

Causation and the Objectification of Agency

Christoph Schulz

Submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of PhD

February 2015

Table of contents

1	Introduction	6
2	Outline of the thesis.....	10
3	What is causation – the conceptual question.....	15
3.1	Levels of abstraction	19
3.1.1	Levels of abstraction for causal judgments	21
3.1.2	The limits of the method of levels of abstraction	22
4	Theories of causation.....	24
4.1	The choice of <i>difference-making</i> for addressing the conceptual problem.....	32
4.2	Analysis of Markovian models of causation.....	35
4.2.1	The causal Markov condition	35
4.2.2	The problems of Causal Bayesian Networks and the Causal Markov Condition	38
4.2.3	The epistemic problems.....	39
4.2.4	The ontological and conceptual problems.....	44
4.2.5	Criticism of the objectivist difference-making theories.....	52
4.3	Problems and objections against the agency-theory	59
4.3.1	Circularity	59
4.3.2	Anthropomorphism	59
4.3.3	Two additional concerns.....	59
5	The Problem of Circularity	63
5.1	Circularity of ‘bringing about’	63
5.2	A parallel analysis – Becoming informed	70
6	Thermodynamics of acting and information processing	74
6.1	Causation by information	76
6.1.1	Intuitions concerning mechanistically and informationally driven processes.....	76
6.2	Natural agency and its properties.....	82
6.2.1	Thermodynamics, spontaneous processes, and entropy	83
6.2.2	The connection between thermodynamics and causation.....	84
6.2.3	Thermodynamics and actions	86
6.2.4	Maxwell’s demon as paradigmatic causal agent	88
6.2.5	Internal computational model	93
6.2.6	Causation by information revisited.....	96
6.2.7	Concept acquisition.....	99
6.2.8	Free action	102

6.2.9	Concept identification.....	103
6.3	The properties of causation: asymmetry, locality, regularity.....	106
6.3.1	Asymmetry.....	109
6.3.2	Locality.....	110
6.3.3	Regularity.....	111
7	Causal judgments in LoA1, LoA2, and LoA3.....	115
7.1	Objectification - Switching from LoA1 to LoA3.....	119
7.1.1	Causal judgements concerning particular past events.....	122
7.1.2	Counterexamples against chance-raising.....	129
7.1.3	Over-determination.....	131
7.1.4	Counterexamples from causal diagrams.....	134
7.2	Different constraint-levels inform different causal concepts.....	138
8	Summary, final conclusion, and outlook.....	141
8.1	Summary.....	141
8.2	Final Conclusion and outlook.....	142
9	Bibliography.....	144

List of abbreviations

CBN	–	causal Bayesian networks
CMC	–	causal Markov condition
CM	–	causal Markov condition in Woodward and Hausman’s weaker formulation
CI	–	constraint level of coefficient invariance
EI	–	concept of intervention (James Woodward’s version)
LI	–	constraint level of level invariance
LoA	–	level of abstraction
MD	–	constraint level of modularity
MOA	–	method of abstraction
MOD	–	the modularity thesis (the connection between manipulability and CMC)
PI	–	concept of intervention (Judea Pearl’s version)
PSE	–	principle of precise specification of events
SEM	–	structural equation model
TLC	–	token level cause, according to agency-theory as proposed in this thesis

Abstract

This dissertation defends the so-called 'agency-approach' to causation, which attempts to ground the causal relation in the cause's role of being a means to bring about its effect. The defence is confined to a *conceptual* interpretation of this theory, pertaining to the concept of causation as it appears in a causal judgement. However, causal judgements are not seen as limited to specific domains, and they are not exclusively attributed to human agents alone. As a methodological framework to describe the different perspectives of causal judgments, a method taken from the philosophy of information is made use of – the so-called 'method of abstraction'. According to this method, levels of abstraction are devised for the subjective perspective of the acting agent, for the agent as observer during the observation of other agents' actions, and for the agent that judges efficient causation. As a further piece of propaedeutic work, a class of similar (yet not agency-centred) approaches to causation is considered, and their modelling paradigms – Bayesian networks and interventions objectively construed – will be criticised. The dissertation then proceeds to the defence of the agency-approach, the first part of which is a defence against the objection of conceptual circularity, which holds that agency analyses causation in causal terms. While the circularity-objection is rebutted, I rely at that stage on a set of subjective concepts, i.e. concepts that are eligible to the description of the agent's own experience while performing actions. In order to give a further, positive corroboration of the agency-approach, an investigation into the natural origins and constraints of the concept of agency is made in the central chapter six of the dissertation. The thermodynamic account developed in that part affords a third-person perspective on actions, which has as its core element a cybernetic feedback cycle. At that point, the stage is set to analyse the relation between the first- and the third-person perspectives on actions previously assumed. A dual-aspect interpretation of the cybernetic-thermodynamic picture developed in chapter six will be directly applied to the levels of abstraction proposed earlier. The level of abstraction that underpins judgments of efficient causation, the kind of causation seemingly devoid of agency, will appear as a derived scheme produced by and dependent on the concept of agency. This account of efficient causation, the 'objectification of agency', affords the rebuttal of a second objection against the agency-approach, which claims that the approach is inappropriately anthropomorphic. The dissertation concludes with an account of single-case, or token level, causation, and with an examination of the impact of the causal concept on the validity of causal models.

1 Introduction

This dissertation is the product of a research project whose goal was to deliver an examination of the concept of causation. I have understood that examination explicitly as *conceptual*, in contrast to providing an ontological account of causation, or searching for solutions concerning more specific problems of the epistemology of causation, which take the content of causation for granted to some extent. I had started the project in the way a classical project of conceptual analysis would have been conducted, by providing continuous refinements of an initial analysis that belongs to a certain class of approaches to causation, and to which one would have to commit at the outset. The continuous refinements would be due to the pressure put on an initial attempt of analysis by conceived counterexamples, which takes it for granted that there is a univocal intuition that provides the verdict in those cases, and whose results the theory is supposed to capture. Some philosophical literature on causation still proceeds like that, although it rarely refers to itself as 'conceptual analysis'. My method does not consist in a continuous refinement of some stage of the account in the light of examples that more and more inform the account. Instead, the initial attempt captures the idea very generally: when we apply the concept of causation we actually apply the concept of agency. As a further constraint of that identity, I add that applying the concept refers to its use in a judgment that can be made explicit. After this identity will have been defended against some typical objections, whereby it will be demonstrated that such an account is *possible*, a naturalistic turn is then taken in order to search for the most likely origins of the concept of agency, in order to show that the account is also *plausible*. But why turn to an explicitly conceptual understanding of an analysis of causation? The downsides are clearly that

- a) it looks like conceptual analysis as commonly understood, and therefore will most likely ultimately result in a failure (even more likely because causation is such a basic idea that seems to resist further analysis);
- b) the analysis, if explicitly understood in contrast to how its object is ontologically realised, might produce an *unconstrained concept*;
- c) the result might be subjective, which is again especially worrying for something supposedly 'objective' like causation.

But also the advantages are very obvious:

- 1) The ontology and the epistemology of causation are hard problems. Humanity has been thinking about them for two and a half millennia now. By means of a distinct conceptual understanding of the question we might be able to break down the difficulty a little bit, by distinguishing certain aspects of causation from cognate concepts like regularities, laws of nature, determination, dependence, etc. Such an analysis is not unconstrained at all. If some specific and practical problems can be solved without addressing the ontological aspects, then these aspects maybe do not play such a crucial role for *all* of the aspects of causation, or maybe it does play a role only for one of the *other* concepts from the above family of concepts germane to causation.
- 2) In the debate on causation, sometimes one has the impression of what Dennett (1991), in the context of phenomenology, refers to as a situation where 'controversies degenerate into desk-thumping cacophony, with everybody talking past everybody else.' Although it might sometimes be wrong-headed to proceed according to a clear-cut sequence of conceptualising the object of study in order to then go into the field and see whether the concept is vacuous or not, I think the least one could say about causation is that it would be worth mapping the conceptual territory

and taking serious at least the not obviously inconsistent interpretations of causation. Then one can properly address elements of this map and make a point more precisely, such as expounding a concern of internal coherence, or worries that the term is sound but has no corresponding object in nature. Also, the structure of such a map might tell a story of how the different ways of understanding causation are related to each other. Is it really the case that they are mutually incoherent? Are capacity-based approaches in conflict with manipulation theories, and do counterfactual approaches rule out nomological accounts?

As it will be clear in the course of the chapters, I think of my account rather as reconstruction than analysis (although I see no substantial problem in referring to it by the term 'analysis'). I would also feel comfortable with saying that analysis – analysing how the concept of causation is used – *guides* the conceptual reconstruction. As far as the ontological questions are concerned, the kind of examination that I am offering will have little to do with them. As an example, there is the question whether there are real causal capacities of objects, which give rise to observable regularities by virtue of the objects' arrangements, or, reversing the explanatory direction, whether there are real causal laws of nature that merely make us ascribe certain dispositions to objects. My analysis will not cut that deep into such metaphysical issues.

The account, although I spend a large section on non-circularity, also presupposes some realities that are sometimes discussed in the causal context. I will not analyse where the asymmetry of time stems from. Hence I will not expound the problem of whether the direction of time is independent of the direction of thermodynamic entropy and of causation or is just given by the way causal agents usually happen to reason. Besides that, I rely on the notion of 'dependence' in the logical sense, and thus take it as a non-causal term that can be used to explain causal matters without danger of circularity. And I take the reality of thermodynamics for granted, according to which we are able to tell apart the unfolding of events in forward and in reversed playback while watching video footage. My analysis of causation will build on this unexplained phenomenon. All these assumptions can be attacked, however, and one could claim that causation plays a role in their grounding. But the role causation plays in my theory does not have this wide a range. It rather resembles a more 'light-weight' concept like those expounded in Russell (1913), or more recently in Norton (2003) and also in Maudlin (2007). I treat it as an epiphenomenal concept that can be explained in terms of agency, and agency in turn can be embedded into a thermodynamic framework.

Thus, while abstracting from some classically ontological questions, there is still a metaphysical backdrop that the account – my specific interpretation of the agency-approach to causation – fits in. This metaphysical backdrop is a cybernetic view of the world, interpreted according to the dual-aspect theory, which refrains from giving precedence to either the subjective or the objective view of the world. According to this backdrop, I consider the world as a system whose entropy rises in time, and which comprises smaller systems, some of which are agents. The reality of these sub-systems as *agents* in this world is made possible by a loophole in physics that allows the unlikely phenomenon of acting by living organisms. Acting is part of a cybernetic feedback cycle (or control-circuit) that defines the *identity* of an agent, and the locality that the second law of thermodynamics imposes on that structure defines the *boundaries* of the agent. The material input into the cybernetic structure is free energy, and it is the asset that allows the agent to maintain its unlikely structure. But the input can also be informational feedback, in which case the agent learns something, which is tantamount to updating its model of its environment. The model, and the

contingent information that informs a particular output of the model, are necessary for acting sensibly and for exerting control. In an entropic world there is nothing like uninformed acting if the actions are supposed to yield a result. Since a single agent is not the only player operating in that entropic environment, other agents seeking their advantage also have to be taken into account in the agent's model. When we see *causes* at work, we apply our conceptual scheme of acting, and causing becomes a derived concept that depends on acting; this is the central claim of the agency-theory. I interpret the imposition of such a scheme as a *partial* imposition of the scheme of a complete cybernetic feedback cycle, whenever we apply it to objects in order to turn them into 'causes' even when the objects are bereft of the capacity to process the feedback stemming from their imputed 'actions'.

When one analyses what an action is, one has to deal with the question of teleology. A natural teleology is provided by cybernetics by the idea of the purpose of self-preservation: the unlikely structure exists for maintaining the unlikely structure. This can be rephrased directly into a thermodynamic language. In fact, cybernetics and thermodynamics are really two sides of the same coin. I will show how this account of teleology, grounded in entropy reduction, can be used to ground a naturalised account of the concept of action, illustrated by the thought-experiment of Maxwell's demon, which I will adapt in such a way that its agent completes a thermodynamic cycle.

That cybernetic feedback is sometimes informational perfectly fits the purpose of taking into account the reality of higher-order utility: not everything we seek is physical entropy reduction. For example, if a chemist has discovered a substance of which she thinks it might be toxic, and she does not want to test it on other human beings or animals, then she might decide to consume a diluted quantity of that substance herself. It is *not* the well-being of her body that she seeks by doing so, but general knowledge, and the implications for possible, more serious *future* situations that involve that substance. This is the path to explain higher-order utility, while ultimately staying grounded in entropy reduction. Of course, it is not easy to prove the overall consistency of such an account of teleology, but it shows that it would be too easy to rebut this account on the basis of seeing no obvious connection to physical entropy. The issue is similar to money and real values: whereas money promises the prospect of being convertible into consumable goods, information promises to be convertible into adequate behaviour at some later time. (But our existential position makes it sometimes hard to say when and if this is going to happen.) Although *information* plays a central role in this thesis, the account is thus not a pan-informational one.

Next to these remarks concerning ontological aspects, there need to be some clarifications concerning the scope of the *epistemology* of causation considered in this thesis. Although the account is informed to a large extent by thermodynamic considerations, the account is not considered to be a contribution to the philosophy of science, or to the epistemology of causation. The scope of the thermodynamic chapter is limited to the second law of thermodynamics, which has a direct impact on the crucial question of whether a process is spontaneous or non-spontaneous. Since acting will be interpreted as bringing about proximal non-spontaneous processes, there is an immediate impact on action and its concept, too. But the particular causal knowledge that regiments appropriate interventions in specific physical domains or higher-order domains supervening on physics is not analysed. Again, doing so by-passes many epistemological questions which would have to be considered indispensable were the scope not limited to the conceptual question as I construe it in chapter 3.

The thesis can be broadly distinguished into three parts. I will first select the agency-theory of causation as the one best fitting my purposes of conceptual reconstruction, and I will defend it

against the objection of circularity, postponing the objection of anthropomorphism. At that stage, I embrace subjectivity (and with that, to some extent, anthropomorphism). The level of abstraction correspondingly employs subjective variables. I will subsequently make a detour into the thermodynamics of actions and information processing, in order to find an explanation of where the concept of action plausibly comes from. After that excursion, I will have at my disposal further information governing the two other levels of abstraction, i.e. the one for judging *other* agents' actions, and the derived level of abstraction to deal with efficient causation. These three levels of abstraction and their mutual relations are the basis for solving the conceptual problem of causation and action, and each of the three will be shown to be appropriate to answering specific problems connected to causal judgements. In the third part, I will apply the levels of abstraction to concrete examples of causal judgments, and subsequently, questions of causal structure.

2 Outline of the thesis

In this chapter I provide a comprehensive outline of the argument of my thesis. There is also a concise summary in the last chapter that might help the reader further with grasping the overall structure of the argument.

The conceptual problem of causation

The thesis seeks to deliver an answer to the question of whether the notion of causation has an internal structure that can be disclosed. This question shall be understood explicitly at the *conceptual level*. I take it that answering this kind of question amounts to making sense of how an agent could explain a judgment in terms that do not directly involve the term that is to be analysed. What this means in the context of causation will be explained in more detail in chapter 3. The conceptual problem is distinguished sharply from both epistemological and ontological problems of causation. It does not concern the former class, since the account does not ask where the agent has the knowledge from that might be required to form the judgment, or whether the agent is justified in making the judgment. Likewise, the ontological problem of causation is a different one: the agent does not need to have an idea of what *really* connects an event *A* and an event *B*. Instead, the approach seeks an account of a minimal semantics of the causal relation. The argument draws heavily on the idea of perspectives and their impact on judgments. The philosophical tool used to describe different perspectives, namely 'levels of abstraction', will also be introduced in chapter 3, and the three levels relevant to the thesis will be introduced there.

Why agency-account of causation

After providing an overview of theories of causation, I will give, in section 1 of chapter 4, my arguments for why I decided to treat causation as a relation of *difference-making*. In section 2 I will analyse the characteristics of Bayesian networks, an important sub-class of difference-making theories. I will criticise their inherent structural feature of the causal Markov condition, as well as their objectively understood notion of intervention. I will then turn to the agency-theory of causation as the most plausible candidate for a successful conceptual analysis, as this notion is understood here.¹ At that stage, the result is merely one from elimination of alternatives. The choice of agency then needs to be positively defended against two major objections: anthropocentricity and circularity.

Defence of agency-approach

The objection of anthropocentricity will be addressed later on, by the account of 'objectification' (in section 7.1). The objection of circularity will be dealt with in section 5.1. In that section I will provide an explicit account of how the agency-account is to be understood without ending up in a conceptual circle. According to this understanding of the agency-account, physical channels mediate causal influences through time and space, giving rise to physical observables, into which an agent can intervene. This intervention is (unlike interventions in what I call 'Objective Interventionism') *not* a causal relation between an agent and an observable, but is a free and direct action on its target. The idea behind 'free and direct action' on the one hand, and 'correlations between physical observables' on the other hand, is that the model the agent uses to evaluate her causal judgement

¹ This approach looks as if we single out agency as the only viable theory of causation. But this singling out concerns only the eligibility of a theory for my specific approach to the conceptual analysis of causation, viz. the singling-out is relative to the question I ask.

(thereby delivering an answer to the conceptual question of causation) is the same model that excludes determining antecedents to the agent's decision, which is thus rendered 'free' from the perspective of the agent. The model likewise excludes intermediate causal variables between the decision to act and the target of the intervention, therefore rendering the action 'direct'. An analogous reasoning (section 5.2) is applied to the process of information transfer and an agent's 'becoming informed of a contingent matter of fact'. This account has the flow of information depend on physical correlations, and has *perception* consist in a direct interaction with a physical signal carrying information about a prior event. At this stage, we have an account of a *binary* causal judgment that allows a certain level of analysis on the basis of agency and correlations of physical observables. The approach so far has been only logical; it was presupposed that the concepts of causation, agency, information, etc. were already given and meaningful, such that their dependence-relations could be highlighted.

Natural origins of the concept of agency

An assumption made in section 5.1 implies that causal flow and informational flow are physically identical. Thus it seems plausible that information can cause something, and that this variant of causation has something to do with actions. Such a view would tie the informational perspective (looking at physical differences that allow inferences to what has happened at an earlier time) to the agency-perspective (considering differences that make some difference at a later time). Binding the two views together alongside their respective semantic aspects (semanticisation and direct action respectively), yields the composite concept of 'causation by information'. If the agent explains her own action *as* an action, then the action is judged to be constrained by information, while the action turns into an event if the two processes of semanticisation of information and making a decision based on that information are supplanted by a physical model of efficient causation. The composite concept of causation by information, however, hinges on the two notions of semanticisation and direct action, which have, at that stage, a kind of dubious status, since they involve unobservable *relata* and depend on a first-person perspective. By looking at examples of natural action we will want to corroborate that actions are indeed informationally triggered, and whether this is necessarily so.

There are thus several motivations to turn to an examination of the natural origins of agency and its concept, a lead that I will follow in section 6.2. These motivations are:

- 1: We need to understand better the phenomenon of causation by information.
- 2: We need a plausible account of concept acquisition, since an account of concept acquisition by simple ostension, as suggested by Huw Price and Peter Menzies, is wanting. One of the problems of ostension is to explain how an agent recognizes her behaviour *as* an action, rather than an event.
- 3: We need a better account of how the notion of a 'free action', a notion that is required by the agency-account, could arise in the first place, and how it can be reconciled with actions that are constrained by information transfer – a seeming contradiction in terms.
- 4: We need to get further information about agency. The previous logical analysis left us with quite a thin causal concept that relied on the notions of actions and observable correlations. We now need to gather further information about agency that could, in turn, also constrain our concept of causation. Also, we will want to reconcile the agency-account with accounts of causation that do not

incorporate a free agent's action explicitly, and we will want to be able to cover causal judgments that go beyond the case of a binary causal connection between A and B.

These four points motivate taking a closer look at biological agents and how they are constrained in their acting. This, in turn, leads to an excursion into the physics (thermodynamics in particular) of acting in sub-sections 6.2.1 to 6.2.3. According to the stance of physics, all biological agents have to regularly bring about *non-spontaneous* processes to maintain their structure through time, among achieving other, higher-order goals. I will consider bringing about non-spontaneous processes as goals which acting typically seeks to achieve. After expounding the importance of that idea for the further course of the argument, the remainder of section 6.2 will concern the results of asking the questions that correspond to motivation 1 to 4, as given above. An overview over the respective results is given by the paragraphs below:

Concerning 1: Causation by information is first introduced as a composite concept. If an agent interacts with a signal that carries some meaning for the agent, this prompts the agent to take action. Seen from the first-person perspective, there is nothing surprising in that concept. But to judge in purely observational terms whether an informationally triggered action has taken place is a harder question. How do we tell that it is the information carried by the physical signal and not the physical features of that signal that do the causing? *Maxwell's demon* will serve as a paradigmatic model of an agent, and the corresponding scenario as a model for agency. The sorting operations, which the demon performs, are basic, binary decisions resulting in an observable intervention into the physical world. It will be shown that these actions are necessarily informationally constrained, and that the information constraining the action is necessarily physically embodied. There are two ways to interpret what we see when the demon is performing the work: either the demon abides by laws of efficient causation, or by a computational rule, which affords an interpretation according to the concept of causation by information. The explanation resulting from these considerations will serve as a general template to explain the phenomenon of causation by information in observational terms.

