûs*, *53*(4), 852–883.

List, C. (2022). Special-science counterfactuals. *The Monist*, *105*(2), 194-213.

Locke, D. (2012). Quidditism without quiddities. *Philosophical Studies*, *160*, 345–363.

Lowe, E. J. (2012). What is the Source of Our Knowledge of Modal Truths? *Mind*, *121*(484), 919–950.

Lévy-Leblond, J.-M. (1976). One more derivation of the Lorentz transformations. *American Journal of Physics*(44), 271–277.

Mallozzi, A. (2021). Two notions of metaphysical modality. *Synthese*, *198*(S6), 1387–1408.

Maudlin, T. (2007). *The metaphysics within physics*. Oxford, New York: Oxford University Press.

Maudlin, T. (2020). A Modal Free Lunch. *Foundations of Physics*, *50*(6), 522–529.

McKenzie, K. (2020). A Curse on Both Houses: Naturalistic Versus A Priori Metaphysics and the Problem of Progress. *Res Philosophica*, *97*(1), 1–29.

McLoone, B. (2019). Thumper the Infinitesimal Rabbit: A Fictionalist Perspective on Some "Unimaginable" Model Systems in Biology. *Philosophy of Science*, *86*(4), 662–671.

McLoone, B. (2020). Calculus and counterpossibles in science. *Synthese*, *198*(12), 12153–12174.

McLoone, B., Grützner, C., & Stuart, M. T. (2023). Counterpossibles in science: an experimental study. *Synthese*, *201*(1), 27.

McMullin, E. (1985). Galilean Idealisations. *Studies in History and Philosophy of Science*, *16*(3), 247–273.

Meehl, P. E., & Scriven, M. (1956). The concept of emergence. In *Minnesota Studies in the Philosophy of Science* (Vol. 1, pp. 239–252). University of Minnesota Press.

Meek, C., & Glymour, C. (1994). Conditioning and Intervening. *The British Journal for the Philosophy of Science*, *45*(4).

Menzies, P. (1989). Probabilistic Causation and Causal Processes: A Critique of Lewis. *Philosophy of Science*(56), 642–663.

Menzies, P. (1998). Possibility and conceivability: A response-dependent account of their connections. In R. Casati (Ed.), *Europen Review of Philosophy, Volume 3: Response-Dependence* (pp. 255–277). Stanford: Csli Publications.

Menzies, P., & Price, H. (1993). Causation as a Secondary Quality. *The British Journal for the Philosophy of Science*, *44*(2), 187–203.

Messer, L. C. (2016). *Natural experiment | observational study.* Retrieved from `https://www.britannica.com/science/natural-experiment`

Morgan, M. S. (2004). Imagination and Imaging in Model Building. *Philosophy of Science*, *71*(5), 753–766.

Morrisson, M. (1999). Models as autonomous agents. In M. S. Morgan & M. Morrisson (Eds.), *Models as mediators: perspectives on natural and social sciences* (pp. 38–65). Cambridge: Cambridge University Press.

Mumford, S. (2004). *Laws in nature*. London, New York: Routledge.

Murphy, A. (2020). Toward a Pluralist Account of the Imagination in Science. *Philosophy of Science*, *87*(5), 957–967.

Myers, J. (2021). The epistemic status of the imagination. *Philosophical Studies*, *178*(10), 3251–3270.

Ney, A., & Albert, D. (Eds.). (2013). *The wave function: Essays in the metaphysics of quantum mechanics*. New York: Oxford University Press.

Nguyen, J., & Frigg, R. (2022). *Scientific Representation* (1st ed.). Cambridge University Press.

Nichols, S., & Stich, S. P. (2003). *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds*. Oxford: New York: Clarendon ; Oxford ; Oxford University Press.

Näger, P. M. (2016). The causal problem of entanglement. *Synthese*, *193*(4), 1127–1155.

of Encyclopaedia Britannica, T. E. (1998). *Kepler's laws of planetary motion | astronomy.* Retrieved 2022-08-04, from https://www.britannica.com/science/Keplers-laws-of-planetary-motion

Oppenheim, P., & Putnam, H. (1958). Unity of Science as a Working Hypthesis. *Minnesota Studies in the Philosophy of Science*(2), 3–36.

Pearl, J. (1988). Embracing causality in default reasoning. *Artificial Intelligence*, *35*(2), 259–271.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge, U.K.; New York: Cambridge University Press.

Physics, O. U. (2016). *Tidal Force, University Physics Volume 1, fig. 13.23.* Retrieved from https://commons.wikimedia.org/wiki/File:CNX_UPhysics_13_06_TidalForce.png

Pincock, C. (2012). *Mathematics and scientific representation*. New York: Oxford University press.

Pincock, C. (2015). Abstract Explanations in Science. *The British Journal for the Philosophy of Science*, *66*(4), 857–882.

Plutynski, A. (2006). Strategies of Model Building in Population Genetics. *Philosophy of Science*, *73*(5), 755–764.

Potochnik, A. (2010). Explanatory Independence and Epistemic Interdependence: A Case Study of the Optimality Approach. *The British Journal for the Philosophy of Science*, *61*(1), 213–233.

Price, H. (2004). Models and Modals. In *Laws and Models in Science.* King's College Publications.

Price, H. (2007). Causal Perspectivalism. In H. Price & R. Corry (Eds.), *Causation, Physics, and the Constitution of Realty: Russell's Republic Revisited* (pp. 250–292). Oxford: Oxford University Press.

Prior, E. W., Pargetter, R., & Jackson, F. (1982). Three Theses about Dispositions. *American philosophical quarterly (Oxford)*, *19*(3), 251–257.

Psillos, S. (2004). A Glimpse of the Secret Connexion: Harmonizing Mechanisms with Counterfactuals. *Perspectives on Science*, *12*(3), 288–319.

Psillos, S. (2011). Living with the abstract: realism and models. *Synthese*, *180*(1), 3–17.

Putnam, H. (1972). The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*(7), 181–193. (Reprinted in Putnam (1975))

Putnam, H. (1975). The Meaning of 'Meaning'. In *Mind, Language, and Reality* (pp. 215–271). Cambridge: Cambridge University Press.

Redhead, M. (1995). *From physics to metaphysics*. Cambridge; New York: Cambridge University Press.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of California Press.

Reutlinger, A. (2011). A Theory of Non-universal Laws. *International Studies in the Philosophy of Science*, *25*(2), 97–117.

Reutlinger, A. (2013). Getting Rid of Interventions. In *A Theory of Causation in the Social and Biological Sciences* (pp. 104–124). London: Palgrave Macmillan UK.

Reutlinger, A. (2014). Why Is There Universal Macrobehavior? Renormalization Group Explanation as Noncausal Explanation. *Philosophy of Science*, *81*(5), 1157–1170.

Reutlinger, A. (2016). Is There A Monist Theory of Causal and Noncausal Explanations? The Counterfactual Theory of Scientific Explanation. *Philosophy of Science*, *83*(5), 733–745.

Reutlinger, A., & Andersen, H. (2016). Abstract versus Causal Explanations? *International Studies in the Philosophy of Science*, *30*(2), 129–146.

Rice, C. (2015). Moving Beyond Causes: Optimality Models and Scientific Explanation: Moving Beyond Causes. *Noûs*, *49*(3), 589–615.

Robertson, K. (forthcoming). Autonomy Generalized; or, Why Doesn't Physics Matter More? *Ergo*.

Robinson, D. (1993). Epiphenomenalism, laws & properties. *Philosophical Studies*(69), 1–34.

Romero, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese*, *192*(11), 3731–3755.

Rosen, G. (2002). A study of modal deviance. In T. S. Gendler & J. Hawthorne (Eds.), *Conceivability and possibility* (pp. 283–307). Oxford: Clarendon.

Russell, B. (1912). On the Notion of Cause. *Proceedings of the Aristotelian Society*, *13*, 1–26.

Russell, B. (1927). *The analysis of matter*. London: Kegan Paul.

Saatsi, J., & Pexton, M. (2013). Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations. *Philosophy of Science*, *80*(5), 613–624.

Saatsi, J., & Reutlinger, A. (2018). Taking Reductionism to the Limit: How to Rebut the Antireductionist Argument from Infinite Limits. *Philosophy of Science*, *85*(3), 455–482.

Salis, F. (2020). The New Fiction View of Models. *The British Journal for the Philosophy of Science*, *72*(3), 717–742.

Salis, F., & Frigg, R. (2020). Capturing the Scientific Imagination. In A. Levy & P. Godfrey-Smith (Eds.), *The Scientific Imagination* (pp. 17–50). New York: Oxford University Press.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Salmon, W. (1998). *Causality and Explanation*. Oxford: Oxford University Press.

Schaffer, J. (2005). Quiddistic knowledge. *Philosophical Studies*(123), 1–32.

Schaffer, J. (2016a). Grounding in the image of causation. *Philosophical Studies*, *173*(1), 49–100.