Concerning 2: The thermodynamics of informationally constrained acting will show that there is a strong pressure for biological entities to develop the concept of agency (and then, according to the hypothesis, causation), in order to coordinate actions or plan future actions from an internal model. Whereas the necessity of informed anti-entropic acting already applies to the lowest of biological agents, higher-level biological agents soon have to incorporate information about other agents and their doings into their internal models, such that the agent's competitiveness is not just limited to entropic environments, but extends to environments in which different agents compete for limited resources. Such more complex scenarios require acting to be informed just like entropic environments do.

Concerning 3: The preceding sections will have explained how information can, at least to some extent, determine behaviour, while they still allow for the construal of an action *as* an action, rather than an event. As opposed to this, a 'free action' is a realisation of several possibilities by an agent *not* constrained by preceding conditions. In the present framework, choices will be interpreted as possibly constrained by determinants; however, these determinants are not represented in the agent's own internal model. Therefore, the free decision has to be construed as unconstrained in the agent's model.

Concerning 4: Considering actions in a naturalised context requires a complementary partitioning of a considered space into agents and their immediate environment. Entropy reduction can be brought about only at the cost of the environment. This constitutes the fact that causation depends on **local** rather than global factors. Next to that, the **asymmetry** of an agent's action is reflected in the explanatory asymmetry of thermodynamic processes of different entropic order. Thirdly, it will have been learned that information needs to be represented physically at every stage of its processing, which can be proved by the counterexample of Maxwell's demon against the second law of thermodynamics, which can be conceived unless the assumption of physical representation is made. Since an agent is physically limited, this also means that an agent can only represent a limited number of causal laws. In order to reuse empirical knowledge, which knowledge of specific causal relations usually is, the agent needs to connect past with future instances of a causal connection through reference classes. For the conceptual (but not the ontological) problem, 'A causes B' therefore connects *types* A and B, not tokens. Given the right reference classes, a causal connection therefore gives rise to observable **regularities**. A second argument for the regularity of causal relations, understood at the conceptual level, will be the reflective or prospective, and therefore *abstract*, context, in which we evaluate causal relations.

Connotations of agency and other constraints on the concept

At this stage (section 6.3), after completing the detour via thermodynamics, we will have gathered further evidence for how to evaluate a causal claim based on agency. The first set of constraints we can now assume to hold corresponds to what is sometimes called the 'connotations' or 'platitudes' of causation: an A in 'A causes B' is a local factor (**locality** of causation), the causal relation is **asymmetrical**, and there is **regularity** in causal relations (causation connects types rather than tokens). That the account operates at the type level, rather than the token level, is to be taken in the conceptual sense that pervades this framework, not in the ontological sense. That means the deeper metaphysical question of whether causation really connects either local tokens or universals is not addressed. The regularity of causation requires the agent to recognize which causally relevant properties are 'instantiated' (again not in the ontological sense, but relative to its internal model of evaluation). Observable properties must allow an inference to the presence of causally relevant properties (for example: I observe the typical form, colour, size, etc. of what I take to be a stone, in order to infer that I could use that object to make a glass plate shatter by throwing it against it). The more informed concept of acting, and, by extension, causing, can also be used to evaluate judgments that construe actions as information-constrained. In addition, we will be able to make sense of the role of 'free action', which appears as an auxiliary concept in the agency-formula, and which had previously been understood only intuitively.

Causal transitivity, extensional identity, and objectification of agency

Up until that stage, we have a conceptual apparatus in place that allows analysing judgments concerning actions and binary causal relations at the type level. We can also make sense of judgments that concern informationally constrained actions, reconciling the two seemingly contradictory aspects of an action as both determined by information and, at the same time, free. With this, the level of expressiveness of causal judgments covered by the account is still quite confined. To extend it, we must first accommodate causal transitivity. Since causal judgments of the form 'A causes B' are, on the basis of this account, to be understood on type level and in a probabilistic sense, an agent can make the judgment 'A causes C', on the basis of his judgments 'A causes B' and 'B causes C', if the Bs are successfully identified. In a causal chain 'A -> B -> C', the B

has a dual role. It follows depending on A if we evaluate the ' $A \rightarrow B$ '-part, and it occurs spontaneously if we evaluate the ' $B \rightarrow C$ '-part. The question naturally arises how and why this seemingly contradictory way of evaluating B does not instil a contradiction into the real course of events. A similar problem can arise even with judgments concerning *our own* actions, previously considered free and now considered determined in the light of new evidence. Three different levels of abstraction, as introduced in chapter 3, can now replace the talk concerning different perspectives employed until then. It will be seen that the aforementioned, seeming contradictions are in fact no genuine contradictions. I will call the proposed account, a compromise between the objectivity and the agency-dependence of causation, the 'objectification of agency', which will rebut the anthropomorphism objection against the agency-approach (section 7.1).

Token level causation, the structure of causation, conclusion

With the objections of circularity and of anthropomorphism the main obstacles of a conceptual reconstruction are out of the way. Besides the defence against these major objections, a positive account has been delivered that corroborates the assumption that causal judgments might be agency-judgments. But some further problematic cases have to be accounted for. A problematic issue for an agency-account seems to be the token level case, ' a caused b '. In the general case, an agent's interventions can have varying success, which is possibly due an incompletely observed causal background, or possibly due to genuine indeterminacy inherent in the causal scenario. In such cases of varying success, every intervention is merely a chance-raiser of the effect. In section 7.1.1 I will look at two different ways of how probabilistic causation can be interpreted. One scenario (genuine indeterminacy) has all causal contributors in place and *then* determines the factor of remaining indeterminacy, which will be more or less biased by the setting of all the causal contributors; the other scenario (epistemic indeterminacy) takes the cause in question to make a difference to the effect deterministically while some of the other possible causal contributors are left unobserved. I will then demonstrate that the conceptual problem of causation can be addressed, i.e. sense can be made of judgments concerning such causal claims related to particular past events, without having to clarify which one of the two scenarios is real. I will also show agency's role in analysing causal claims that seem to preclude applicability of the conceptual apparatus of agency. The following sections (7.1.2 to 7.1.4) all address further specific problems of accounts that rely on chance-raising and difference-making, and thus challenge the agency-theory as well. Following the defence of the agency-theory against these problematic cases, I will look in section 7.2 at the impact of the concept of causation as derived from agency to questions concerning the structure of valid causal models. Rather than being a necessary condition, the modular models complying with the causal Markov condition will have to be regarded conceptually as a particular sub-class of a more general concept of causation. Therefore, a similar consideration like the one relating epistemic and metaphysical indeterminacy applies with respect to the structural question: modular, but also certain non-modular systems (level invariant systems) count as causal systems. I conclude with the assessment that the agency-account has been defended successfully against the major objections, if its interpretation is confined to the conceptual question of grounding causal judgments.

3 What is causation – the conceptual question

There is a certain sense in which one can understand the ‘conceptual problem of X ’. According to the concept-as-abilities-view (see Margolis (2014)), having the concept of X implies the ability to judge correctly that the concept of X applies in a specific situation. Accounting for *why* it applies, however, implies that the concept can be analysed in terms of *other* concepts.

To give an example of my reading of this problem, I can prompt an interlocutor to account for why she has denoted a geometrical figure, which is visible to both of us right now, by the word ‘square’. A possible answer might be: ‘Because my concept of a square implies having a rectangle with equal sides, and it seems that this figure in question satisfies these conditions.’

A concept can apply to an object in a narrow sense, like a tangible thing, or to something more abstract like a relation, as it is the case with ‘causation’. On this understanding of the conceptual problem, there is often no direct ontological commitment involved in analysing concepts, since the conditions that make the application of a concept viable could be devoid of such commitments. For example, this is often the case with *operational* definitions of concepts. The ontological commitment might, of course, be established indirectly, by virtue of the concepts that analyse the analysandum. Some philosophers writing on the subject of causation, who want to highlight this aspect of analysing causation, denote that problem by ‘semantic problem’ instead of ‘conceptual problem’.²

The concept-as-abilities-view, as far as the ontology of concepts is concerned, combined with the classical theory (again, see Margolis (2014)), as far as the structure of concepts is concerned, comes closest to the understanding of concepts needed to follow the argument of my thesis. The argument could also be thought of as consisting in the bet that the classical, or definitional, theory of concepts applies to the concept of causation, such that it can be analysed according to a set of constituents. Of course, the problems of such an approach are well known (see Sloman (2005)). I will address these worries later on in the thesis (e.g. in 6.3.1).

For now, another aspect is crucial, and this concerns the interpretation of the conceptual problem as distinct from ontological and epistemological questions, so here is another example illustrating this distinction: if my interlocutor has made a judgment involving the word ‘dog’, and if I prompt her to explain the use of that word, I can expect to learn something about my interlocutor’s concept of a dog, such as ‘social carnivorous animal’, ‘barks when in distress’, ‘lives in a kennel in the backyard’, etc. Concepts of an object can vary across different agents. The concept of a dog might be more or less fine-grained, with more or fewer conditions conjoined by conjunctive or disjunctive conditions. As will be made clear when analysing the concepts of ‘causation’ and ‘becoming informed’ in their respective sections, the conceptual problem is to be distinguished *both* from the epistemological *and* the ontological problem. Sticking with the dog example, at this problem-level I am not concerned with whether the agent’s dog-categorisation is actually correct, or how the agent might have verified that the conditions of dog-ness are satisfied by a candidate animal. This would concern the epistemological problem. Likewise, the conceptual question does not address what *really* makes an object a dog, whether dogs are instantiations of Platonic ideas, or examples of idealisations of survival strategies in certain ecological niches. What the question does address is the

² In the literature on causation, both expressions can be observed to be used for the same thing, e.g. James Woodward usually talks about the ‘semantic problem’ (e.g. on page 38 in Woodward (2003b)), whereas Dowe (2000) uses the term ‘conceptual’; both authors contrast this with epistemological and metaphysical problems of causation.

list of criteria that the agent thinks an object satisfies to comply with a corresponding concept. Making these criteria explicit can serve as an explanation of a corresponding judgment.

As far as causation as an object of inquiry is concerned, the answer to the question that prompts an epistemic agent to account for his judgment 'A causes B' might be: 'I think I could manipulate a B via an A, but not vice versa.' In fact, a very similar understanding of causation is defended by this thesis, but the analysis of causation by means of manipulation is understood at a conceptual level only. Again, in contrast to giving meaning, or content, to a causal judgment on the basis of other concepts, which is how I understand the conceptual analysis of causation, one can also ask how the agent has *verified* its judgment.³ One could ask what methodology was used to disclose the causal structure that the judgment refers to. I consider these problems as belonging to the epistemology of causation. One can also search for the truth-maker of the judgment. What establishes the seeming connection between A and B, and is this connection a kind of entity, an element of our ontology that best explains reality? These are classical, ontological questions of causation, which are again not addressed.

Hence I want to provide a conceptual analysis that is ontologically non-committed, and even leaves room for error in a causal judgment. This approach has the obvious disadvantage of yielding weaker claims than those resulting from a theory that makes definite assertions concerning the ontology of causation, or recommends a methodology for disclosing causal structure. But the conservative nature of the approach also has advantages, the biggest of which would be the clarification of what is actually meant by 'analysis of causation'. Defining a clear-cut conceptual level of the problem of analysing causation enables us first to set up a list of features and constraints that our concept has to satisfy in order to qualify as a possible object of inquiry. For example, this could be 'causation = correlation plus X', with X as some kind of symmetry-breaker. Then, the second step would be to find empirical evidence for the existence of such an object. The concept might turn out to be vacuous; it might also turn out to have objective counterparts in reality. The concept of causation can be thought of as comprising a family of related, smaller concepts, which satisfy different constraints but all equally qualify as subspecies of causation in general. Such a web of concepts would be similar to the inventory of an ontology in a Quinean sense, but allows elements that one might want to exclude from such an ontology, like relational objects. It also permits vacuous concepts that are clearly not part of a scientific ontology in a Quinean sense.

Analysing causation in terms of manipulability yields *prima facie* an *operational* characterisation of causation. This circumvents certain concerns that might be raised against the approach of explicitly distinguishing conceptual from metaphysical analysis that is more ontologically committed. For example, Quine (1960) asks whether such a clear-cut approach is possible if we investigate questions concerning definite scientific objects like neutrinos. If we were to find out that processes involving neutrinos actually feature particles with a mass, does that mean we have not been engaged with investigations into the properties of neutrinos, given that we had defined neutrinos, our object of investigation, as particles without mass beforehand? Or have we found out that neutrinos do, or at least can, have a mass? But causation understood conceptually in manipulative terms seem to be immune against these concerns, since empirical findings only concern how the means-end-relationship is established, not the conceptual nature of that connection. Similarly, a worry of a conflation of primary and secondary intension (Putnam (1973),

³ Henceforth I will refer to the agent (in the general, including non-human case) in the grammatically neutral gender form unless the context suggests human agency.

see also Chalmers (1996)) does not apply, and therefore the worry that causation is a rigid designator that might undermine what we have thought of as the essence of our object of investigation. The corresponding problem here is that the manipulability aspect of causation relates to the metaphysics of causation like the dispositional features of water relates to H₂O. Water's dispositions enable an investigator to identify her object of investigation via its primary intension. But once water's chemical structure, its secondary intension, has been determined, the meaning of 'water' is fixed.

There are two ways to respond to this worry. Either causation, if understood manipulatively, is seen merely as an instrumental concept that has only primary intension. Then, again, empirical information only adds further information without putting in doubt the starting point of the analysis. This is the template of the objection in Searle (1983) to the twin-earth thought experiment. But even if there is a secondary intension involved in that structural concept, there is reason for saying that whatever ontologically constrains agency, in the sense of a secondary intension, constrains causation, too, but *via* agency, causation's primary intension.

Two other worries that I would like to obviate right from the start concern the idea of 'unconstrained analysis' (even if the analysis is confined to the primary intension) and, closely related, the concern of subjectivity. First, I do not see that this kind of conceptual analysis is unconstrained in a pathological sense. Although it is true that an ontological underpinning of a causal judgment is a stronger constraint, the idea of a successful manipulation is constrained to some extent, and grounded in experience, although judging a manipulation that takes place merely hypothetically satisfies a corresponding constraint only in a qualified way. Similarly, it would be unfair to say that causal connections, construed merely as relations of means and end, and established by links whose nature we completely abstract from, depend only on an agent's belief. It is true that the judgment concerning a causal link depends on an agent's belief, but implicit in that belief is the idea that such a connection might at some point be verified or has in fact been verified by the agent uttering the judgment.

That the approach of distinguishing a conceptual and an ontological level of analysing causation (maybe in contrast to other objects of investigation that do not allow this clear-cut distinction) really is expedient can be shown by two conceptual distinctions within the family of causal concepts. One contrastive pair is the distinction between two types of probabilistic causation. On the one hand, there is in a causal connection a kind of genuine, observer-independent uncertainty about whether an effect will happen given that the cause, and all other factors that determine the causal background, have been fixed. On the other hand, there is an alternative, epistemic interpretation of the uncertainty about whether the effect will be triggered. The observer happens to be ignorant of the determining additional factors or unable to measure them correctly (section 7.1.1). Another contrastive pair concerns modular causation and causation that is merely level invariant (section 4.2). We see an ongoing debate concerning both of these issues (Hausman and Woodward (1999), Cartwright (2002), Hausman and Woodward (2004); Steel (2005), Drouet (2008)), but it is often unclear whether a claim like 'Causal relations satisfy the causal Markov condition' is meant normatively, viz. is meant to make recommendations about what makes a causal model a *good* causal model, or whether the claim is to be understood much stronger, viz. that in nature there is no causation that does not comply with the causal Markov condition. As far as both contrastive pairs are concerned, my approach makes the recommendation to first map the conceptual territory and assess which concepts of causation are available. These should be checked for inner consistency and be assigned adequate names. *Then* it depends on empirical findings

whether these concepts are vacuous or have natural representatives. Whereas this procedure seems to work well for the debate on how to interpret probabilistic causation, either epistemically or metaphysically (after all, none of the two concepts is outrageously inconsistent), the debate on modularity and the causal Markov condition is for some curious reason not so clear-cut.

3.1 Levels of abstraction

A useful method from the philosophy of information (Floridi (2010)) is the method of levels of abstraction, which posits that it is desirable to stick consistently with a conceptual scheme according to which a problem is addressed. In particular, the answer to a question should be given in a way that corresponds to the question asked. This, of course, does not preclude the approach from tackling a problem in an iterative way that changes the focus of an investigation, and asking new kinds of questions in the light of new evidence. But such a cycle of rephrasing questions should be made explicit. Levels of abstractions are used in this thesis to express in a semi-formal fashion what can more colloquially be called ‘perspectives’ or ‘contexts’.

To first provide an example of where some debates of causation sometimes take a wrong direction, one can point to the fact that it is often difficult to tell confidently at which level of ontological commitment a theory operates. When Nancy Cartwright says that nature can assign probability distributions for observables of a causal model in an unconstrained fashion (in Cartwright (1993); see also section 4.2.4.1 for further details on this issue), she is obviously starting from a different set of assumptions than the defenders of the causal Markov condition. For example, she might be implying that, from general assumptions about causation, there is no constraint intrinsic to our notion of causation that necessitates the causal Markov condition to hold. In other words, nature might assign probabilities in a certain way, but we cannot expect that to happen *a priori*. Defenders of the causal Markov condition likewise often fail to argue in a way that makes it transparent whether they are assuming that all systems naturally behave in accordance with that condition and equate these systems with ‘causal’ systems, or whether the causal Markov condition is inherently given by the concept of causation.

As a second example, a similar instance of talking past each other can be observed in Woodward (2003a) and Pearl (2003), which shows that the two authors seem to have a very different understandings of what it means to ‘define causation’, without, however, making it clear what a satisfying definition consists in.⁴

Levels of abstraction are an attempt to circumvent these unnecessary problems, and a relatively convenient one. To provide an overview of the gist of the method of abstraction, I will briefly summarize the account given in Floridi (2008). Floridi devises ‘six key concepts necessary to explain the method of abstraction, namely, typed variable, observable, level of abstraction, behaviour, moderated level of abstraction, and gradient of abstraction.’

- A *typed variable* is a conceptual entity, which can be seen as the atomic building block of description of some system under investigation. The property of being typed is meant to specify exactly which kinds of value the variable can assume. For example, this prevents delivering a qualitative description of some feature of an object, when a quantitative description is required.
- Being an *observable* adds the feature of being not only a typed variable, but an interpreted one. This is supposed to prevent confusions like those arising from different units of measurement, like imperial vs. metric, which can entail that sameness in values of variables results in different specifications of the described system.
- A *level of abstraction* is a set of observables eligible to model a system under investigation.

⁴ This impression is further corroborated by Woodward’s reference to Pearl’s online discussion with readers, especially with regards to the question ‘Has causality been defined?’, which Pearl answers in the affirmative. The page can now be found at <http://bayes.cs.ucla.edu/BOOK-2K/singpurwalla.html>

The definition of observables, or setting up the level of abstraction (henceforth also abbreviated by 'LoA'), is only the first step in studying a system. The second step consists in deciding what relationships hold between the observables.

- The *behaviour* of a system, at a given LoA, is defined to consist of a predicate whose free variables are observables at that LoA. The substitutions of values for observables that make the predicate true are called the *system behaviours*. Behaviour can also regiment relations between observables of a LoA. In some cases, the relations between observables can be functions, such that, when one observable assumes a certain value, another observable's value is uniquely determined by that assumption.
- A *moderated LoA* is defined to consist of a LoA together with a behaviour at that LoA.
- Finally, a gradient of abstraction regiments relations between the observables of different LoAs. For example, two LoAs can share some observables between each other, while describing other features of a system by exclusively maintained observables. The two special cases are 'disjoint' and 'nested' gradients of abstraction. Disjoint gradient of abstractions consist of LoAs of disjoint sets of observables, while nested gradients of abstractions afford a hierarchy of LoAs where each value in a more abstract LoA is preserved in a more fine-grained description at the LoA following in the hierarchy.

A philosophical example Floridi provides as a possible application of the method of abstraction is a gradient of abstraction consisting of the Cartesian *res extensa* and *res cogitans*. This is an example of a disjoint LoA. Another example is the set of Kantian antinomies – seeming contradictory statements, for each of which a proof can be delivered. Again, the antinomies can be shown to arise from disjoint LoAs, within which the proof proceeds correctly according to the rules of logic (the behaviour of the LoA). The relativity of the result – relative to the choice of LoA – correctly represents Kant's idea that our scheme of reason is inadequate to tackle absolute metaphysical questions.

As far as the two aforementioned examples from the philosophical debate on causation are concerned, it seems it would help to specify whether the causal Markov condition is supposed to govern models or real systems. If the first were the case, then the causal Markov condition is part of the behaviour of the level of abstraction chosen to model a causal system, and therefore a prior constraint. In the second of the above examples, one should specify what a good definition consists in, viz. how the variables of the conceptual explanation LoA are supposed to relate to each other. E.g., eliminative, contextual, indirect definitions would all allow for different relations of terms.

What I will borrow from the method of abstraction are two ideas: first, we have to make clear what kind of answer we have to expect from the kind of conceptual analysis that I am proposing. For example, in the previous section I have made clear in which way I abstract from ontological assumptions, and therefore from providing truth-makers for causal claims. Secondly, it can be useful to choose different kinds of levels of abstraction for different purposes. When this is done, the relation between the different levels of abstraction should be made clear, in order to prevent confusion.

I will show that the agency-theory of causation can be acquitted of the objection of conceptual circularity and conceptual regress, by showing that raising the objection involves what corresponds to an inappropriate use of the method of levels of abstraction. I'll show a similar misuse

in the context of information, since the two cases, causation by action, and becoming informed via information transfer, are analogous (see sections 5.1 and 5.2).

Given below is a preliminary list of problems which arise if incommensurable levels of abstraction are applied to the same situation and the results are compared without referring to the correct level of abstraction that yielded the result:

- An action (or more generally, an event) seems to be spontaneous *versus* seems to be caused mechanistically
- An action seems to immediately influence its target *versus* appears to affect its target via causal intermediaries
- An event seems to be triggered by information about a past matter of fact *versus* seems to be mechanistically triggered by the physical properties of a signal, bypassing any possible semanticisation of the information
- An intermediate event B in a causal chain A causes B causes C has contradictory properties of being dependent on a causal predecessor *versus* occurring spontaneously
- There is an apparent over-determination in mental causation, when we posit both mental and physical causal antecedents

3.1.1 Levels of abstraction for causal judgments

I claim that correct application of the method of levels of abstraction can resolve the seeming contradictions mentioned in the previous section. To indicate the solution, I will briefly anticipate my results. I am going to use three different levels of abstraction, which correspond to different perspectives of looking at causal phenomena:

LoA1 is defined as the level of abstraction that is used for a judgment that concerns one's own action. It posits an absolutely direct and either uncaused (section 5.1), or *informationally* caused (section 6.1) action, and an indirect effect of that action. Information addressee, acting agent, and epistemic agent (the one forming the causal judgment) are *identical*.

This first-person LoA allows for an occurrence of an action either in a completely unconstrained way or in such a way that the action occurs as constrained by information whose relevance is relative to the agent's goal. The action cannot be regarded as coerced by an efficient cause, in which case the event fails to qualify as an action. Also part of this LoA, since we are considering causal relations, is a variable *E*, which references an event external to the action itself. *E* is the event indirectly brought about by the action, which is tantamount to being *caused*.

LoA2 is defined as the level of abstraction for describing causation by information (see section 6.1) when judging another agent's doings. LoA2 employs an objective perspective; now the epistemic agent is *not identical* with the acting agent as at LoA1. Also as opposed to LoA1, which features unanalysed semanticisation of information, LoA2 needs a structural account for why a piece of information itself can sometimes be considered as the cause of an event, rather than the physical signal merely *carrying* that information. It does require some account of identity, e.g. identity as persistence, too. But at this level of abstraction, the identity needs to span the roles of information addressee, acting agent, and beneficiary of feedback (which can be physical or referential) originating from the action. This is the case if the agent embodies a cybernetic feedback cycle. The

required notion of identity across time makes Markovian models insufficiently expressive⁵ and therefore incompatible with this LoA.

LoA3 is an objective view that does not feature and does not *need* a concept of identity of any sorts. Accordingly, this is the LoA that is concerned with depicting efficient, or Markovian⁶, causation (Markovian LoA, 'LoA3'). What corresponds to immediate action at LoA1 is at LoA3 an (uncaused) assumption of a value of the cause-variable, which is tantamount to 'wiping out equations' (see section 4.2) for the special case of modular systems. Variables of LoA3 are instantiations of causally relevant properties (see chapter 6.3).

LoA3 covers, in its simplest form, efficient causation in the sense of a simple dependence of an effect on a hypothetical action by the agent who is also the one judging the action. My interpretation of the agency-account of causation is tantamount to saying that this LoA evaluates *binary* causal propositions in the same way as LoA1 does, except for dropping the identity condition of LoA1. Application of LoA1 gives meaning to a judgment that concerns efficient causation involving two variables. Much of 6.3 and subsequent sections deal with extending this LoA, such that more complex causal models can be constructed. By adding further constraints to causation, different sub-concepts of causation can be constructed, which all retain the basic notion provided by the first-person LoA.

Between these LoAs there are relations (explained in more detail in chapter 7). LoA1 can always be applied to evaluating a binary LoA3 claim. LoA3 can always be applied to LoA2, since no matter whether information is disseminated or an efficient (non-informational) cause is arbitrarily set, the repetitive measurement of an effect will give rise to a regularity (since regularity will be shown in section 6.3 to be a necessary characteristic of causation). However, doing so changes the explanatory focus, and the kind of causal explanation that LoA3 delivers might be in conflict with the expectation of an interlocutor that asks for an explanation according to LoA2.

3.1.2 The limits of the method of levels of abstraction

The method is useful for designing concepts, when concepts are seen as intimately connected with models of evaluating judgmental claims, but there is no deeper ontological import in the way the method of abstraction is applied in the course of my analysis. I give two examples that show the limits of the approach.

First, as it will be seen in sections 6.1.1 and 6.2.6, an interesting question is whether an intrinsic property of a physical signal or its relational property of carrying information is the explanatorily relevant property for explaining observed behaviour. My suggestion will be to look at different ways of representing the information internally, such that for a specific representation a suitable computation needs to be implemented by an agent in order to extract the matter of fact that determines the adequacy of the agent's subsequent behaviour. If the signal's representation is arbitrary, it cannot be causally relevant; instead it will merely be a reference to the causally relevant property, which is realised elsewhere. But the arbitrariness of the representation can only be shown by comparing an ensemble of agents, or, alternatively, a class of similar situations (see 6.2.2), therefore it depends on an extrinsic property of the signal whether the agent processes the

⁵ Since Markovian models like Bayesian networks forbid cyclic structures in graphs, identity has to be asserted extra-logically by a corresponding interpretation of the variables involved. There is no symbol for and no relevance of identity across time in Markov chains.

⁶ Markovian means *memoryless* causation in this context, in contrast to the causation via referring back to a past event, which LoA2 depicts. Thus, I am not implying with that term that the full implications of the causal Markov condition are satisfied, if LoA3 is applied.

information semantically or merely seems to do so. The latter will be the case if the agent's internal mechanism is configured in such a way that the right response to the causally relevant matter of fact is brought about, but no semantic processing of information ever happens – in other words, if the agent could not represent the relevant matter of fact by *other* means than those given by the rigid way the internal mechanism is wired. If we build an artificial intelligence and the ethical question arises whether the intelligence works rather like a mindless toaster or rather like us, the scheme resulting from applying the method of abstraction would not settle the matter since it seems that we would need an *intrinsic*, not an extrinsic criterion of being semantically enabled. In other words, the policy-maker raising the question would hardly be satisfied by the claim that an answer to his question depends on how to look at the problem, which is, according to the method of abstraction, *relative* to a LoA. At least such an answer will be as unsatisfying as a pan-psychic account of toasters or a materialistic account of how human beings process information. Nevertheless, in the section concerning concept acquisition I will rely on the scheme given by LoA2, but in the sense of a necessary, not a sufficient criterion for the successful acquisition of the concept of agency.

The second problematic use of the method of abstraction concerns the reduction of entropy, crucial to my suggested model of a paradigmatic action. Reducing uncertainty concerning a contingent matter of fact can be modelled by reduction of *informational entropy*, and informational entropy in turn allows for using infinitely many different levels of abstraction (e.g. measuring the height of a person with arbitrary precision). But this knowledge can refer to physical matters of fact that are independent of how we model the concerned system. In particular, the physical entropy of a system relative to its surroundings determines whether certain thermodynamic processes are possible. For example, whether the wheels of a perpetual motion machine will turn *by themselves* does not depend on our way of modelling the machine, but on its absolute state. This holds true as well for the general background assumption about our physical world, i.e. the continuous rise of entropy.

The two problems are similar in that in these cases we do not seem to be free to choose a level of abstraction in order to get to the core of the problem; rather the problems seem to be 'absolute'. Regarding the point of 'absolute questions', I therefore do not concur with Floridi's point that these kinds of questions cannot be meaningfully asked.

As an example that shows the limit of levels of abstraction (see Floridi (2008)), consider the question whether a software agent can be deemed autonomous, interactive, and adaptive, if its observable behaviour has these features. This example is discussed in Floridi (2010), section 4.1. According to Floridi's account, once the code of the software agent were revealed, it would thereby be shown that it has been 'simply following rules' and the ascribed features would have to be jettisoned. This might be enough to choose an appropriate 'stance', in a Dennettian sense, in order to describe the system, but does not inform us about how the system experiences *itself*. A solution that purely relies on levels of abstraction turns on the unwarranted assumption that an agent is free to impose a level of abstraction concerning its own state. I would even claim that these problems generally show the limits of purely *informational* approaches to science or metaphysics – at least if information is understood, as in this thesis, as a relational (or referential) notion.

4 Theories of causation

There are currently many different causal theories that are still being developed further. I'll present two different ways of mapping the theories, in order to get an overview of them. The first approach will be called 'vectorial' and describes a more abstract view, which maps causal theories by means of their properties, according to which they can be distinguished from each other. The approach that is more orientated to the classification found in the literature on causality will be referred to as the 'denotational' one.

Vectorial description

The dimensions of description belonging to the vectorial view can be given as follows:

1. theories of event-based causation vs. theories of property-based causation (very similar to, and sometimes also referred to as token- vs. type level causation, or singularist vs. regularity theories of causation),
2. theories of productivity vs. theories of relevance (the same dichotomy is sometimes represented as theories of dependence vs. theories of difference-making),
3. theories of causal mechanisms vs. theories of probabilistic causation (quite similar to productive causation vs. relevance-causation, but rather pertaining to different methodologies of finding causal structures),
4. epistemic theories of causality vs. ontologically committed theories of causality.⁷

Regarding 1. 'Event-based', 'token level' or 'singularist' theories of causality posit that two *relata* *A* and *B* are causally connected via an intrinsic relation, which is called 'causal'. When attributing this property to the considered relation there is no reference to other instances of the same relation, or to instances of similar relations like the one between *A* and *B*. 'Property-based', 'type level' or 'regularity' theories of causality claim that the causal aspects of a relation between two *relata* are deducible only from other instances of the regularity, or from the fact that the *relata* are instantiations of abstract entities. In other words, according to those latter theories, a causal relation is an extrinsic feature of the connection between *A* and *B*.

Regarding 2. Theories of causal productivity claim that causes *engender* their effects, and are therefore 'productive'. Examples of causal over-determination and examples of causal pre-emption do not contradict those theories, since causes are considered sufficient for the effect, rather than necessary. On the contrary, theories of relevance (also known as theories of dependence or difference-making) stress that a cause must be counterfactually relevant to, or necessary for, the effect.

Regarding 3. Not entirely independent of 1 and 2, there are two different methodological approaches to investigate causation. One starts from looking for mechanisms connecting cause and effect, while the other is based on the concept of probability. Mechanistic approaches consider probabilities merely as an indicator of a causal connection and, in most of these approaches,

⁷ More independent characteristics can be found, e.g.: reductionist vs. non-reductionist theories, unified theories vs. theories of causal plurality, Humean regularity theories and non-Humean property or law-based theories of causality

probability plays only a peripheral or no role at all. Conversely, probabilistic theories often abstract from mechanisms.

Regarding 4. The difference between epistemic theories of causality and ontologically committed theories can be paraphrased by the question whether causality should be considered to be ‘in the system’ or ‘in the model’. Since a model can always be attributed to an epistemic agent, epistemic theories aim to solve the problem of how an agent acquires knowledge of a causal connection between two *relata*, whilst ontological theories aim to clarify what criterion qualifies a connection as causal, independent of any observer.

Denotational description

The four pairs of categories outlined above form dimensions of characterisation, which, when combined according to the corresponding values, can help to locate a specific causal theory. The alternative, denotational way of referring to a specific theory uses the names usually found in the literature. This approach yields the following list of causal theories as prominently discussed, but the list is far from being exhaustive:

- I) theories of probabilistic causation,
- II) process theories,
- III) invariance theories
- IV) manipulation theories,
- V) mechanistic theories,
- VI) the theory of epistemic causation, and
- VII) counterfactual theories of causation.

A brief description of all the mentioned theories which features their more distinguished elements follows.

I - Probabilistic causality: the probabilistic theories’ starting point is the observation that a cause *C* raises the probability of the effect *E* taking place: $\Pr(E|C) > \Pr(E|\text{not } C)$. In the strong variant of the probabilistic approach, this formula is the basis for reducing causality to probability. In weaker versions, the stipulation of reducibility is given up. However, a strong connection between probability and causality is maintained.

II - Process theories describe causality by means of their time-dependent feature, i.e. as a process. A causal interaction involves the exchange of an entity, which can be a ‘causal mark’ (in Salmon (1984)), or a ‘conserved quantity’ (in Dowe (1992)).

III - Invariance theories employ the notion of ‘invariant generalisations’ that can be observed in causal systems. Depending on the specific theory, those ‘invariances’ can refer to the stability of a functional relationship between parts of a system against disturbances, or against interventions on exogenous elements of a model. Similarly, relative to the further characteristics of a concrete causal theory, the invariances can be interpreted nomologically, or mechanistically.⁸

⁸ Some authors, like Cartwright (2004), assert that ‘invariance theories’ are a category of their own. The description I delivered is meant to indicate that ‘invariance’ can also be seen as an aspect of theories that

IV - Manipulation theories are based on the claim that ‘causes can be used to bring about their effects’. In the strong version, manipulation is confined to human agency. An example of the strong version can be found in Price and Menzies (1993). The weaker version considers instead the notion of interventions as direct influences on some system parts. Woodward (2003b) further weakens this requirement by including *hypothetical* interventions, which relates this particular manipulation theory closely to the counterfactual theories.

V – Mechanistic theories focus on mechanisms, which play the role of a necessary substrate of any cause’s influence on its effects.

VI - The theory of epistemic causation developed by Williamson (2005) is a Bayesian theory.⁹ It is related to probabilistic theories of causality, but is distinguished from them by strictly adhering to the ‘degree-of-belief’ interpretation of probability. Also, it incorporates rationality principles, like maximum equivocation and belief update, which, combined with background knowledge, describe the evolution of causal knowledge of an epistemic agent.

VII - Counterfactual theories are committed to the identity ‘A causes B = If A had not happened, B would not have happened’. In order to compare the actual world with the corresponding counterfactual one, a metric is needed, which, in Lewis’ case, is the ranking of possible-worlds, elaborated in Lewis (1979).

In order to provide an example of how to square these theories with the aforementioned four dimensions of characterisation, consider that most of the variants of probabilistic theories of causality in (I) are based on a concept of type level causation, and they describe causes as relevant rather than productive. Whether they are epistemically or metaphysically committed depends on the specific variant of probabilistic theory, in particular on the interpretation of probability. As another example, process theories of causality in (II) are metaphysically committed: they describe productive causes and events in space-time, and they operate on token level.

Prominent work and authors

Following the aforementioned classifications, I will present an overview over the relevant work on causation done by some prominent authors.

The origin of the regularity view on causation is often traced back to the work of David Hume (1748). According to Psillos (2002), Hume’s analysis of the problem of causality resulted in two causal concepts, one pertaining to causality in the mind and one pertaining to causality in nature. Modality enters in the former, but not in the latter concept, since causality in nature is nothing but regularities. Two important subsequent adherents of the regularities view on causation are John Stuart Mill and John Mackie. Mill (1843) presented a more elaborated version of a regularity theory by distinguishing a set of causal factors (rather than one causal object, like Hume) as positive

could be attributed to another category. The other denoted categories seem more like theories of their own distinct kind, not overlapping as much with the others.

⁹ Williamson (2005) refers to epistemic causality as a form of ‘objective Bayesianism’, which, according to Williamson, has its origin in the writings of Bernoulli, Laplace and John Maynard Keynes.

conditions, and the absence of negative conditions as necessary (and sometimes sufficient) to produce the effect. He therefore combines the productive and difference-making view on causation. To the inferential methodologies, he contributed the method of agreement: 'If *ABC* bring about *abe*, and *CFG* bring about *efg*, then *c* is the cause of *e*', and the method of difference: 'If *ABC* bring about *abe*, and *AB* only bring about *ab*, then *C* is the cause of *e*' (ibid.). Mackie (1974) further refined Mill's ideas by treating causes as 'INUS conditions' – insufficient but non-redundant parts of an unnecessary but sufficient condition. Psillos offers another modern variant of the regularity theory in Psillos (2002).

The aforementioned theories are all regularity theories and are grounded in the idea of 'Humean supervenience', according to which causation is construed as a relation that supervenes on observable, spatio-temporal relations. By contrast, there are non-singular approaches to causality not based on Humean supervenience, represented by David Armstrong and Fred Dretske, among others. In Dretske (1977) a view is elaborated that takes singular causal events as instances of relations between universal properties. Armstrong (1983) is a strong realist concerning universal properties by positing existence not only of first-order universal properties, but of hierarchies of properties. Observable regularities are explained by relations between the universals themselves, which gives his theory a modal quality that the theories grounded in supervenience lack. According to Gillies (2002), regularities, mathematically formulated as probabilities, are indicators of causation, but causation must itself stand as an independent and irreducible concept.

The first, influential singularist theory of causation was developed by Ducasse (1968). Paraphrasing his account, a cause of particular change *E* is another particular change *C* that alone occurred in the spatio-temporal environment of *E*. Other, modern theories of singular causation include the process theories of both Wesley Salmon and Phil Dowe. In their theories, the truth maker of a particular causal claim is derived from observation of the intersection of two world-lines of objects that either exchange a causal mark or a conserved quantity, respectively, whilst there is no reference to a regular succession, or to a law of nature.

The aforementioned process theories are also theories of *causal productivity*. A similar, but methodologically different account is Stuart Glennan's theory of causal mechanisms. This theory of singular causation and causal productivity is defended in Glennan (2009). Likewise, Machamer, Darden and Craver's mechanistic theory (Machamer, Darden et al. (2000)) holds that causes are productive, but unlike Glennan's, their concept of mechanisms cannot be said to be based on a singularist view.¹⁰

While both the process and mechanistic theories fit more naturally with the productive view of causation, the probabilistic theories are all theories of causal relevance (or difference-making). However, the two characteristics of causal theories, of being based on probability or on causal relevance, are not identical as can be seen with counterfactual theories, which are theories of difference-making but appeal to counterfactuals rather than probabilities. Moreover, the counterfactual stance relates to the singular case rather than regularities or universals. A similar precaution stops one from relating all probabilistic theories to theories of regularity. When interpreted as chances, probabilities raising an effect can refer to the propensity of a single event.

Representatives of probabilistic theories of causality are Patrick Suppes (cf. his seminal work, Suppes (1970), in which the basic idea of causes as probability-raisers are introduced), Christopher

¹⁰ Russo and Williamson call such theories "top-down" as opposed to "bottom-up", like Glennan's. The former corresponds to a generalist, the latter to singularist approach.

Hitchcock, the early Nancy Cartwright, and authors who work in the field of causal Bayesian networks, which should more adequately be called causal Markov networks. Peter Spirtes, Clark Glymour and Richard Scheines (Spirtes, Glymour et al. (1993)) developed causal Bayesian networks in parallel to Judea Pearl. Spirtes et al. developed the first algorithm for inferred causation ('PC-algorithm'). They also formulated a set of assumptions (minimality, faithfulness and the causal Markov condition; see section 4.2.1 for details) that must be satisfied for correct application of Bayes nets methods. Since their work is focused on quantitative questions of causation, they do not extensively discuss the problems revolving around different interpretations of probability. However, since they are, like Judea Pearl, mainly interested in a calculus for causal inference, their concept of probability is best understood as related to frequencies of actual observations.