Schaffer, J. (2016b). *The Metaphysics of Causation.* Metaphysics Research Lab, Stanford University. Retrieved from `https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/` (2016b)

Scheibe, E. (2001). General Laws of Nature and the Uniqueness of the Universe. In B. Falkenburg (Ed.), *Erhard Scheibe: Between Rationalism and Empiricism. Selected papers in the philosophy of physics* (pp. 276–288). New York: Springer. (first edition: 1991)

Schelling, T. C. (1971). Dynamic models of segregation†. *The Journal of Mathematical Sociology*, *1*(2), 143–186.

Schroeter, L. (2019). *{Two-Dimensional Semantics}* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2019/entries/two-dimensional-semantics/

Shech, E. (2018). Infinite idealizations in physics. *Philosophy Compass*, *13*(9), e12514.

Shoemaker, S. (1984). *Identity, Cause, and Mind: Philosophical Essays*. Cambridge: Cambridge University Press.

Sider, T. (2011). *Writing the book of the world*. Oxford, New York: Oxford University Press, Clarendon Press.

Smith, D. C. (2016). Quid Quidditism Est? *Erkenntnis*, *81*(2), 237–257.

Sober, E. (2011). *A Priori* Causal Models of Natural Selection. *Australasian Journal of Philosophy*, *89*(4), 571–589.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, Massachussets: MIT Press. (first edition: 1993)

Stalnaker, R. C. (1968). A Theory of Conditionals. In N. Rescher (Ed.), *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)* (pp. 98–112). Blackwell.

Starr, W. (2021). Counterfactuals. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/counterfactuals/.

Statham, G. (2018). Woodward and variable relativity. *Philosophical Studies*, *175*(4), 885–902.

Stoljar, D. (2001). Two Conceptions of the Physical. *Philosophy and Phenomenological Research*, *62*(2), 253–281.

Strevens, M. (2008). *Depth: an account of scientific explanation*. Cambridge, Massachussets: Harvard University Press.

Strevens, M. (2019). The structure of asymptotic idealization. *Synthese*, *196*(5), 1713–1731.

Strotz, R. H., & Wold, H. O. A. (1960). Recursive vs. Nonrecursive Systems: An Attempt at Synthesis (Part I of a Triptych on Causal Chain Systems). *Econometrica*, *28*(2), 417.

Sugden, R. (2009). Credible Worlds, Capacities and Mechanisms. *Erkenntnis*, *70*(1), 3–27.

Sullivan, E. (2019). Universality caused: the case of renormalization group explanation. *European Journal for Philosophy of Science*, *9*(3), 36.

Suppes, P. (1970). *A probabilistic Theory of Causation*. Amsterdam: North Holland Publishing.

Suárez, M. (Ed.). (2009). *Fictions in science: philosophical essays on modeling and idealization* (1. publ ed.) (No. 4). New York, NY: Routledge.

Tahko, T. E. (2011). In defence of Aristotelian metaphysics. In T. E. Tahko (Ed.), *Contemporary Aristotelian Metaphysics* (1st ed., pp. 26–43). Cambridge University Press.

Tan, P. (2017). Interventions and Counternomic Reasoning. *Philosophy of Science*, *84*(5), 956–969.

Tan, P. (2019). Counterpossible Non-vacuity in Scientific Practice. *The Journal of Philosophy*, *116*(1), 32–60.

Taylor, E. (2021). Explanatory Distance. *The British Journal for the Philosophy of Science*, 714–788.

Taylor, H. (2017). Powerful qualities, the conceivability argument and the nature of the physical. *Philosophical Studies*, *174*(8), 1895–1910.

Taylor, J. E. (1976). The Structure of Singularities in Soap-Bubble-Like and Soap-Film-Like Minimal Surfaces. *The Annals of Mathematics*, *103*(3), 489–539.

Teller, P. (2009). Fictions, Fictionalization, and Truth in Science. In M. Suárez (Ed.), *Fictions in Science: Philosophical Essays on Modelling and Idealization* (1st ed., p. 13). Routledge.

Thomasson, A. L. (1999). *Fiction and metaphysics*. Cambridge: Cambridge University Press.

Toon, A. (2012). *Models as make-believe: imagination, fiction, and scientific representation*. New York: Palgrave Macmillan.

Vaidya, A. (2008). Modal Rationalism and Modal Monism. *Erkenntnis*, *68*(2), 191–212.

Van Cleve, J. (1983). Conceivability and the Cartesian argument for dualism. *Pacific Philosophical Quarterly*, *64*(1), 35–45.

Van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford; New York: Oxford University Press.

Van Inwagen, P. (1998). Modal Epistemology. *Philosophical Studies*, *92*(1), 67–84.

von Wright, G. H. (1971). *Explanation and Understanding*. Ithaca, NY: Cornell University Press.