Judea Pearl is one of the developers of Bayesian belief networks (see Pearl (1988)), which he subsequently applied to causal contexts. He also devised a mathematical theory of interventions, in particular the so-called 'do-calculus'. The latter is an extension of conventional statistics, which is unable to express the difference between evidential and interventional information.¹¹ Another paramount achievement of his is the treatment of counterfactuals in the context of Bayesian networks. The later Pearl changed his stance with regards to the interpretation of probability in a causal model, and interprets connections between nodes of a causal graph physically, despite his scientific background, which lies in the degree-of-belief favouring Bayesianism (cf. Pearl (2001) for a detailed account of this change of stance). Most of the theorems of his major work, Pearl (2000), require causal background knowledge or causal beliefs in addition to measured data, and therefore he does not address the fundamental epistemic causal problem of bootstrapping the build-up of causal knowledge. However, his work is interesting with respect to the development of a unified causal theory, since it straddles both the mechanistic-vs.-probabilistic categorisation, as well as the dependence-vs.-productivity categorisation. That being said, a constraint in his approach is that he considers only modular systems, which comply with the causal Markov condition. James Woodward considers his work (see Woodward (2003b)) to be complementary to Pearl's. It addresses the philosophical implications in more depth, and focuses more on causal explanation as opposed to Pearl's focus on causal inference. Hausman and Woodward (1999) developed an invariance theory of causality that is applicable to modular systems. They analyse the common implications of structural equation systems, which are causally interpretable, as well as the causal Markov condition. This is the pivotal condition of Bayesian networks and other theories derived from the probabilistic approach. It should be noted, however, that both Woodward and Pearl do not further justify their assumption that structural equation systems and probabilistic models based on correlational data are equivalent ways to describe causal systems. The accounts of the two authors differ, however, in their interpretation of what a mechanism is (see section 4.2.5). Both Pearl's and Woodward's theories are probabilistic and mechanistic theories. Similar to Salmon's process theory, which is sometimes referred to as 'mechanistic' (e.g. in Psillos (2002)), one should notice that this attribution could cause confusion with those causal theories for which mechanisms are the central object of enquiry, like Glennan's and Craver's.

Whereas Pearl departed from the original programme of probabilistic causation in favour of a more ontologically committed approach, Nancy Cartwright developed a sceptic stance against one of the structural constraints connected to the programme. This constraint is the 'screening-off'

¹¹ This distinction refers to observational data and data acquired in experimental settings respectively, e.g. in the context of randomized clinical trials.

principle, which is one of the corollaries of the Markov condition. This scepticism is explained in Cartwright (1999a) and Cartwright (2001), then in more technical detail, and in response to Hausman and Woodward (1999), in Cartwright (2002). Her own account is one of causal plurality, developed in Cartwright (1999b). In Cartwright (1999b) and Cartwright (1989) she also introduces the concepts of ‘capacities’ and ‘nomological machines’, which supply the situational contexts for capacities to be productive. Her approach is explicitly singularist.

All of the aforementioned theories are ontologically committed to some degree. Pearl, Gillies, Armstrong, Glennan and Kevin Hoover (who outlines his causal theory in Hoover (2001), and subsequently applies it to econometric questions) are particularly committed to some form of realism in their theories. The contrasting approach can be described as ‘epistemic’, or ‘model-relative’. Its recent advocates are Jon Williamson and Steven Sloman (cf. Sloman (2005)). Williamson started with a genuine Bayesian and empiricist approach, but recently turned his attention to incorporating mechanisms into his causal theory, making this theory another variant of a hybrid between a mechanistic and probabilistic theory. However, it remains an epistemic theory. According to the Russo-Williamson thesis in Russo and Williamson (2007), the role of mechanisms is that of a second pillar, next to probabilistic difference-making, on which to ground causal inferences.

Problems of current causal theories

One way to show the shortcomings of causal theories is by means of devising *counterexamples*, a strategy often used in the literature. Every causal theory formalises the concept of causation to some degree of precision. If a considered situation is judged to be causal according to the formal concept of a theory, whilst it is judged to be non-causal according to the general understanding of the term, the theory produces a type I error, or false positive, with respect to a causal claim. If the converse is true, the theory produces a type II error, or false negative. In both cases, the range of the formal causal concept differs from the range of the un-formalised concept. The following list arranges some counterexamples that can be found in the literature as type I error and type II error counterexamples against theories with corresponding properties.

1.1 Regularity theories

Type I error counterexamples: the stock counterexamples against regularity theories involve non-causal associations, also known as ‘spurious correlations’, or correlations in a time series, see for example Eliot Sober’s much discussed example of covarying Venetian sea levels and British bread prices (Sober (2001)).

Type II error counterexamples: intuition asserts certain causal relations that have no observable instance, see Christopher Hitchcock’s example that ‘eating one kilogram of uranium 235 causes death’ (Hitchcock (1995)).

1.2 Singularist theories

If a single causal event is considered an instance neither of a regular association, nor of a type level property, nor of causal law, *type I and type II error counterexamples* can be constructed showing that the causal concept becomes *vacuous* in terms of difference between a causal and a coincidental association.

2.1 Theories of causal relevance

Type II error counterexamples: unsophisticated theories of causal relevance do not capture cases of causal over-determination and causal preemption, since, in these cases, the cause does not make a difference to the effect. Here is an example from Hall (2004): ‘Two children, Billy and Suzy, are throwing rocks at a bottle. Suzy’s rock hits the bottle first, just before

Billy's. Suzy's rock causes the bottle to break, even though Billy's would have done so if she had missed.' I discuss causal over-determination in section 7.1.3

2.2 Theories of causal productivity

Type II error counterexamples: cases of causation by omission, e.g. the statement 'my failing to brake caused the accident' from Glennan (2009). The reasoning behind examples of this kind is that negative causal factors cannot be considered 'productive' to bringing about the effect.

3.1 Probabilistic theories of causation

If a difference between informational and causal relevance of a cause to an effect is denied (as argued by Hausman and Woodward (1999)), examples of non-deterministic causation with effects of a common cause are *type I error counterexamples*¹² against probabilistic theories, e.g. Cartwright's factory example, described in Cartwright (1993). I discuss this example in section 4.2.4.

Type II error counterexamples involve cases in which a productive event is an instance of a factor that generally lowers the probability of the effect. One such example is due to Deborah Rosen (cited from Hitchcock (2002)): 'A golfer badly slices a golf ball, which heads toward the rough, but then bounces off a tree and into the cup for a hole in one.' This example will be dealt with in detail in section 7.1.3.

3.2 Mechanistic theories

The process-based variants of mechanistic theories face the *type I error counterexamples* of causal irrelevance, of which one instance is given in Hitchcock (1995): a pool player chalks a cue stick, strikes the cue ball with the stick, the cue ball strikes the eight ball, which drops into the pocket. In the course of this process the cue stick transfers a chalk mark from the stick to the ball and from the first ball to the second. The chalk mark counts as a 'causal mark' according to Salmon's process theory and as a 'conserved quantity' according to Dowe's theory; however, the mark is causally irrelevant to the effect.

Type II error counterexamples against mechanistic theories that define mechanisms as complex aggregations of causal intermediaries between cause and effect are examples that imply causal relations that depend on fundamental laws, or examples of higher-level causation for which a mechanism is not clearly identifiable, or unknown. Such cases are found in epidemiology or econometrics.

The counterexamples in the list indicate that the concepts of causation proposed by a causal theory fare well in certain classes of systems, whilst they fail to imply correct interpretations of causal events in other systems. The above list has been given mostly for doxographical reasons. Of course, proponents of theories that are problematic in some respect have reacted to the problems, but it is not my intention to cover these kinds of debates in further detail. What I find interesting to notice, however, is that, in the above list of authors, no one seems to be engaged in 'conceptual analysis' in a proper sense.¹³ They also do not impute this endeavour to their perceived opponents, which, one would expect, implies that specific counterexamples are not to be considered as critical blows to a

¹² The false positive account is due to attributing a causal relationship between the two effects, which are actually causally independent. The false attribution results from the conditional and unconditional dependence of the two effects, from which a causal relation follows according to the logic of probabilistic causation.

¹³ An exception might be Hume himself, if we interpret his project accordingly.

theory. But often this conclusion is drawn from counterexamples. For example, the agency-theory of causation is often readily dismissed just by pointing to its alleged inability to deal with examples of causation that do not involve human agency.

Next to devising counterexamples, one can raise more generic caveats against every theory of causality. Based on the denotational mapping, I will present some of these caveats in the following paragraph.

Probabilistic causality was made problematic by numerous counterexamples. This forced continual adjustments of the original simple approach of explaining causes as probability, which lead to a considerable complication (see Hitchcock (2002) for a survey). Many original proponents of this approach have abandoned it in favour of other approaches. It also inherits the ongoing debate on the nature of probabilities, the concept to which proponents of this theory hope to reduce causation. *Process theories* have a very confined scope of application (mostly physical processes) due to their low level of description, which involves the identification of causal marks or conserved quantities. *Invariance theories* struggle with disentangling the two tasks of conceptual clarification of causation and the methods of identifying causal structures. In the case of *manipulation theories*, the variant of agent manipulation theories needs an account of how to accommodate cases of causation where human agents are not involved, whereas the broader theory that depends on 'intervention' seems question-begging with regards to the question of how to define intervention in non-causal terms. The *theory of epistemic causation*, like many other theories, lacks a discussion about the nature of the causal *relata*, since merely treating these as variables in an agent's model seems to be an unsatisfying interpretation. Also, some critics of epistemic causality hold against it its non-commitment to metaphysical foundations of causation.¹⁴

¹⁴ As opposed to other theories of causation, it is difficult to find our construct *counterexamples* against one or the other type of theory from the dichotomy of epistemic and ontologically committed theories, which is why they are missing in the above list of type I and type II errors. For example, a prima facie suggestive move would be to try to construct a counterexample against an ontologically committed theory if a causal explanation in this theory turns on causal properties that never manifest themselves, such that the license to employ the property in the explanation depends on a belief in the existence of the property. But since the empiricist dogma, which requires observability of an existing property, is rarely maintained by most contemporary philosophers, this criterion will not do.

4.1 The choice of *difference-making* for addressing the conceptual problem

To address the conceptual problem of causation, one of the desiderata would be that the theory should provide a simple criterion for what determines a causal judgment, whose simplest form is the assertion whether a causal connection between two observables exists or not. Difference-making theories offer such a prospect. Consider the concepts of *correlation*, or *covariance*. There is a clear recipe for finding out whether two observables are correlated, consisting of a stage covering their measurement, and a mathematical procedure of calculating the statistical relationship. Also, there already is a 'difference-making' in an evidential sense involved, such that when we see the one observable, the likelihood of correctly predicting the value of the correlated variable is increased. But correlation is not causation. Nevertheless, if we could find an additional factor to be added to a confirmed correlation of two observables, the prospect of simplicity of the approach to the conceptual problem would probably be satisfied. More complicated specifications of how to evaluate a causal judgment would then, hopefully, correspond to a more specified concept of causation, derived from the simple root-concept given by difference-making.

There are several theories of difference-making on offer, as was seen in the previous section. Probabilistic causation started from the observation that causes raise the probabilities of their effects, but to distinguish evidential from causal difference-making turned out to be an insurmountable difficulty, such that probabilistic causation is now a dead research paradigm (see Pearl 2000)). The counterfactual theories of causation tie causation to counterfactual propositions. Hardly anyone doubts that counterfactuals play some role in causation, and even those theorists that explicitly endeavoured to get along without them (Salmon (1994)) had to retract their original attempts. But the evaluation of counterfactuals either require starting from an elaborated calculus (Lewis (1973)) or are extensions of existing frameworks to accommodate the token level counterfactual case (Pearl (2000), chapter 7). Although this is no argument against the validity of these theories, they do not satisfy my desideratum of providing a simple conceptual basis from where to proceed with a conceptual reconstruction of causation.

With this the manipulationist theories remain in the set of theories to choose from. Manipulation offers a simple notion of telling apart correlations from instances of causation, at least in order to rule out those cases when spurious correlations are due to a common cause. Imagine a simple arrangement of two observables, a switch and a light bulb. A manipulation, leaving open at this stage whether due to agency or due to an objective intervention, targeting the switch, enables the manipulator to control the state of the light bulb. We can introduce the slight complication that the system behaves probabilistically, in order to accommodate a wider range of cases. With the modification, the light bulb sometimes changes its state in any of the two possible directions, going on from off, or vice versa. And even the switch can be allowed to change states by itself. Still, after a number of trials, a manipulator can tell whether she feels empowered to control the light via the switch to some degree, or not. I argue that this is a sufficient criterion to underpin a causal judgment.

A manipulator in such a setup does not need to have any concrete idea of a mechanism connecting the two observables. We can think of some kind of lever and some other observable manipulated by the lever as in no visible or even conceivable way connected to each other. The feeling of control would still be there, as long as the manipulator receives the information about the change of state of the indirectly manipulated observable. That is why we can at this stage, while we are looking for simple criteria to underpin a causal judgment, rule out mechanisms.

There are some further observations corroborating that this approach might serve as an appropriate starting point of conceptual reconstruction of causal judgments. Consider the case of watching a Harry Potter film. Scenarios that feature the casting of a magic spell always involve the wielding of a wand, sometimes underpinned by the recitation of Latin phrases, and then something else at another place happens. Although we have no idea of the mechanism connecting the two events, we are able to tell that the spell caster was the one causally responsible for the subsequent occurrence. We therefore judge that a causal relation obtains between the two events, although a magical causal connection can be seen as the very opposite of a mechanistic connection, in which case we would probably deny that magic was at work. In that sense, the example is even more strikingly in favour of a manipulationist view of causation than cases in which we are, for now, just unable to identify some mechanism. Causation as manipulation thus seems to survive the journey to a counterfactual world like Harry Potter's.

A similar point concerns those scientific-mechanistic explanations that have supplanted the unscientific attempts of explaining natural events based on the workings of god(s) or other anthropomorphic powers. But causation has not been introduced by these new explanations; rather it is the case that mechanistic causes have merely replaced gods as causes.

An observation related to these previous examples concerns what one might want to call, echoing Menzies (1996), the 'platitudes of causation', except that, for now, they are merely platitudes of manipulations. These are the locality of manipulations (the target of the manipulation is distinct from other observables), its asymmetry (there is manipulability only from the cause to the effect, but not vice versa), and its regularity (for probabilistic cases, manipulability implies that the cause is *generally* an appropriate means to bring about the effect; for deterministic case, this would be true *a fortiori*). I will, in section 6.3, try to recover all three platitudes as properties of causation, via an objective description of agency and its subsequent identification with causation.

At the current stage of the argument, I am implying that manipulation is at least a specific case of causation. But even against this weak claim the scenario of Newcomb's paradox (Nozick (1969)) can be held. In this decision game, a player is given the choice of taking home either a box whose content he cannot see, or that box and an additional second box, visibly containing a thousand dollars. The opponent player consists in an intelligence that can predict the first player's choice with absolute precision, and it will make sure that the opaque box, which is part of the prize in any case, is empty if the first player chooses greedily, i.e. chooses to take both boxes. Now it seems that the player's choice is a manipulation that affects the state of the second box – it is empty given the greedy choice, and it contains a million dollars given the modest choice. The conditions of manipulability of the effect (the state of the second box) via the cause (the choice) are met; therefore the two events must be related as cause and effect. But this contradicts the causal intuition that the state of the second box has already been fixed before the moment the choice is made. Hence the manipulability criterion conflicts with a strong intuition about causation, which says that a manipulation cannot influence the past, i.e. the moment when the prize money was put inside the box. Newcomb's paradox does not undermine the manipulability criterion, though. Manipulability implies that an agent can *arbitrarily* manipulate the target – in fact, that is what an agent should do given it wants to make sure the effect stems from its action, not from a possible common cause of the action and the putative effect. But Newcomb's setup primarily considers a hypothetical *rational decision* problem. It asks us what we would do, given that we are told the rules

of the game and shown our options. Replace the situation's pondering of the dominance principle against the expected utility principle by a 'decision' to abide by the result of a coin toss, or any other device that produces genuinely random result, and a predicted decision will become an inconsistent idea, given we assume an ordinary structure of time without backwards causation. Backwards causation would mean that the choice would cause the prediction to *have been* according to the chosen alternative, which does, however, not contradict a causal interpretation of manipulability.

That being said, I do not consider further the possibility of backwards causation at any stage of my argument. Next to the idea of logical or mathematical dependence, considered as an idea independent of causation, I also take the arrow of time to be an extra-theoretical constraint from the point of view of a judging agent. That means that the causal arrow is logically aligned only to the arrow of thermodynamics, but not to the temporal arrow (cf. Price and Weslake (2009), for a discussion of the relationships between the different arrows).¹⁵ With the temporal arrow fixed, we can assume the past to be fixed and out of reach for any agent-manipulations. For the Newcomb's paradox, this entails that the case of an intervention causing something *to have been different*, can be excluded as a possible alternative. In this regard I concur with the defeatist stance of Tim Maudlin, who also resorts to positing the reality of the directed time as an additional constraint in Maudlin (2007).

¹⁵ I am emphasising that the connection which I see between the arrow of thermodynamics and the arrow of causation differs from how Huw Price thinks about that connection. See sections 6.2.2 and the summary section 9.2 on that issue.

4.2 Analysis of Markovian models of causation

In this section I will examine an important class of probabilistic difference-making theories that also allow for the integration of a notion of intervention. These are the theories of causal Bayesian networks and structural equation models, which both obey the causal Markov condition. Since I am interested in the connection between causation and manipulation, I will scrutinize in particular, but not exclusively, the work of the authors James Woodward and Daniel Hausman, who believe that the causal Markov condition is grounded in our notion of intervention. I will keep an eye on the conceptual problem of causation, but in this section I will also deal with the epistemological and ontological problems, in order to assess the relevance of all aspects of the theories and for a better assessment of the strengths and weaknesses of the remaining theories to choose from. After discussing some canonical problems of the causal Markov condition I will examine whether this condition can be said to follow from some concept of intervention, which will be answered in the negative. Instead, the so-called ‘modular systems’ complying with the causal Markov condition are *particular* causal systems that satisfy additional constraints not following directly from the idea of an intervention.

4.2.1 The causal Markov condition

In most examples of causal relations, causation entails a correlation between the causal relata. Probabilistic theories of causation rely on this fact to formulate causal relations by means of conditional probabilities. The probability of an effect is raised or lowered if one conditions on one of its causes, depending on whether the cause brings about or prevents the effect, respectively. Two effects of a common cause are, barring specific circumstances, correlated with each other. However, the principle of the common cause (Reichenbach (1956)) implies that, with such a structure, the cause screens off one effect from the other. Likewise, distal causes are screened off from effects by conditioning on proximate causes of the effect. Both cases comply with the intuition that further information about other effects or distal causes does not increase our expectation of the occurrence of the effect, once we are informed about its direct causes.

Reichenbach considered the relata of a causal relation to be *event types*. Common causal language is generally more flexible, interpreting the relata also as objects, properties, or facts, whereas Causal Bayesian Networks (Pearl (2000), p. 21; henceforth *CBN*), which share some of the principles of probabilistic theories of causation, uniformly represent relata as random variables. Since *CBN* are directed acyclic graphs, the notation of graph theory is used to express the pivotal structural condition of a causal model, the so-called Causal Markov Condition (henceforth *CMC*). *CMC* enables one to express, in a single formula, the structural constraints of unconditionally correlated effects of common causes, the principle of the common cause, and the screening-off of distal causes. One can make use of these probabilistic constraints to infer, to some degree, the causal structure on the basis of observational data. However, in the general case, the correct causal graph cannot be identified from a set of possible graphs that could all explain the observed data, unless more information is gathered or additional assumptions are made. Among the latter we find temporal data, prior causal background knowledge, knowledge of mechanisms, or interventional data. Except for interventions, these sources of information about causal structure circumvent the epistemic problem of inferring the model from data, while interventions presuppose additional causal knowledge, which also concern the problem of when to presuppose that *CMC* holds, as the further analysis will show.

CMC is central to the work of Spirtes, Glymour et al. (1993), who formulate the condition as follows:

Let G be a causal graph with vertex set V and P be a probability distribution over the vertices in V generated by the causal structure represented by G . G and P satisfy the Causal Markov Condition if and only if for every W in V , W is independent of $V \setminus (\text{Descendants}(W) \cup \text{Parents}(W))$ given $\text{Parents}(W)$. (p. 29)

Hausman and Woodward (1999) formulate a similar, but weaker, condition, which they also call causal Markov Condition (abbreviated by *CM*). Note that Steel (2006) discusses the ramifications of the deviation from the formulation of Spirtes et al. These are significant in some contexts, but not relevant to this chapter and will be disregarded in the following pages. Here is Hausman and Woodward's formulation:

For all distinct variables X and Y in a variable set V , if X does not cause Y , then $P(X|Y \wedge \text{Parents}(X)) = P(X|\text{Parents}(X))$. (p. 523)

In this definition, $\text{Parents}(X)$ denotes the subset of V containing all the direct causes of X in V . In order to make explicit the implications of *CM*, Hausman and Woodward (1999) formulate two corollaries of the *CM*, which they call *CM1* and *CM2*:

CM1 If X and Y are probabilistically dependent, then either X causes Y or Y causes X or X and Y are effects of some common cause Z in the set of variables V .

CM2 If $\text{Parents}(X)$ [...] is non-empty, then, conditional on $\text{Parents}(X)$, X is probabilistically independent of every variable except its effects. (p. 524)

The two corollaries express the unconditional dependences and the conditional independences of causal relations, respectively. *CM1* implies that both cause and effect and two effects of a common cause are correlated, whilst *CM2* expresses the principle of the common cause and screening-off of distal causes by proximate causes. The structural condition derived from *CMC* by conversion is called *faithfulness* (Spirtes, Glymour et al. (1993), p. 13), or *stability* (Pearl (2000), p. 48). This condition implies that the only independences, in a causal model, are those that can be derived from the graph. For example, a *CBN* that features faithfulness has no unconditional independence between two variables of which one is the descendant of the other, a property that would not necessarily follow from *CMC* alone. *CMC* requires a condition called *causal sufficiency*, which posits that all common causes of variables in V are themselves in V .

Often, *CBN* are discussed in connection with a modelling technique called structural equation models (*SEM*). A *SEM* consists of a set of equations whose variables are evaluated in an order that reflects the causal influences between the observables of the modelled system. If the

equations are all linear, a parametric description of the system can be written in the following form:¹⁶

$$X_1 = u_1$$

$$X_2 = a_{21} X_1 + u_2$$

...

$$X_n = a_{n1} X_1 + a_{n2} X_2 + \dots + a_{n(n-1)} X_{n(n-1)} + u_n$$

The X_i are the causal variables of the model, the a_{ij} represent the strength of the causal influence of X_j on X_i , and the u_i are error variables, which encapsulate causal influences that are not part of the model. A *SEM* requires knowledge of the functional relations between variables, and thus it is often modelled on the basis of further background assumptions that are reasonable in the given domain. A *CBN* can also be constructed on the basis of conditional probabilities alone. Apart from this representational issue, the two models can be treated as equivalent within finite ranges of the continuous variables. A proof of the equivalence of the two formulations can be found in (Druzdzel and Simon (1993)), and, accordingly, many authors, including James Woodward, Daniel Hausman, Judea Pearl, and Jon Williamson, treat the two paradigms as alternative but equivalent formulations of the same set of problems. If a causal graph is constructed on the basis of a *SEM*, the compliance of the graph with *CMC* requires assumptions about the error terms of the equations. In models with deterministic causal relations, *CMC* is valid only if the error terms are independent of each other in all combinations.

Different constraint-levels can be formulated in the context of a *SEM*, and they reflect different levels of expressiveness of a corresponding causal graph. Higher expressiveness means that more information about the system can be read off the graph. Those constraint-levels can be called *level invariance*, *modularity* and *coefficient independence* in order of ascending expressiveness, so that the latter level implies the former. Notice that Hausman and Woodward (1999) use the same denotations with slightly different interpretations; see footnotes concerning *level invariance* and *modularity*. Here is a brief presentation of the levels.

Level invariance of an equation states that the equation remains invariant under interventions on the underlying system. If the functional expression does not represent correctly the direction of the causal influence, or if a variable is intervened on for which there is an unrepresented causal connection that passes through this variable, then level invariance is not satisfied. In the following sections, I will denote by '*LI*' the constraint-level at which every equation of the model is level-invariant.¹⁷

¹⁶ The parameters are indexed according to matrix notation. The left and right subscripts of the parameters denote the indices of the corresponding free variables and dependent variables respectively.

¹⁷ Notice that *LI* is to be understood as the 'constraint-level of level invariance' rather than 'level invariance' simpliciter, as defined by Hausman and Woodward. They define level invariance as a property of an equation, not as a property of a system of equations. Also, notice that a better choice for this condition would probably be to call it 'invariance under intervention', since 'level invariance' suggests that the equation remains invariant if one of the right hand variables changes its value no matter for which reason. Probably, Woodward and Hausman have refrained from calling it 'invariance under intervention' because their notion of an intervention already requires, as they believe, a modular system, which is subject to a stronger constraint. I have chosen to keep things simple and stay closer to Woodward and Hausman's chosen terminology.

Modularity (MD) says that in every equation of a SEM, a left hand variable can be set to a specific value within an allowed range of possible values, without disturbing any other equation of the system, including those equations where the intervened on variable appears as a free variable (on the right hand side). This is often interpreted by saying that all variables are determined by independent mechanisms, which can also be independently manipulated.

Coefficient invariance (CI) entails that no coefficient in the system is a function of another coefficient, in particular, coefficients within a single equation of a SEM. If this condition is satisfied, the causal model represents disjunctive causes exclusively.

On the basis of the technical vocabulary defined above, we can now consider the list of problems affecting CBN and CMC.

4.2.2 The problems of Causal Bayesian Networks and the Causal Markov Condition

In what follows, I will distinguish between three types of problems: epistemic, conceptual and ontological. Subsequently, I will analyse a list of canonical problems identified in the literature and map them to the three aforementioned types.

The *epistemic problems* consist in inferring the correct model of a system with unknown causal structure on the basis of different sources of information and structural assumptions. For the purposes of this section I will define the *conceptual problems* more generally than in the preceding and following chapters, as consisting in analysing the concept of causation and formulating criteria that determine a relation between two observables as a causal relation. This version is a bit more general since it is detached from the context of grounding an agent's judgment concerning a causal relation. Finally, the *ontological problems* address the question of whether causal relations are a feature of the real system as opposed to the model that represents it.

When a causal structure is to be inferred on the basis of observed probabilities, additional assumptions must be made, like relying on CMC and the related conditions of faithfulness and causal sufficiency. The common characteristic of the problems denoted by 'epistemic' is the deviation of the inferred model from the correct model. One should notice that, in such cases, there is no controversy involved about what the correct model, e.g. a causal graph, should look like. This contrasts with the conceptual problems. Those concern the question of what characterises a causal relation, which, in turn, has a bearing on what a correct causal model of a specific system should look like. The conceptual problems are partly descriptive, partly normative. They are descriptive in so far as formal definitions of criteria of causal relations attempt to capture what we naturally mean by 'causation'; they are normative in so far as, concerning more controversial aspects of causation, and especially the notion of modularity, a decision on the conceptualisation is required. The ontological problems are speculative from the perspective of graph theory and statistics, since they concern additional assumptions made about the properties of causal systems that can neither be expressed in those kinds of models nor be derived from the measured data. For example, this concerns the question whether there is irreducible causal indeterminism. We can discuss conceptually both deterministic and indeterministic causation, as long as the concepts are coherent. Whether the former or the latter is, ontologically speaking, vacuous, is a different issue. Therefore, it is crucial to distinguish the epistemic problems from those problems that concern conceptual disagreement, and, again, the latter from the problems that depend on ontological commitments.

Based on the conceptual clarifications given so far, the following two lists of problems of CBN and CMC found in the literature can be discussed. Together, they provide a basis for a unified taxonomy. While Hausman and Woodward (1999) consider problems affecting CMC, Cartwright

(2004) presents a list slightly broader in scope, by considering problems concerning *CBN* in general, which can result from breaches of *CMC*, faithfulness or causal sufficiency. Hausman and Woodward defend *CMC*, whereas Cartwright argues that *CMC* cannot be endorsed as a general structural condition for causal systems.

According to Hausman and Woodward's formulations, the list of cases where *CMC* encounters problems can be given as follows:

HW1: the probabilistic dependency between X and Y might be merely accidental;

HW2: a dependency may arise when one mixes two subpopulations in which X and Y are causally independent;

HW3: dependencies that do not reflect causal connections may arise when the wrong variables are measured; and

HW4: correlations in quantum mechanics appear not to have any causal explanation.

According to Cartwright's formulations, the list reads as follows:

C1: positive and negative effects of a single factor cancel each other;

C2: factors can follow the same time trend without being causally linked;

C3: probabilistic causes produce products and by-products;

C4: populations are over-stratified (e.g. they are homogeneous with respect to a common effect of two factors not otherwise causally linked);

C5: populations with different causal structures or different probability measures are mixed.

Notice first that *C2* is a particular instance of *HW1*, and that *HW2* directly corresponds to *C5*. Next, it will be shown that there are two basic categories of problems: *HW1/C2*, *HW2/C5*, *C1* and *C4* are epistemic problems, whereas *HW3*, *HW4* and *C3* will be discussed as predominantly conceptual or ontological problems, although they also have an epistemic aspect. The first category, the predominantly epistemic problems, show the shortcomings of *CMC* when it is part of an algorithm of inferring causal relations from data. Regarding the question of the importance of *CMC* to the concept of causation, this category of problems can be dealt with in less detail. *C3* and *HW3* require the most comprehensive philosophical discussion, the result of which will be to show that three relevant constraint-levels of causation can be formulated, of which only the two stronger ones satisfy *CMC*. The weakest constraint-level does comply with an intervention-based concept of causation; however, it does not imply the 'arrow-breaking' (Steel (2006)) account of interventions. *HW5* refers to the so-called EPR-paradox (Einstein, Podolsky et al. (1935)) and involves paradoxical implications of a seemingly causal connection between the measurements of two entangled particles at distant locations. Since this problem requires an in depth discussion of possible interpretations of quantum mechanics, and since it seems to concern predominantly the ontological level, I will not discuss it in this chapter.

4.2.3 The epistemic problems

The following list is a categorisation of the epistemic problems that underscores the common

features of the problems belonging to a specific group:

EP1: the basis of causal inference consists of a probability distribution table that can be, under general assumptions, considered ‘improbable’. Such a table leads to wrong conclusions about the underlying structure that produces the statistical data.

EP2: the set of considered variables itself is not informative enough. More variables must be incorporated into the causal model to prevent a distorted causal picture.

Other epistemic problems (which are not part of the aforementioned list from Hausman, Woodward and Cartwright): There can be other reasons for why the structural constraints of observational data alone are insufficient to infer the correct causal graph. Among those are observationally equivalent causal structures, or deterministic structures which do not allow the screening-off of spurious effects

The following table provides an overview of the categorisation based on the introduced nomenclature:

Category	corresponding canonical problem in Hausman and Woodward's list, or Cartwright's list	related epistemic problems
<i>EP1</i>	<i>HW1/C2, C1</i> (given the cancellation of positive and negative effects is accidental)	other forms of breaches of faithfulness of data
<i>EP2</i>	<i>HW2/C5, C4, C1</i> (given the cancellation of positive and negative effects are designed or brought about by selection bias)	breaches of causal sufficiency
<i>Other epistemic problems</i>	-	observational equivalence, deterministic structures with mutual screening-off of spurious effects

Table 1: Overview of the epistemic problems

The common feature of all *EP1*-problems is the lack of informativeness of the probability distribution governing the observed variables with respect to the relations of probabilistic dependence and independence. *EP1* captures accidental dependencies, breaches of faithfulness of data, and accidental cancellation of positive and negative effects. With *EP1*-problems, the epistemic agent cannot rely on the probabilistic relations alone to infer the correct target model. *EP2* captures breaches of causal sufficiency, mixing and over-stratification. It also captures cancellation of positive and negative effects, given that those effects are not accidental, but caused by selection bias or purposeful design, respectively. All *EP2*-problems feature a set of considered variables that result in a distorted causal model.

Given their nature, the epistemic problems can be solved on the basis of further background information and further assumptions, but not on the basis of assuming *CMC*, faithfulness, and causal sufficiency alone, which merely enable the exploitation of probabilistic dependence and independence. Incorporating additional assumptions, or resorting to other sources of information

like causal background knowledge, interventional data, temporal data, enables the correction of the model structure (e.g. rejection of a dependence as non-causal, or reversal of a direction).¹⁸

In *EP1*, the epistemic agent is misguided in applying *CMC* and the related structural conditions to the probability distribution, because a deviation of the correct model from these conditions is taken to be 'improbable'. The improbability of the data can be either due to accidental dependencies between the observables, or because the system configuration itself is unlikely. Accidental dependencies arise if a sample-correlation does not represent a population-correlation. Either those non-causal dependencies are genuinely accidental, or they are due to a selection bias. The latter problem, however, belongs to *EP2*, not *EP1*, since the experimenter, who selects the observational data, exerts a causal influence on the variable values, and should therefore be considered as part of the causal model.

Besides unconditional, non-causal dependences, misleading data can result from mixing two heterogeneous sub-populations. Seemingly paradoxical results of such a mix have been described in statistics as 'Simpson's paradox' (Simpson (1951)). The particular consequence for causal relationships is that a statistical indication of dependence can turn into independence. The reason for that is that *faithfulness* of the data is assumed, and unconditional independence of causally related variables is not expected. Nevertheless, it can result from the influence of the third variable. Either the third variable is a common cause of the other two causally related variables (*EP1*, example 2), or it is part of a second causal path between the other variables, which is Cartwright's problem of 'cancellation of positive and negative factors' (*EP1*, example 3). In both cases, misled by the assumption of faithfulness, plausibility reasons countenance the false inference that the third variable might be a common effect of two independent causes (see Spirtes, Glymour et al. (1993), page 41).

¹⁸ It follows that the discussion of epistemic problems given here merely covers their classification, and not how they are solved. Solutions to the problems exist, but the scope of this section is neither the possible extensions of the Markov condition, nor its backing up with further techniques, in order to infer correct models. Likewise, a discussion of why the target models are the correct ones, would be out of place. In each of the individual cases, the target models, from which the inferred models differ, can be considered correct.

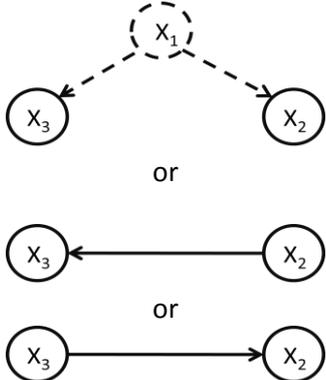
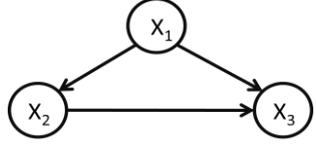
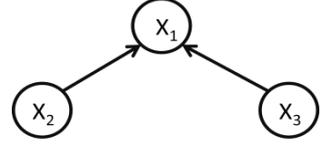
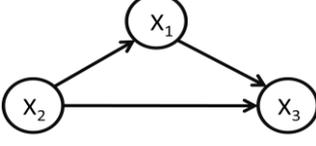
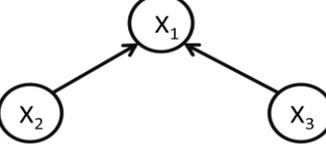
EP1	
Correct Model	Inferred Model
<p>EP1, example 1: accidental dependencies</p> 	
<p>EP1, example 2:</p> 	
<p>EP1, example 3:</p> 	

Table 2: Examples of EP1

A characteristic of *EP2* is an unobserved confounding variable. In such cases, a positive causal influence appears as if it were a preventative, and vice versa. Either the unobserved variable is a common cause, which is a breach of ‘causal sufficiency’ (see *EP2*, example 1) or the unobserved variable is the effect of one of the observed variables and the cause of the second (*EP2*, example 2). *EP2*, examples 3 and 4, are instances of *Simpson’s paradox*, due to ‘mixing two heterogeneous subpopulations’. Simpson’s paradox is given a causal explanation (as in Pearl (2000), chapter 6), because the population-variable is causally linked to the other two variables. In both examples, the two observed variables appear to be independent, although they are actually causally linked. As with examples 1 and 2, this can be due to the omission of either a common-cause-variable (*causal insufficiency*), or of a second directed path that includes the confounder.

In example 5, the inferred causal structure appears to be level invariant, but in fact it is not, as it can be seen once X_1 is intervened on. The model, which merely represents one causal

connection between X_3 and X_2 , could not make sense of such an expected behaviour of the system.

If a causal system features cancellation of positive and negative factors, then, if the cancellation happens accidentally, the criterion of faithfulness, but not the criterion of causal sufficiency, is breached. However, if the cancellation is brought about on purpose, the problem can be dealt with similarly to the problem of selection bias, which would be an instance of causal insufficiency with respect to an extended 4-variable causal graph, including a variable for the manipulator. Over-stratification can be dealt with similarly to causal sufficiency. Yet it is not a common cause, but a common effect which is conditioned on. This is the reverse of EP2, example 3 – two independent causes appear to be causally related.

EP2	
Correct Model	Inferred Model
EP2, example 1 	
EP2, example 2 	
EP2, example 3 	
EP2, example 4 	
EP2, example 5 	

Table 3: EP2, featuring three variable problems with the confounder absent from the model. A causal arrow with '+'-sign denotes a positive causal influence, a '-'-sign denotes a negative connection

Successful application of the screening-off criterion requires non-degenerated conditional

probabilities. Different values of different variables must occur in a sufficient number of combinations in order to have the proximate cause screen off distal causes and secondary effects from any single direct effect. If the data features such kinds of combinations, the correct undirected graph and all the v-structures (Pearl (2000), p. 19) can be inferred. This is the case of *table 4, example 1*. The directions of the arrows in the causal graph, with the exception of v-structures, are non-decidable if only observational (non-experimental) data are available. If one cannot condition on sufficient combinations of values, the joint probability table is not informative enough to allow the inference of the correct skeleton. Such is the case with deterministic causal relations without unrepresented additional causes. Observational data alone then do not allow holding one of the three variables fixed, while varying the other two. The result of screening-off tests in such a case is mutual screening-off in all combinations, from which no structure can be inferred. *Table 4, example 2* presents yet another case, one in which we happen to know the causal direction between three variables. In such a situation, a minimal candidate of a causal model would be incorrect, because it will leave out a possible direct causal connection from X_2 to X_3 . Such a case of breach of level invariance is described in Hausman and Woodward (1999)p. 543).

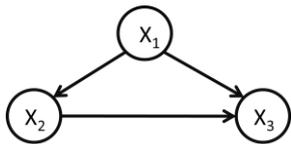
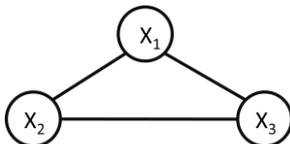
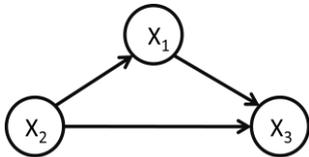
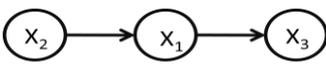
other epistemic problems	
Correct Model	Inferred Model
<p>Example 1</p>  <pre> graph TD X1((X1)) --> X2((X2)) X1((X1)) --> X3((X3)) X2((X2)) --> X3((X3)) </pre>	 <pre> graph TD X1((X1)) --> X2((X2)) X1((X1)) --> X3((X3)) X2((X2)) --> X3((X3)) </pre>
<p>Example 2</p>  <pre> graph TD X2((X2)) --> X1((X1)) X2((X2)) --> X3((X3)) X1((X1)) --> X3((X3)) </pre>	 <pre> graph LR X2((X2)) --> X1((X1)) X1((X1)) --> X3((X3)) </pre>

Table 4: Examples of other problems classified as epistemic

4.2.4 The ontological and conceptual problems

The ontological and conceptual problems will be discussed in direct correspondence to the problems C3 and HW3:

Category	corresponding canonical problem in Hausman and Woodward's list, or Cartwright's list
indeterministic causation with joint effects	C3
measuring the wrong variables	HW3

Table 5: Overview of the ontological and conceptual problems

4.2.4.1 Indeterministic causation with joint effects

As we saw before, two distinct corollaries can be formulated on the basis of *CMC*, which Hausman and Woodward called *CM1* and *CM2*. It is instructive to formulate two subsequent corollaries based on *CM2*, one concerning screening-off of distal causes, the other concerning screening-off of joint effects:

CM2.1 with respect to knowing whether the effect occurs, knowing a distal cause does not add information to knowing a proximate cause, unless there is an additional causal connection from the distal cause to the effect;

CM2.2 knowing additional effects of common causes does not add information to knowing the common cause.

The problem discussed in this section concerns essentially *CM2.2*. The corollary of *CMC* could be further explained as follows: additional effects are not *causally relevant* to one considered effect, but knowing about an additional effect, so one might think, could contribute to a better prediction of whether a considered effect has been brought about. This, however, would contradict the causal Markov condition, which posits screening-off once we condition on the common cause. *CMC* makes no difference between causal and informational relevance, and therefore, negates informational relevance of other effects.