Wallace, D. (2011). The Logic of the Past Hypothesis.. (To appear in B. Loewer, E. Winsberg and B. Weslake (ed.) *The Probability Map of the Universe: Essays on David Albert's "Time and Chance,"* publication date: 01/10/2023)

Walton, K. L. (1990). *Mimesis as make-believe: on the foundations of the representational arts*. Cambridge, Massachussets: Harvard University Press.

Weisberg, M. (2007a). Who is a Modeler? *The British Journal for the Philosophy of Science*, *58*(2), 207–233.

Weisberg, M. (2007b). Three Kinds of Idealization. *The Journal of Philosophy*, *104*(12), 639–659.

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

Wigner, E. P. (1967). *Symmetries and reflections: scientific essays of Eugene P. Wigner*. Bloomington and London: Indiana University Press.

Williamson, T. (2007). *The philosophy of philosophy* (No. 2). Malden, Massachussets: Blackwell Publishing.

Williamson, T. (2016). Modal science. *Canadian Journal of Philosophy*, *46*(4-5), 453–492.

Wilson, A. (2018). Grounding Entails Counterpossible Non-Triviality. *Philosophy and Phenomenological Research*, *96*(3), 716–728.

Wilson, A. (2021). Counterpossible Reasoning in Physics. *Philosophy of Science*, *88*(5), 1113–1124.

Wolff, J. (2013). Are Conservation Laws Metaphysically Necessary? *Philosophy of Science*, *80*(5), 898–906.

Wood, C. J., & Spekkens, R. W. (2015). The lesson of causal discovery algorithms for quantum correlations: causal explanations of Bell-inequality violations require fine-tuning. *New Journal of Physics*, *17*(033002), 1-29.

Woodward, J. (2000). Explanation and Invariance in the Special Sciences. *The British Journal for the Philosophy of Science*, *51*(2), 197–254.

Woodward, J. (2003). *Making things happen: a theory of causal explanation*. New York: Oxford University Press.

Woodward, J. (2004). Counterfactuals and Causal Explanation. *International Studies in the Philosophy of Science*, *18*(1), 41–72.

Woodward, J. (2007). Causation with a Human Face. In H. Price & R. Corry (Eds.), (pp. 45–65). Oxford, New York: Oxford University Press, Clarendon Press.

Woodward, J. (2010a). Agency and Interventionist Theories. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford Handbook of Causation* (1st ed., pp. 234–263). New York: Oxford University Press. (2010a)

Woodward, J. (2010b). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*(3), 287–318.

Woodward, J. (2013). Laws, causes, and invariance*. In S. Mumford & M. Tugby (Eds.), *Metaphysics and Science* (pp. 48–72). Oxford University Press.

Woodward, J. (2014). A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment). *Philosophy of Science*, *81*(5), 691–713.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, *91*(2), 303–347.

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*(4), 1047–1072.

Woodward, J. (2017). Interventionism and the Missing Metaphysics. In M. Slater & Z. Yudell (Eds.), *Metaphysics and the Philosophy of Science* (pp. 193–228). Oxford University Press.

Woodward, J. (2018a). Laws: An Invariance-Based Account. In W. Ott & L. Patton (Eds.), *Laws of Nature* (Vol. 1, pp. 158–180). Oxford: Oxford University Press.

Woodward, J. (2018b). Some Varieties of Non-Causal Explanation. In A. Reutlinger & J. Saatsi (Eds.), *Explanation beyond causation: philosophical perspectives on non-causal explanations* (Vol. 1, pp. 117–138). Oxford: Oxford University Press.

Woodward, J. (2020). Physical modality, laws, and counterfactuals. *Synthese*(5), 1907–1929.

Woodward, J. (2021). *Causation with a Human Face: Normative Theory and Descriptive Psychology* (1st ed.). New York: Oxford University Press.

Woodward, J., & Hitchcock, C. (2003). Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs*, *37*(1), 1–24.

Woodward, J., Loewer, B., Carroll, J. W., & Lange, M. (2011). Counterfactuals all the way down? Marc Lange: Laws and lawmakers: Science, metaphysics, and the laws of nature. New York: Oxford University Press, 2009, 280 pp. *Metascience*, *20*(1), 27–52.

Yablo, S. (1992). Mental Causation. *The Philosophical Review*, *101*, 245.

Yablo, S. (1993). Is Conceivability a Guide to Possibility? *Philosophy and Phenomenological Research*, *53*(1), 1-42.

Zeh, H. D. (1999). *The physical basis of the direction of time* (3rd ed.). Berlin, New York: Springer.