Probabilistic causation involving products and by-products and its implications to *CMC* have been brought to attention by means of an example featuring a factory in Cartwright (1993). The factory produces two chemicals, a product and a by-product, according to a genuinely probabilistic process. The example therefore involves three random variables C , E_1 and E_2 , denoting the cause and two effects, respectively. The specification of the probabilities is set as follows: $P(E_1, E_2|C) = 0.8$ ¹⁹, $P(\text{not}(E_1), \text{not}(E_2)|C) = 0.2$, therefore $P(\text{not}(E_1), E_2|C) = P(E_1, \text{not}(E_2)|C) = 0$. According to the background story, the probability of $P(C)$ is implied to be around 0.5. $P(C)$ is required as an independent piece of information, since C is designated as the exogenous cause-variable. Likewise, the probabilities for the two effects conditioning on $\text{not}(C)$ would be needed as independent parameters for a full specification, but these pieces of information are not relevant to the argument against *CMC*, which is already implied by the given partial specification. The example therefore

¹⁹ $P(C)$ and $P(\text{not}(C))$ is shorthand for $P(C=\text{true})$ and $P(C=\text{false})$, respectively.

features four independent probability parameters, and inferring the causal relations based on *CMC* would yield a graph with two causal connections, one between E_1 and E_2 , and one arc connecting either E_1 with C or, E_2 with C . However, since the cause-effect relationships are provided with the example, and are set to be (C causes E_1) and (C causes E_2), the correct causal graph features, rather than the inferred structure, a fork-like structure with an outgoing arrow from C to E_1 and E_2 , respectively, and no arrow between E_1 and E_2 . The factorized form of the joint probability function, causally ordered, is therefore $P(C, E_1, E_2) = P(C) P(E_1, E_2|C)$. This factorized form is not further reducible since there is no screening-off between E_1 and E_2 . The cause-effect relationships are given, thus the example does not expound the problems of inferring causal relations of any kind.²⁰ What is instead provided by the example is an already interpreted causal system, represented by a causal model that does not comply with *CMC*.

The reason for claiming that the causal structure of the example is a violation of *CMC* is given by the inequality between $P(E_1|C, E_2) = 1$ and $P(E_1|C) = 0.8$, which follows from the probability specification. To defend her counterexample, Cartwright (1993) argues that the causal relations are specified in the form of $P(E_1, E_2|C)$ and not by $P(E_1|C)$, $P(E_2|C)$, independently. Thus, since the independence of the two effects E_1 and E_2 , conditioning on C , only follows for the special case of

$$P(E_1, E_2|C) P(\text{not}(E_1), \text{not}(E_2)|C) = P(E_1, \text{not}(E_2)|C) P(\text{not}(E_1), E_2|C),$$

violation of *CMC* follows, because the indicated equality is not reconcilable with the probability specification. Cartwright (1993) suggests that no further constraints can be imposed on the data by saying that 'nature must fix a joint probability over the whole event space. [...] Nothing in the concept of causality or probabilistic causality constrains how it should be done.' The latter statement, however, is exactly what is at stake, and could be refuted according to causal concepts that differ from the one that Cartwright seems to favour.

The factory example has been criticised, in order to re-instantiate *CMC*, on the following grounds:

1. A different interpretation of probability can be employed for which *CMC* holds.
2. The causal specification is alleged to contradict the probability specification because product and by-product are not distinct variables, which is said to be required by *CMC*.
3. The example can be attacked on speculative grounds, because:
 - a. it can be claimed to be vacuous; or
 - b. it is stipulated that a more informative set of variables can always be found in cases analogous to the factory example, and employing such a kind of variable set restores the *CMC*
 - i. either by incorporating a further variable that turns the fork of the original causal model into an *interactive fork*,
 - ii. or by incorporating two *pseudo-hidden variables*, thereby showing that the example violates the assumption of independent error-variables.

Concerning 1., one can directly criticise the conclusion that the example is a counterexample against *CMC* by interpreting the probabilities in *CMC* as pertaining to physical chances. This replaces evidential relevance by causal relevance for the evidence variable that is conditioned on. The

²⁰ If it did, the problem would fall into class *EP2*, since a wrong structure would result from assuming the causal Markov condition.

distinction between causal and informational relevance relates to what are sometimes called interventional and evidential probabilities (e.g. Sloman (2005)). Given this interpretation of probability, $P(E_1|C, E_2) = P(E_1|C)$ also holds true for Cartwright's factory, because E_2 is of no causal relevance regarding the chance of E_1 occurring. This is obviously different from being a factor in chance-fixing E_1 . Of course, $P(E_1|C, E_2) > P(E_1|C)$ continues to hold true in an interpretation of informational relevance, in accordance with the probability specification. Hausman and Woodward do indicate the option of interpreting the factory in this way in Hausman and Woodward (1999), but they reject this kind of interpretation of *CMC*.

Hausman and Woodward insist that causal and informational relevance are the same because of an alleged incompatibility with the condition *CM1*. Accepting the full specification of the example is therefore not an option in an attempt to defend *CMC*. According to their argumentation, either a more informative cause specification can be found, from which E_1 or E_2 follow deterministically, and in which case *CMC* would be true (see also 3.b.i.). This corresponds to the solution by the so-called 'interactive fork' (see Pearl (2000), page 62). Or the causal dependence of (E_1, E_2) on C is genuinely indeterministic. In this case, it can either be posited that E_1 and E_2 are not distinct, or they are distinct and therefore E_2 can be intervened on without directly influencing E_1 . They subsequently undertake a case distinction for the latter case between the two relevant possibilities: $P(E_1|C \& \text{set } E_2) = P(E_1|C \& E_2)$, or $P(E_1|C \& \text{set } E_2) = P(E_1|C)$, where 'set E_2 ' stands setting rather than observing E_2 . Both possibilities are subsequently rejected.

Arguably, the distinction between causal and informational relevance occurs more naturally if, in contradiction to the usual approach of theories of probabilistic causality (see Hitchcock (2002) for an overview), a causal concept attributes a crucial role to state transitions occurring in time. It is according to such a notion of causality that E_2 can provide additional information about E_1 , due to the property of being a document of a probabilistic process that has, in fact, happened, rather than being a factor in chance-fixing the event that is about to happen. That being said, this way of restoring *CMC* does not solve the epistemic problem of uncovering the structure for a system whose causal specification is unknown, rather than conceived, as in the case of the factory.

Concerning criticism 2 in the above list, one part of the argument against the difference of causal and informational relevance used in criticism 1 hinges on the idea that E_1 and E_2 might not be distinct variables. If this can be proved to be true, it constitutes an independent argument against the example, which would be violating the correct specification of variables. Hausman and Woodward see a strict connection between *CMC* and independently manipulable mechanisms in which one of two effects of a joint cause can be brought about by intervention without influencing the mechanism between the cause and the second effect. They call this assumption 'MOD' in (Hausman and Woodward (1999)), and it states the following: a direct intervention on one variable invalidates the equation that defines the variable as dependent, but leaves all other equations intact, including those in which the variable intervened on appears as an independent variable. If such an intervention is possible, then *CMC* is satisfied. MOD, if applied to a graph-theoretic model, implies an 'arrow-breaking' notion of an intervention. Following this account, if an effect is brought about by a mechanism that hinges only on the cause-variable, then another effect of the same cause that is brought about by *another, independent* mechanism cannot bear any relevance on the first effect, and hence, conditioning on the common cause, the second effect should screen off. If this is not the case, then the mechanism is indeed one and the same for both effects, which should then be represented as a joint variable rather than two distinct variables. The example they employ for such a kind of joint variable is the atomic decay, which produces the emission of two protons and two

neutrons that, together, form the joint variable of an alpha-particle.

Hausman and Woodward (1999) propose a proof that *MOD* implies *CMC*. Cartwright (2002) rejects this proof by showing that *MOD* does not bear any relevance in the derivation of *CMC* from the additional assumptions, from which *CMC* follows anyway. The disagreement, which the factory example can help to retrace, ultimately results from different notions of an intervention, the crucial concept implicit in *MOD*. Applying the *MOD* condition to a parameterised model of the factory specification underscores that *MOD* does not bear on the validity of *CMC*, if *MOD* comes with the wrong notion of intervention. The model in Cartwright (2002) (variable denotation is adapted here for the sake of consistency)

$$\begin{aligned}
 C &= u_c \\
 E_1 &= a_{E1} C + u_{E1} \\
 E_2 &= a_{E2} C + u_{E2}
 \end{aligned}
 \tag{eq. 1}$$

uses two binary random variables to specify the probabilistic connection between *C* and (*E*₁, *E*₂), with the following distribution: $P(a_{E1} = 1) = P(a_{E2} = 1) = 0.8$ and $P(a_{E1} = 1 | a_{E2} = 1) = 1$. The error term *u*_{*c*} is not relevant for the current discussion, and *u*_{*E1*} and *u*_{*E2*} are set to 0 according to the specification²¹. Depending on one's specific notion of a proper intervention, one can argue that *MOD* is applicable to this model – one can intervene on *C* (but, since *C* is the root node of the model, that is the same as conditioning on *C*, and therefore provides no relevant information); the interventions on *E*₁ and *E*₂ are likewise possible, e.g. via manipulating the error-terms. No equation, except the one that determines the variable intervened on, is invalidated. But this intervention is completely uninformative as for determining the independence of mechanisms, since both *E*₁ and *E*₂ only appear on the left side of the equation. Hausman and Woodward disagree with such a conclusion since they endorse a necessary condition of an intervention as defined in Spirtes, Glymour et al. (1993). The necessary condition is the characteristic of the intervention of breaking all causal connections from the previous (pre-intervention) causes to the effect, whose value is set by the intervention-variable, which is typically the, or part of an, error variable. Since there are good reasons to think that the two parameters *a*_{*E1*} and *a*_{*E2*} are not only correlated, but co-referential, it would follow that it is not possible to break the causal connection from *C* to *E*₁ without breaking the connection from *C* to *E*₂. Therefore, while it is possible to vary *E*₁ by varying *u*_{*E1*}, this would not count as a proper intervention.

Independently of which definition of intervention should be applied to *MOD* in the factory example, one could still accept the conceptualisation of distinct variables as effects of independent mechanisms. One would thereby by-pass the epistemic problem and argue conceptually. However, this severely limits the scope of applicability of a *CBN*. Considering the alpha-particle example, the two neutrons and the two protons could engage in different causal relations in descendant nodes in an extended causal graph that traces further causal interactions. It is unclear how the treatment of such products and by-products by means of a joint variable would handle those kinds of

²¹ Alternatively, one can also mathematically model the same situation by correlating the error terms *u*_{*E1*} and *u*_{*E2*} of the two effects of *C* instead of the two parameters *a*_{*E1*} and *a*_{*E2*}, which then would have to be set to 1 unconditionally. That would create the same probability distribution. In order to keep the analysis less complicated, I ignore the variant of correlation the error terms.

circumstances. The strategy of finding a proximate and deterministic common cause of both effects, which might be possible for non-subatomic (e.g. chemical) reactions resulting in product and by-product, does restore *CMC*. However, it conflicts with the underlying semantics of causal paths as independent mechanisms.

A different strategy from the two aforementioned ones is to accept the consistency between the probabilistic and the causal specification, and also endorse the informational interpretation of *CMC*, but to attack the example on ontological grounds.

Concerning criticism 3.a, a straightforward way to attack the probabilistic and causal specification of the example is to deny its occurrence in the real world, as suggested by Glymour (1999). In her response, Cartwright (2002) refers to macro-level examples from scientific practise, which are, in turn, rejected by Hausman and Woodward (2004). Whether those examples are accepted or rejected ultimately hinges on additional definitions of conditions that causal relations must satisfy, and therefore amounts to a conceptual decision.

Regarding 3.b.i, a view which is a bit less strict implies that the kind of structure outlined in the factory example does exist on the level of coarse-grained variables, but that those variables can be replaced by more informative variables that restore *CMC*. This strategy interprets the factory case as an example of an 'interactive fork' (see Pearl (2000)). The probabilistic part of the causal connection between common cause and joint effects is disposed of by formulating an indeterministic relation between the cause-variable and a new, now proximate and common-cause-variable, which is connected, deterministically, to the joint effects via the fork-structure. This view invariably leads to the requirement to apply a *CBN* at the level of deterministic proximate causes for every fork-structure with arrows that do not represent independent mechanisms.

Using the interactive fork, however, is a measure of last resort for all causal theories whose underlying ontology is based on mechanisms, which is the case for Pearl as well as for Hausman and Woodward. Considering causal forks like the one of the factory example, *CM2.2* can be valid due to two very different reasons: either because the causal relations of the fork are deterministic, or because the mechanisms, by which the common cause brings about the two effects, are independent of each other. Employing the interactive fork refers to the former; it makes *CMC*, as Cartwright (2002) puts it, 'trivially true', since the complete information about whether the effect will occur is encapsulated in the cause-variable, leaving no additional informational relevance that the other effects could contribute. Whichever additional variable is integrated into the model in order to produce the interactive fork, it is not guaranteed that this integration turns the incoming arrows into the two joint effects into distinct mechanisms, which is how the mechanistic ontology interprets the arrows. The deterministic relation merely *conceals* the fact that the mechanism might in fact be the same for both effects.

As far as criticism 3.b.ii is concerned, Steel (2005) applies the concept of *pseudo-hidden variables* to the factory example, in order to show that a premise of *CMC*, independence of error-variables, or, according to Steel's terminology, *exogenous variables*, is not satisfied. The concept of pseudo-hidden variables is designed to make deterministic and indeterministic systems more easily comparable. Steel endorses Cartwright's specification according to eq. 1, but interprets the coefficients a_{E_1} and a_{E_2} as parents of E_1 and E_2 , respectively, not as direct causes. This unconventional treatment of causal graphs has certain advantages, since it allows one to apply knowledge about the behaviour of systems in deterministic models to indeterministic ones, but the inclusion of parents that are not also causes of the descendants does complicate somewhat the interpretation of the causal graph. Moreover, identifying the probabilistic coefficients with parents of variables in a graph

is itself an interpretation in which one does not necessarily have to follow Steel. Besides those conceptual problems, Steel's approach, similar to the interactive fork solution, raises the same epistemic problem of identifying the pseudo-hidden variables with concrete observables, such that *CMC* could be restored on their basis.

Very likely, other cases similar to Cartwright's factory involve connections between joint effects that are non-causal. When product and by-product are interpreted as a chemical compound that is decomposed into its parts in the course of a chemical analysis, the relation that results in a probabilistic correlation is not causal. One could argue that graphs, for which the only connections between correlated variables are causally interpreted arrows, will necessarily distort either the causal picture, or lead to the wrong conclusion about the probabilities that the graph produces. Accordingly, one would have to rule out such variables prior to a causal analysis that is based on probabilities. The next section discusses this and similar problems that involve measuring the wrong variables.

4.2.4.2 *Measuring the wrong variables*

Whilst the previous section discussed the problems with the second of the two corollaries of *CMC*, and identified it as the more problematic one, this section is wider in scope and addresses the problem of the appropriateness of the variables in the first place, to which *CMC* can be applied.

Hausman and Woodward (1999) refer to several examples for which wrong variables are part of a causal model. Among them is an often cited example from Salmon (1984), according to which two billiard balls share a quantity of momentum determined by a third ball that is struck by the cue, as well as some other examples that are taken from (Arntzenius (1993)). Examples with an analogous structure can be found in Cartwright (1989) and in chapter 4 of Williamson (2005). What all the examples have in common is a breach of *CMC* due to the screening-off condition failing in the light of non-causal dependences between common effects variables. Those non-causal dependences can be logical, mathematical, semantic, or based on non-causal constraint-relations.

In order to reject those kinds of variables as a valid basis for a causal model, Hausman and Woodward employ the concept of independent manipulability, according to their *MOD* condition. The concept of independent manipulability hinges on the concept of an intervention. The latter can be interpreted as an introduction of an additional cause-variable into the causal model that enables one to set the value of an individual variable. The introduced cause behaves like an error variable in so far as it is independent of all other hidden cause-variables. Also, it switches off the influence of all other explicit causes of the variable it intervenes on. In the structural equation, this relates to substituting the function of all pre-intervention variables that determine the dependent variables by the function describing the newly introduced intervention variable. According to Hausman and Woodward, if mechanisms are independently disruptable, proximate causes screen off both distal causes and additional effects of the proximate causes from their effects, which validates all corollaries of *CMC*. Therefore, a necessary condition of appropriate variables posits that all effects are results of independent mechanisms. If effects are related according to one of the aforementioned non-causal relations, such independently working mechanisms do not exist, and the non-causally related variables should be ruled out or treated as a joint variable instead.

This definition of intervention according to the *EI*-criterion is analogous to the one given in Hausman and Woodward (1999), to which several authors have replied. It is defined as follows:

(EI) A necessary and sufficient condition for a process I to count as an intervention on the value of some variable X possessed by some individual i with respect to some second variable Y is that I changes the value of X possessed by i in such a way that if any change in the value of Y occurs, it only occurs through the change in the value of X and not via some other route. Graphically, this amounts to the requirements that

- (i) I is the only cause of X – all other arrows into X are broken,
- (ii) any directed path from the intervention variable I to Y must go through X ,
- (iii) any directed path from any cause Z of I that goes through Y must also go through X , and
- (iv) I leaves the values taken by any causes of Y except those that are on the directed path from I to X to Y (should this exist) unchanged.

Cartwright (2002) further refines the definition of an intervention by adding the restriction that no mechanisms, or functions, are altered while intervening on a target variable. Even though this extension prevents some wrong conclusions about causal structures that would arise when **EI** is applied, Cartwright rejects the very approach of an operational definition of appropriate variables.

Steel (2006) criticises the arrow-breaking account of intervention, since in practise the other causes that are not part of the intervention will often merely be conditioned on, rather than controlled for. Bogen (2004) expressed scepticism whether atomic interventions on a single target variable are possible in practise. The conditionals in the definition of **EI** are counterfactual. Counterfactual elements in causal theories have conceptual advantages, since it is necessary to cover such examples of causal relations for which it is not possible to intervene on the cause. Also, some of the aforementioned problems are thereby accounted for. But the strategy raises the epistemic question of whether the methodology is logically complete, given some causal structures are unknown a priori, an issue which Woodward, in his alternative **EI**-account of interventions, purports to solve. As Psillos (2002) emphasizes in his analysis, Woodward's amalgamated account suffers from the same problems of circularity as the pure interventionist and counterfactual accounts.

It remains to be seen whether a non-circular account can be developed based on the idea of independently disruptable mechanisms, but it seems doubtful that this is possible without the screening-off criterion implicit in *CMC*, since it provides significant information to distinguish different possible causal models. This may be the reason why many authors are unwilling to give up this structural condition of models that are based on causal graphs and probability distributions.

Williamson (2005) proposes an alternative, epistemic account of dealing with the aforementioned counterexamples against *CMC*, the 'qualified causal Markov condition'. According to this formulation, *CMC* need not be satisfied in all circumstances, and is subject to a causal epistemology of an agent. The qualified *CMC* is embedded in a framework that distinguishes causal belief and probabilities, which are interpreted as degrees of belief. Since the quantitative causal analysis is constrained a priori by the agent's causal belief, which are represented as directed acyclic graphs, the agent is able to rule out the non-causal constraints that would otherwise invalidate the causal inference due to the failure of *CMC*.

Summing up this discussion of *CMC* and the way it is linked to interventions, I can neither agree with Woodward and Hausman's idea that *CMC* is grounded in interventions, nor that a correct causal model requires *CMC*. Modular systems, as can be shown by means of a *SEM*, satisfy the causal Markov condition. A reasonable physical interpretation of independent equations, espoused by Woodward and Hausman, are independently manipulable mechanisms. But simple deterministic

systems, like Pearl's example of '(wet grass) causes (slippery grass)', also comply with *CMC*, and there is by no means a 'mechanism' between the relata that could be manipulated or switched off. On the other hand, the commonsensically correct causal model of Cartwright's factory (see diagram 4.2), which also complies with a manipulationist interpretation of the causal concept, does not require *CMC*. It seems that the idea of independently manipulable mechanisms is a precaution to prevent the conflation of explanatory relevance with causation, but the criterion is too strong and not as strongly linked to *CMC* as Woodward and Hausman hope it is. Modular systems are particularly nice to handle causal systems, but the constraint would be too strong if we take it as a necessary condition for causation. If the conditions of *CMC*, faithfulness, and causal sufficiency are met by the data, the epistemic task of discerning causes and effects circumvents some of the problems that otherwise arise. However, given a variable set that is to be ordered causally, there is a priori no reason to assume that those conditions are generally met, in particular, since *CMC* abstracts from the nature of the causal relata.

4.2.5 Criticism of the objectivist difference-making theories

I have mentioned some problems of various theories already in the course of the overview of theories of causation at the beginning of the chapter. To these I have added the particular problems stemming from the assumption of the causal Markov condition. Now, since I have opted for the manipulationist approach in section 4.1, this section will be the place to distinguish further the theories within this class. The interventionist approach and the agency-approach can be considered as subclasses of manipulationism, as it was suggested by Woodward (2009). I will call the former 'Objective Interventionism' to further stress the aspect in which the two approaches differ. I attribute the authors studying causal Bayesian networks, like Judea Pearl and Peter Spirtes, Clark Glymour and Richard Scheines, and the philosopher James Woodward to the class of Objective Interventionism. In a nutshell, the problem addressed in the remaining parts of the chapter consists in the way the account of interventions can serve as a basis to solve the conceptual problem of causation successfully. I will start with a closer look at the work of James Woodward.

4.2.5.1 James Woodward

James Woodward construes one of the aspects of his theory as 'interpretive or semantic' (see, for example, Woodward (2008), Woodward (2003a), or the introductory chapter and page 38 in Woodward (2003b)). Although he does not, as my thesis does, explicitly tie the idea of such an analysis to interpreting causal *judgments*, he is likewise ontologically uncommitted and suspicious about an ultimately real level at which causation operates (see, for example, Woodward (2007); also page 36 in Woodward (2003b)). His interpretative or semantic aspects seem to refer to disclosing the meaning of certain technical concepts that belong to modelling paradigms that he himself espouses, like causal Bayesian networks and structural equations as considered in the previous sections. The directed acyclic graphs of Bayesian networks feature arrows that assign different roles to the nodes at their heads and tails respectively and therefore indicate a connection that is stronger than just an observable correlation between the two variables. In a completely analogous way, structural equations indicate, besides equality in values, different roles of the terms at the left and at the right hand side of the equation. It seems that Woodward's account of interventionism seeks to provide the semantics of this asymmetry. This motivation of his would be in line with his view that there is a certain lacuna (see endnote 19 of Chapter 2 in Woodward (2003b)) in the accounts provided by Pearl and Spirtes, Glymour and Scheines, whose accounts he endorses to a large extent apart from this perceived lacuna. In his view, interventions, if taken merely as intuitively understood,

primitive notions, cannot be the basis of providing the semantics of a causal arrow in a causal graph, and therefore cannot be the basis of the semantics of the causal relation itself.

It is instructive to look at a correspondence between Woodward and Pearl in (Woodward (2003a) and Pearl (2003)) regarding that issue. Woodward, in his critical notice, first quotes Pearl (2000), according to whom an intervention (dubbed **PI**) can be defined as follows:

(**PI**) The simplest type of external intervention is one in which a single variable, say X_i , is forced to take on some fixed value x_i . Such an intervention, which we call 'atomic', amounts to lifting X_i from the influence of the old functional mechanism $x_i = f_i(p_{a_i}, u_i)$ and placing it under the influence of a new mechanism that sets the value x_i while leaving all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by $do(X_i = x_i)$ or $do(x_i)$ for short, amounts to removing the equation $x_i = f_i(p_{a_i}, u_i)$ from the model and substituting $X_i = x_i$ in the remaining equations. (p. 70)

Woodward endorses **PI** to some extent, but identifies two problems of an account of intervention as given by **PI**: one concerns the aforementioned semantic (or interpretive) aspect, the other an epistemological aspect. The conceptual problem, as I call the former aspect, consists in the fact that Pearl takes the notion of a causal mechanism (as represented by an arrow or a functional relationship) as primitive and defines the notion of an intervention in terms of this primitive, thus losing any possibility of using the notion of an intervention to characterize the notion of a causal mechanism or relationship.' (Woodward (2003a), p. 330) The epistemological problem derives from the fact that an intervention is allegedly defined relative to a graph whose structure is already known to be correct, and therefore such an intervention cannot serve to construct the graph in the first place.

I will discuss **EI**, which is relevant to Woodward's necessary and sufficient condition of what it means to say that one variable causes another (Woodward (2003b), page 45), alongside his criticism of Pearl's definition. I will also consider the epistemological next to the conceptual problem, since the qualitative question of when one is entitled to draw an arrow from one variable to another has an importance for the conceptual question as well.

Woodward's assessment of Pearl's account is not entirely correct in both regards. Regarding the epistemological problem, he correctly observes that in the examples of the relevant chapter 3 in Pearl (2000) the correct causal graph is already given, and the calculations refer to quantifying the amount of causal influence of one variable on another. Probably, this is because Pearl assumes that some pre-interventional data about the joint distribution of variables in a considered system are always available, and in that case his criterion of 'stability' (Pearl (2000), p. 24) require that an arrow (or at least an undirected edge) is drawn between two variables. Having said that, it seems that nothing in the mathematical formalism of Pearl's *do-calculus* stops us from drawing an arrow if an intervention really changes the target value distribution – or even if an updated calculation based on observational data brings that to light. But this would be merely tantamount to correcting an error in the causal inference that constructed the graph in the first place according to the criterion of stability. This is also indicated by Pearl's reply (p. 343 in Pearl (2003)).

The second element of Woodward's scepticism – we are still dealing with the epistemological problem – seems to derive from Pearl's criterion positing that an intervention leaves 'all other mechanisms unperturbed'. Woodward's concern is again with the mechanism connecting the putative cause with the putative effect. An experimenter might not know whether

there is a mechanism between two observables, so that she uses the intervention to discover a possible connection. At that point, an unfortunate misunderstanding kicks in. In Pearl's view, leaving mechanisms unperturbed means leaving variables unperturbed, since these are equivalent to mechanisms. One equation in a structural model, interpreted as a mechanism, determines a variable, so in Pearl's reading there is no existence independently of each other. Woodward obviously has another understanding of mechanism. A mechanism can be a *part* of one equation that determines a variable (see also the analysis in Hausman and Woodward (1999)) such that perturbing a mechanism could mathematically consist in the change of value of one parameter in a term of an equation. This might truly be a possible confounder that could mislead an experimenter into thinking an intervention on a variable has changed its putative effect as well, while in reality the intervention has changed the parameter of a mechanism connecting one of the effect's actual causes. But Woodward's example does not corroborate that this is his point, or at least his would be a badly chosen example. Woodward considers a paradigmatic example of an inappropriate atomic intervention, administering a drug in order to test its effect on recovery of some patients, *without* administering a placebo as well to a control group. The care the patients perceive to be receiving by being treated by the doctor introduces an additional mechanism from the act of administering the drug. It is a causal influence not passing through the cause-variable to be tested – the chemical effect of the drug. Therefore, it breaches condition (ii) in **EI**. Of course, there is no practical ideal intervention in this case, which would directly realize the consumption of the drug by the patient, which is why both **EI** and **PI** are not directly applicable. Practically, that is why the placebo is administered, such that the information the ideal intervention would have delivered can be approximated. The problematic point about Woodward's criticism is that Pearl has a means of dealing with this kind of structure as well, his 'backdoor criterion' (Pearl (2000), p. 79), which includes Woodward's criterion (ii). Crucially, Pearl's criterion does not presuppose any more structural prior knowledge about the causal graph than Woodward's list does, and thus Woodward has not raised any issue against Pearl's account.

This brings us to the second type of concern, the conceptual question whether we can *define* causation in terms of interventions. It is evident that Woodward thinks his treatment regarding this question is more adequate than Pearl's. Pearl uses his notion of intervention interpreted as *replacing mechanisms governing a variable* as primitive. Woodward is worried that something is lost by doing that. The question, however, is: What do we gain by introducing an intervention variable *I* that is explicitly 'causally' related to the cause-variable in the original system, with respect to the interpretive question? Does Woodward have in mind a *recursive* definition of causation? But there is no supplementary element in his definition of intervention that stops the recursive regress at some point, as it is the case with other recursive definitions. For example, natural numbers can be recursively defined (see, for example, Jeffrey (1967)), by means of the successor function, but they require the primitive of 'zero' as an anchor point. Woodward does not provide any hint in that regard, and so his motivation for insisting that *I* must be connected *causally* to the system variable rather than treating intervention as a primitive remains unclear.

I share Woodward's basic concern, though. It seems to be about the question how the whole process of maintaining and updating causal knowledge was bootstrapped initially. The suspicion is that knowledge of a causal structure would not have been there if interventions of some sort had not been performed before, which, in turn, cannot depend on the knowledge of the causal structure intervened on. But *some* prior causal knowledge seems to be required for interventions. Woodward's conviction that interventions do somehow bootstrap causal knowledge appears time

and again in his work. For example, the non-interference with a mechanism connecting the putative cause and its effect is *not* part of Woodward's definition **EI**, since he wants to have the knowledge of a connection *posterior* to the intervention. There has been a debate between Woodward and Cartwright about this point (Woodward (2008), Cartwright (2002)) and I would be inclined to espouse Woodward's view rather than Cartwright's. But the idea of a proper intervention can be put into question nevertheless.

There seems to be neither a real practical nor a real theoretical purpose that his definition of **EI** could serve – at least not over and above **PI**. In the clinical trials example, **EI** identifies administering a drug as an inappropriate intervention, since it introduces a confounder. But it does so only given we know the problematic structure beforehand. Also, in Woodward (2003b) and other works by Woodward there are no recommendations that allow circumventing such problems mathematically. On the other hand, **EI** is also insufficient to ground the meaning of causation, since it is a circular definition that does not allow defining causation recursively. But then both the theoretical and the practical use of this notion are in doubt, which leaves little room for any other sense of use.

I also want to challenge the theoretical value of the definition specifically from the conceptual perspective and raise a concern that has, to my knowledge, not yet been highlighted. Apart from the additional constraints on the intervention according to **EI**, there is the simple conceptual question whether something like an atomic intervention (Pearl's wording, but Woodward's notion essentially copies that aspect) is possible at all, at least in the way Woodward defines it, i.e. by representing it as an additional *variable*. Woodward says that this variable is added to the model, and can be thought of as a kind of 'switch' (Hitchcock and Woodward (2003)). It has the values 'on' and 'off', and if we put it on, the cause-variable will be 'on', or will increase its positive value, or the event it stands for will become more likely to happen. The intervention variable does that in virtue of being a new, direct cause of the system's putative cause-variable in the considered system. What is the *semantic* use of such a variable? How does it assume its values 'on' or 'off'? Supposedly by an arbitrary choice of the experimenter or by being directly tied to a random generator that outputs a value of a binary randomization function (for example, see Woodward (2003b)). But why not put the putative cause-variable 'on' or 'off' directly, if the point is the semantics of the drawing of an arrow between the putative cause and effect in the original system? At this point, keep in mind again Woodward's explicitly expressed concern with *semantic* issues in contrast to any epistemological questions. In terms of semantics, there is no insight to be gained from the explicitly *causal* connection between *I* and *A*. For all (semantic) means and purposes, the way *I* and *A* are connected is identical to how *A* and its effect *B* are connected.

In every example Woodward uses throughout the works cited in this section, *I* is a direct cause. But this is of no help, since direct causes are, according to Woodward's definition, model-relative. Due to his non-commitment regarding ontological questions, there is no hint in Woodward's work that he considers a bottom level to exist where a direct causal connection cannot be further fine-grained by allowing the consideration of additional intermediate causes. But other senses in which the causal connection between *I* and *A* could have a somehow distinguished status compared to other kinds of causal connections are not identifiable in Woodward's work, and therefore again the question of the semantic use of *I* arises.

In order to distinguish correlations from causation, we want to have the cause reflect the assumption of an arbitrary value that differs from the value it has had before or would have without a manipulation. But then the intervention should not be represented by a variable *on a par* with the

other variables of the considered model, which implies the same level of observability as can be ascribed to the other variables. Once that is done, the only possible connection between the intervention variable and the target variable is a *causal* connection, whose semantic grounding the intervention was supposed to deliver. In contrast with objective interventions, agency might be of very limited use in concrete causal modelling. But it might also fulfil exactly the role that Woodward wants his interventions to play in his semantic, or interpretive, context: grounding the meaning of a causal judgment at an atomic level, as an unanalysable, non-causal connection between an agent's free decision and an inter-subjectively observable action. Without resorting to the first-person perspective, as the agency-approach does (see section 5.19), it is difficult to make sense of the intervention variable. In particular, in which way can the intervention variable have no causal predecessor? Here another conceptual means that Woodward often makes use of comes into play: a random generator. Either the result of a randomization is used to fix the value of the intervention variable causally, or the intervention variable *is* the result of this randomization.²² In either case, the result of the randomization is an inter-subjectively observable event *distinct* from the target of the intervention *A*, in which case the connection between *I* and *A* can again only be causal.

Woodward asserts that his account is not viciously circular because the explanation of when there is a causal connection between *A* and *B* does not involve mentioning *that* causal connection (e.g., Woodward (2009), p. 254). But it should be clear that the kind of basic semantic grounding that agency proposes cannot be achieved by Woodward's variant of Objective Interventionism.²³ Therefore, in terms of the conceptual problem, Woodward's account adds little if anything to Pearl's account, which he criticises because it presupposes that the experimenter knows the causal structure already.²⁴ Woodward's intervention presupposes a lot of causal structure too, apart from knowledge about the specific connection between *A* and *B*. But this is a far cry from explaining causation as such on the basis of interventions.

4.2.5.2 Judea Pearl

From the conceptual standpoint, I have two objections to raise against the theorists of causal Bayesian networks like Judea Pearl, Peter Spirtes, Clark Glymour, and Peter Scheines. First, they content themselves with mathematical notions, when it comes to defining causation via interventions. Regarding this matter I concur with Woodward that such an account is conceptually incomplete. According to Pearl, causality has been defined according to probabilities and causal graphs, or alternatively, on the basis of structural equations. These are interpreted as mechanisms. It seems that this interpretation is supposed to take into account that causation, or at least its relata, are part of the observable, physical world – otherwise there would be no difference between causation and causal inference. Still, even with the interpretation of equations as mechanisms there remains a lacuna in the account, which relates to the question of what exactly happens when we set

²² It is worth noting that, for Woodward, variables *are* the relata of causation. Therefore, there is no further need to distinguish the cases of a relatum, like an event, mapped to a variable, and a relatum being that variable.

²³ Price (2007) speaks of interventions, objectively understood, as 'Trojan horses' for perspectivalism. I am very sympathetic to that interpretation. Price extends perspectivalism to the problem of grounding causal asymmetry in its temporal alignment, while I confine myself to the problem of the role of subjectivity in evaluating agent probabilities.

²⁴ For the sake of fairness, I need to say that in many other aspects Woodward's theory of causation does add a lot to Pearl's theory.

a variable to an arbitrary value. This operation can be defined relative to a mathematical model, but what is its equivalent in the physical world? The *set*-operation seems to contradict what we consider to be the ‘principle of causality’, namely that every event has a causal predecessor (different formulations, though all capturing the same idea, can be found in Norton (2003), Russell (1913), Eagle (2007), Hitchcock (2007)). My interpretation of the agency-approach will make use of computational models that could be realised by causal Bayesian networks, which makes my account look similar to Objective Interventionism in this regard. But these models are interpreted as the private property of the judging agent. The intervention of the agent, including its motivation for intervening, needs to be described according to *another* model. In this way, setting a variable to an arbitrary value can be interpreted while preserving the principle of causality (see chapter 7, where the idea is cashed out in terms of changes of levels of abstraction).

A second objection against this variant of Objective Interventionism is connected to the upshot of the analysis of section 4.2. It is their incorporation of the causal Markov condition as an a priori constraint of causation.²⁵ This entails that they have to reject correct causal models like the causal structure of Nancy Cartwright’s probabilistic factory example. I have discussed it in all mathematical details in section 4.2, so here I will just add a more commonsensical aspect. Qualitatively, the causal structure of the example has a V-shape with two effects and their common cause. The cause (the production process of a factory) brings about the occurrence of product and by-product. According to a pairwise evaluation of the process, (a) the production process causes the product, and (b) the *same* production process causes the by-product. Visualisation of (a) and (b) individually by means of causal graphs yields:

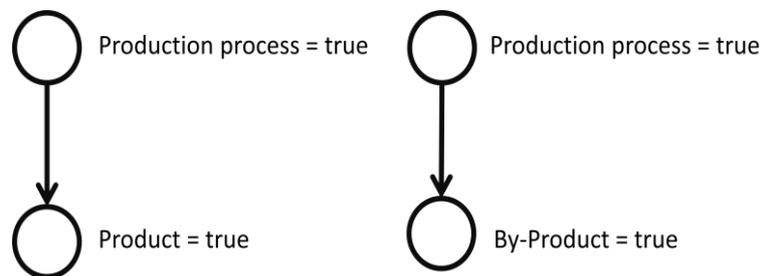


Diagram 4.1

Merging the two diagrams into one yields:

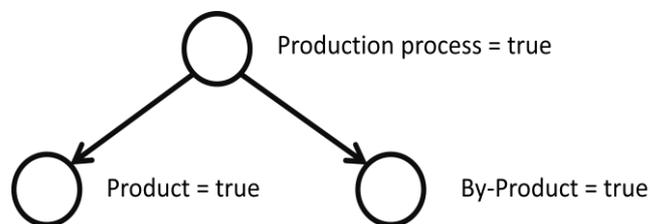


Diagram 4.2

This merging is licensed by the identity of the two cause-variables of diagram 4.1. But the corresponding causal graph is to be rejected according to Objective Interventionism, since every

²⁵ In accordance with that interpretation, Pearl (2000) writes on page 44 that the ‘Markov assumption is more a convention than an assumption, for it merely defines the granularity of the models we wish to consider as candidates before we begin the search.’

causal graph must comply with the causal Markov condition. As 4.2.4.1 has shown, this is not the case with the specific mathematical configuration of this example, so the causal graph is not a graph of a causal Bayesian network, and therefore not a valid graphical model.

There are reasons for assuming the causal Markov condition to hold. Among others, without it the algorithms of causal inference will not be able to deduce the causal structure for probabilistic configurations like the factory example. But the conflict between intuitively judged correctness of diagram 4.2 and the causal Markov condition can be interpreted as evidence that the causal Markov condition is not *directly* required by a concept of causation derived from manipulability, but has to be seen as an additional constraint to be added to the constraints directly following from manipulability.

4.3 Problems and objections against the agency-theory

4.3.1 Circularity

The agency-approach to causation shares the challenge of conceptual circularity with the interventionist approaches. James Woodward admits that his own theory is circular, but not that it is *viciously* circular (see previous section). Since Woodward, on the other hand, raises the objection of circularity, unanimously with some other writers, against the agency-theory of causation, it seems that this theory, in his eyes, embeds that vicious kind of circularity.

There are several accounts of agency on offer. In line with the general tendency of conceptual minimalism of my approach I will discuss the alleged problems of agency by looking at the account by Peter Menzies and Huw Price, who propose a concrete conceptual equation between causation and agency. According to Menzies and Price (1993), an 'event *A* is a cause of a distinct event *B* just in case bringing about the occurrence of *A* would be an effective means by which a free agent could bring about the occurrence of *B*'. The account can be read as a reductive analysis of causation on a metaphysical level, while I will, again, interpret the above characterisation of causation, more conservatively, on a conceptual level only, i.e. in terms of analysing a causal judgment.

Applying Price and Menzies' 'agency-formula', as I will refer to the above characterisation henceforth, the judgment '*A* causes a distinct event *B*' can be equated with the judgment 'For me, a free agent, *A* would be an effective means to bring about the occurrence of *B*'. Circularity allegedly enters the agency-account, since at least the concept of 'bringing about', which occurs twice, seems to require the causal concept, our explanandum. Moreover, assuming 'bringing about' is indeed a causal concept, a second problem occurs: the problem of *regress*. Consider the first occurrence of 'bringing about' in the agency-formula: 'the free agent brings about *A*'. Now, suppose that relation is causal and can be broken down into its constitutive elements by the agency-formula. Let us denote the second occurrence of 'free agent' that thereby arises by *C*. Then a two-fold application of the agency-formula to the causal claim '*A* causes *B*' would yield that this claim meant something like 'a free agent, bringing about an occurrence *C*, such that *A* effectively follows from *C*, would thereby render the occurrence of *A* an effective means to bring about *B*'. Iteration of this procedure of substitution leads to a never-ending regress of conceptual analysis. Thus, construing the operation of bringing about by the free agent in causal terms raises a conceptual circularity *and* a conceptual regress problem.

4.3.2 Anthropomorphism

Anthropomorphism is the second usual objection raised against agency. The content of this objection is the *prima facie* limited scope of (human) agency, and therefore the agency-theory's difficulty to accommodate instances of causation that lie outside the reach of any agent. In particular, it seems problematic to explain causal intuitions concerning the distant past. Hausman (1997) distinguishes the objection of limited scope explicitly from anthropomorphism in a second sense, viz. the anthropocentrically interpreted *asymmetry* of causation.

4.3.3 Two additional concerns

One caveat against picking difference-making as the starting point for developing an account of causation needs to be spelled out explicitly. Sometimes this caveat is formulated by stating that difference-making, which includes agency among a wider class of approaches, 'conflates epistemology with the metaphysics of causation' (Menzies and Price (1993), page 188). But this is

just a particular variant of a wider concern, namely that the difference-making aspect of causation does not capture the essential nature of causation, but merely a corollary of causation, or one of its emergent features. Given the metaphysics of causation, causes happen to make a difference to their effects. That does not mean that causation is essentially difference-making. I take it that this objection impacts not only the metaphysical or ontological problem, but the conceptual problem of causation, too, although to a somewhat smaller degree. To illustrate the point, one may adapt an example from Kant (1787):

‘A straight line between two points is the shortest,’ is a synthetic proposition. For my conception of straight contains no notion of quantity, but is merely qualitative.’

It is not spelled out in the section containing the quoted sentences what Kant means by his qualitative criterion of what it means to be a straight line, but one can assume that it turns on a notion similar to constancy of direction of a line we are drawing between its two endpoints. Thus, what corresponds to predicating a concept synthetically or analytically in Kant’s original intention of his example concerns an analogous distinction to be made in the present context: we can predicate a concept by one of its corollaries, or analyse what is contained in the concept. Some philosophers doubt that any meaningful difference exists between the two operations (see Quine (1951)). But one should at least keep that caveat in mind, since its denial is at least counterintuitive. The property of tracing the shortest distance between two points might be sufficient and necessary to identify a straight line, and as good a criterion as saying ‘the straight line is the one connecting two points while not changing its course’. Still, the latter criterion seems to follow more immediately from the concept of ‘straight’ than the former, and therefore it seems to more correctly capture what is seen as its essential property. Notice that the property of tracing the shortest distance between two points is not just a good indicator for having a straight line, like an animal’s capacity to fly being a good indicator for being concerned with a bird (despite the existence of penguins and bats). In contrast, the criterion is *without exception* sufficient and necessary for belonging to the class of straight lines (at least within Kant’s geometrical context, the Euclidian geometry). That is why the strategy of looking for specific counterexamples will fail in that case, given that a priori this method is a good strategy for demonstrating the accidental nature of a non-essential property.

Clearly, the example concerns the *conceptual* problem of a straight line, and one cannot eschew the problem by delegating it to a metaphysical level to be dealt with at another time. Therefore, if one takes the distinction to be problematic in this geometrical case, the question arises whether there is a *similar* relation of causation’s property of difference-making with another, more directly given property of causation, from which the difference-making logically follows. Concerns like these, although with a more metaphysical or ontological than conceptual flavour, have been expressed in the context of causation by Cartwright (2001), Psillos (2002) on page 103, or Mumford (2009), who writes that standard approaches to causation like those based on counterfactual dependency or probability-raising capture what is ‘symptomatic of causation, not constitutive of causation’.

Since my approach explicitly allows an operational definition of causation, the problem seems to be less severe in the conceptual than in the ontological context. As far as causation understood at type level is concerned, there does not seem to be a serious issue for difference-making approaches. However, explicating what is meant by ‘actual cause’ will raise problems that seriously challenge the approach. I will outline my own solution to that problem in section 7.1.1.

The second caveat concerns the *subjectivity* of a causal judgment based on agency. If I can manipulate the state of a light bulb via manipulating the state of the switch, then there is hardly a question, from my own viewpoint, whether a causal connection exists between the two, whichever physical realisation the connection might have. But the agency-theory makes the bold claim that all causal judgments have this shape. How can the subjectivity of the personal judgment be squared with the much wider scope that such a judgment is assigned according to the agency-theory? There is a sense of identity of the causal agent, for example a human agent who actually performs the manipulation, and the epistemic agent, who formulates a causal judgment based on her own experiences of manipulation, rather than being just an observer. But causal judgements often have inter-subjective relevance and they seem to require objective criteria, not just underlying subjective impressions. We are thus caught in a dilemma. Either we have to opt for an extensionalist ontology, which accepts only objects that are reconcilable with observational facts. Then we have to make sense of the difference which the identity of causal and epistemic agent makes for the causal judgments, on the basis of the non-subjective terms of the extensional ontology. Or we can embrace the subjectivity, and accept the identity as a ground of unity of the experiencing subject and its agency. But then we have to make sense of the seeming objectivity of causation.²⁶ This thesis takes the second horn of this dilemma.

I take it that, in the light of these two additional caveats, there will probably be residual doubt whether causation can really be explained in terms of agency, no matter how successful the analysis will be from this point. I will in the following section analyse causation in terms of agency, then provide an account of how I think the concept of agency arose from our natural history, which makes it look as if agency is in turn, via its secondary intension, reduced to regularities. This might be the case, but I am not sure whether this reduction still covers a dedicatedly *conceptual*, rather than ontological, reading of the problem. Also, regularities and causation are different concepts. If a concept of agency can be shown to be integrated into a picture of nature comprising only regularities, this would be a good sign, as long as the agency-notion of *starting a new causal history* is preserved. I will argue that it is preserved (section 6.2.8), and that this is exactly why causation needs to rely on agency. Therefore, the direction of dependence is conceptually fixed according to my hypothesis. Still, someone might argue that the concept of causation comprises something different. An appropriate analogy might be the perseverance of the notion of gravitational mass, which persists as an independent concept even after we have learned that the theory of relativity explains its identity with inertial mass (not just identity of their *values*, which has been experimentally shown by Galilei already, but of their *concepts*). If agency and causation are an analogous case, then in the end it is up to everyone's posterior intuition whether causation and agency are different concepts or not.

4.3.3.1 Recapitulation of the state of the argument and outlook on further sections

Interpreting the equation 'causation = agency' on a dedicated conceptual level that gives meaning to a causal judgment seems to be a more cautious approach compared to the alternative of tackling ontological questions in the course of the analysis. But there are further downsides. The agency-

²⁶ An example of an approach that tackles the first horn of the dilemma is Quine (1969). An example of tackling the second horn is Williamson (2005). But Williamson's 'objective Bayesianism' has a different focus than mine. While I see the central challenge in making sense of the identity of agency across time, Williamson focuses on how objective values of degrees of belief can arise, given we start from a personal set of beliefs of causal agents, which implies a subjective degree of belief interpretation of probability.

account seems implausible to many philosophers and is therefore rather unpopular. It is also rather surprising for the layman. Thus the equation does not seem to report the simple fact that causal agents explain their judgment based on the concept of agency. A normative reading of the account seems to circumvent that problem, but prompts the question what the normative force would consist in. If agents make use of intuitions concerning agency while forming judgments, even if causal judgments involving natural forces rather than human agents are concerned, then the account would have some normative force. Empirical evidence might be sought in order to corroborate such a claim, but this thesis goes another way and draws heavily on the argument (expounded in section 5.1) that agency has the virtue of being a non-circular theory of manipulation. It takes the advantages of the manipulationist approach without suffering from the obvious circularity of Objective Interventionism. The section that immediately follows that section is an expedition into the thermodynamics of causation, which will serve two purposes. It gives an account of how to objectify agency in order to get a concept of causation that allows the construction of more informative models of causal relations than what the agency-account in its bare sense allows, i.e. qualitative, subjective judgments concerning a binary relation between two variables. But the section also serves the purpose of giving further support to the idea that our causal judgments do in fact involve intuitions concerning agency, by looking at the conditions under which we judge agency *as observers*. Simple organisms, *living* organisms, act on a very basic level already. These considerations allow retracing how the acquisition of the concept of causation might have taken place. If an account of non-circular concept acquisition were successful, that would be a further piece of evidence that agency intuitions underlie our causal intuitions – if further plausible arguments can be made that concept acquisition based on agency feeds into concept application based on agency (see Ahmed (2007)). This part of the argument is speculative, though, and the connection between concept acquisition and concept application is only indicative, not compelling. Even with the prospect of phylogenetic concept acquisition the question of conceptual priority cannot be clearly settled, and it seems to touch intricate, transcendental questions time and again. But even independently of the question of concept acquisition we can synthesise a concept of natural action in such a way that it serves the purpose of tackling the second horn of the aforementioned dilemma. Natural actions can be embedded into an objective view of the world, and the identity of causal and epistemic agent will be replaced, from the perspective of an external observer, by the identity of information processor and actor. This will give us the constraints we need to go beyond the ‘agency-account taken at face-value’. The basic idea I am following is the same by which Huw Price is guided in Price (2007), although the structure of his argument is a bit different from mine.

5 The Problem of Circularity

In section 4.2.5 we have seen the problems of Objective Interventionism concerning the project of solving the conceptual problem of causation. Circularity cannot be circumvented, because ‘intervention’ is a causal concept, and regress cannot be circumvented, because the atomic intervention cannot be atomic. This raises the question whether the notion of agency fares better than the notion of intervention, given that the common basis of both approaches is the idea that a manipulation helps in distinguishing correlation and causation via different evaluations of conditional probabilities in each case. The starting point of the new interpretation of agency that I am defending is based on Menzies and Price’s formulation of the agency-theory (Price and Menzies (1993)), according to which ‘an event *A* causes a distinct event *B*’ can be equated with the judgment ‘For me, a free agent, *A* would be an effective means to bring about the occurrence of *B*’. If we apply this ‘agency-formula’ to a causal judgment we can explain what we actually mean by the judgment. This is one way of making sense of the conceptual problem of causation. The question of circularity then clearly hinges on whether the notion of ‘bringing about’ is causal or non-causal. Apart from this question the concern has been raised whether ‘free agent’ and ‘means (to an end)’ could be causal notions (see Ahmed (2007)).

5.1 Circularity of ‘bringing about’

This section will concentrate on the two occurrences of ‘bringing about’, and I will show that applying two different interpretations to each of the occurrences yields a non-circular analysis of causation. These interpretations entail a conceptual structure of causation that is analogous to Fred Dretske’s semantic information theory with its conceptual analysis of ‘becoming informed’ (Dretske (1981)). According to this theory, the process of becoming informed consists of two parts that are to be clearly distinguished: the transfer of information via a physical channel, and the subsequent formation of a semantic structure, called ‘digitalisation’. Leaving out any one of the two parts renders the concept of becoming informed incomplete (see section 5.2). My interpretation of the agency-formula was inspired by this bi-partitioning of something (causation) that *prima facie* seems monolithic. Both accounts, the one covering causation and the one covering becoming informed, are exactly analogous to each other in this respect. Building on the conceptual foundation to be developed in this section, it can be shown that informational and causal views on events can be integrated into a combined account, according to which information channels and causal mechanisms are identical. Digitalisation and direct action are parts of *causation by information* (section 6.1), a concept that further binds becoming informed and causation to each other. Causation by information is relevant because it helps us make sense of the seeming contradiction of construing some events as actions, while at the same time allowing them to be effects of causes. Causation by information will also be seen to be an important step toward a natural account of agency in section 6.2 and its sub-sections. Both of these aspects will be crucial for my overall argument.

In line with this outlook, this section of the thesis is structured as follows. The first sub-section will be concerned with Menzies and Price’s agency-theory of causation, and I defend my interpretation of this theory against the objections of conceptual circularity and conceptual regress. I will offer a new interpretation of the notion of ‘bringing about’, and will then illustrate the applicability of this interpretation by an example situation. The subsequent sub-section will be an interpretation of Dretske’s account of the flow of information and the subsequent formation of a semantic structure by digitalisation (both taken together henceforth simply referred to as ‘becoming

informed'). The interpretation is selective and also simplifying with respect to the overall theory of knowledge and the flow of information given in Dretske (1981). The intention behind this simplification is again to highlight the conceptual aspects of the problem to explain becoming informed, while Dretske's original aim is the development of a full-scale ontological account of information. What the conceptual problem exactly means in the context of becoming informed will be covered in the corresponding section. I will illustrate the applicability of this interpretation by picking up the example situation of the sub-section dealing with agency, and this time applying the informational rather than the causal perspective.

Sections 5.1 and 5.2 can be seen as self-sufficient conceptual analyses of their corresponding concepts, with special focus on avoiding the problem of conceptual circularity. But they also serve as a foundation for clarifying, in section 6.1 and the subsequent sections, a wider range of problems that involve the relation between information and causation. First and foremost, we require an answer to the question concerning the source of the analogous structure that has arisen from the preceding sections. Information and causation will be shown to be systematically tied to each other by the notions of causation by information and informed action. Information flow and causal flow are seen as identical, as are information channels and mechanisms. Which of the views is assumed when describing an agent's interaction with the observables of a channel depends on purpose and context. The section will also expound how it is possible – and why it is not a contradiction – that an agent can understand its own action *as* an action, although the action is judged to be dependent on, and even caused by, information.

In interpreting the agency-formula in the conceptual sense I probably depart from Menzies and Price's intentions. They do not distinguish the conceptual and the metaphysical level of the problem. For example, on page 188 of Menzies and Price (1993), they say an objection needs to be addressed that 'Agency accounts confuse the epistemology of causation with its metaphysics'. Later, picking up that very objection, they ask rhetorically: 'Doesn't the agency-approach to causation confuse the epistemology of causation with its conceptual analysis?' (p. 192) It seems they distinguish only the epistemological from the metaphysical-and-conceptual level, conflating the latter two. In contrast, I assign the conceptual level of the problem a dedicated treatment.²⁷

The suggested interpretation of the *first* occurrence of 'bringing about' (the free agent's bringing about A) is to see it as an *immediate and free action*, relating an agent's free decision with an observable event without any intermediate observables in the model determined by the context of the utterance of the judgment. The immediacy of the action circumvents the circularity and therefore also the regress problem. I will first offer a logical proof of this claim and then illustrate it by means of an example. Suppose that bringing about *were* a causal notion. Then 'the free agent brings about A' can first be turned into 'the agent's free decision brings about A', by virtue of substituting the first of the two relata by 'the agent's free decision', according to my suggested interpretation of the relatum related to A. Substitution of 'brings about' by 'causes', by virtue of the assumption that the relation is causal, yields 'the agent's decision causes A'. But applying the agency-formula to analyse this putative causal relation would amount to the proposition 'bringing

²⁷ Since the meaning of 'metaphysics of causation' depends on how one understands the term 'metaphysics', it might be argued that Menzies and Price really mean 'conceptual analysis' by the term 'metaphysics'. In that case no issue needs to be raised. In order to make sure that the explanation of what underlies a causal judgment is properly distinguished from the question of the reality of causation independently of its observers, I contrasted the conceptual problem from the epistemological and the '*ontological*' problem. That being said, whenever I use the term 'metaphysics of causation' I mean the 'ontology of causation', not its concept.

about the occurrence of the agent's decision would be an effective means by which a free agent could bring about the occurrence of *A*'. Finally, we substitute by 'the agent's free decision' in the same way as before, which results in a proposition about the agent's free decision to bring about a decision to bring about *A*. But the decision to decide to bring about *A* is the decision to bring about *A* *already*. The attempt to analyse 'to bring about' in the way we analyse 'to cause', on the basis of the agency-formula, yields a repetition of vacuous relations between identical relata. Thus, a semantic interpretation renders the agency-formula inapplicable at this point, for the very reason that the relation of bringing about is not causal.

The series of substitutions and its result can be interpreted in two ways. Either the first instance of 'bringing about' is not a causal concept. Thus the agency formula is, from the point of view of agency-theory's own logic, not applicable. The circularity and regress objections would be rebutted. Or my suggested interpretation of the first relatum, construing it as a *decision* of the free agent rather than the agent itself, is incorrect. But that this relatum is something distinct from an observable, something which is true of the two *causally* connected relata (*A* and *B*), lies at the heart of the agency-theory. Unfortunately, the way in which the agency-theories are often dismissed in overviews over theories of causation leaves the question of the first relatum, on which the circularity objection hinges, unexplored. Objective Interventionism, with its 'intervention variable', which is causally connected to the variable representing the cause of the causal claim, definitely makes the wrong logical move at this point.

Besides resulting from a free decision, I suggest to interpret the means *A* as a *direct* action. Both points require further explanations. The agent's freedom from coercion will be explained in section 6.2.8, where it will be made plausible why the decision to act cannot be an observable of the model that the agent uses to judge the efficacy of its own intervention. Similarly, the agent's 'bringing about' is immediate relative to the agent's own model, while it can appear to be mediated from the point of view of an external observer.²⁸ An agent's model implies that the chain (or, more generally, the web) of observable causal events starts only *after* the immediate action has occurred. I understand the claim of the agency-formula, interpreted on the conceptual level, to imply that an agent can formulate a causal claim '*A* causes *B*' already, and therefore has *some* causal concept. But the agent, at this stage, does not necessarily need to have the ability to analyse his causal concept further (see again chapter 3 on this point). This is where the agency-approach suggests: 'When you say '*A* causes *B*', what you actually mean is 'I, a free agent, can use *A* as a means to bring about *B*''.

We grant the agent (possibly just the illusion of) immediate access to *A* according to its model, whereas the way *B* is connected to *A* is something the agent must be able to justify further. As indicated above, a minimal justification can be a belief in a dependence relation between the observables. This interpretation leads to the (only seemingly) paradoxical view that 'raising one's arm' can be given a causal and a non-causal interpretation. It is causal, since we can tell a plausible mechanistic story explaining how this event unfolds. This, however, is the perspective of an external observer. An external observer might even be able to pin down where and when a decision-event takes place in the brain of the acting agent. In that case not only the immediacy of the action, but also the causal history of a 'free' action would be gone. But from the perspective of the acting agent who also models the event as *her own* action, both the immediacy and the property of starting a new strand of causal history are implicitly given by the concept of action. The identification of event

²⁸ Crucially, these changes of perspective do not correspond to a simple zooming-in into a more fine-grained model, since not only the number, but also the quality of the relata of the model changes during such a transition (see section 7.1)

A as object of immediate action on the one hand and as part of the continuous chain of events according to the principle of causality on the other hand, leads to conflicting intuitions similar to a Kantian *antinomy*. In chapter 7 I will explain by means of different levels of abstraction, which correspond to the above different perspectives, how the conflict can be explained further.²⁹

To illustrate the freedom from coercion and the immediacy of action in a causal model, I will describe an example situation that involves agency. The example affords both a causal and an informational interpretation, which is why I will pick up the example again in the sub-section about information-flow and becoming informed. The example features Alice, who pays a visit to Bob. She wants him to open the door and hence looks for a way to inform him about her presence. A good way to do that would be to make the doorbell ring. But she cannot do that directly. What she can do directly is depress the button and thereby exploit her background knowledge about the causal connection between button and bell. Depressing the button is an adequate measure to make the bell ring, in order to inform Bob about her presence, such that eventually he will open the door. The strategy is very straightforward and commonsensical. But in order to plan an action explicitly one has to resort to some kind of model of the situation in question.³⁰

The given situation involves just a sequence of events (featuring no side effects, branching, etc.), which are – putatively causally – connected. At this stage it must be decided only on how fine-grained such a model has to be in order to represent the sequence of events adequately. In this case, an adequate partitioning might be the highlighting of some conspicuous events, such as: *D* (depressing the button) – *R* (ringing doorbell) – *V* (Bob's vibrating eardrums) – and, finally, *B* (Bob's reaction). Alice's goal is to cause Bob's reaction. Connected to devising a strategy to accomplish that task is the conceptual problem of how Alice could underpin the underlying causal judgement, given that she is prompted to utter her causal beliefs explicitly. It is then, according to the assumption of the agency-approach, possible to equate her explicit causal judgment 'Depressing the button causes the ringing of the bell', and, by virtue of transitivity, 'Depressing the button causes Bob's reaction', with the judgment: 'Alice believes that her bringing about the depression of the button is an effective means to make the bell ring'.

One could raise the question how exactly the agent intends to depress the button. Given the initial model of evaluation, the model according to which Alice plans her action explicitly, there was no need to answer that question in more detail: she does it simply by depressing the button. The explicit question of how to proceed at this point raises the need to create a different model. An adequate answer that follows the new context could be: 'You depress the button by touching it with your outstretched finger'. The model that Alice used to plan her action considered the button as an object under her direct influence. However, given a new context has been fixed by the question 'How can you depress the button?', and a reflection is forced upon the agent about what it actually is that she can directly influence, the hand has become the immediate object of intervention. The movement of the button has become subject to a mechanism that connects the hand and the button, a mechanism that she hopes will work for the sake of her final intention. The impenetrability

²⁹ See also Floridi (2010) for discussions of how the Kantian antinomies relate to the method of levels of abstraction.

³⁰ 'Planning an action' is, in this context, to be understood as *explicitly* planning an action, such that the plan of action can be subsequently justified and explained to others. This is to make sure that the agent possesses at least an intuitive concept of causation and can be prompted to utter his causal beliefs about the situation. Understood in this way, planning and explaining match each other, apart from the difference given by the prospective and retrospective stance, respectively.

of matter can be seen as establishing the required ‘mechanism’ between hand and button, and this will bring about the depressing of the button indirectly. In a normal context, there is, of course, no need to take into consideration a more proximal object than the object we can immediately reach by stretching out our arm. An analogous reasoning in the informational (rather than causal) context makes us consider the object causing the sound to be the one we are acoustically interacting with during an auditory perception, rather than with our auditory sense organs. The shift to a new model of evaluation, which is prompted by the question that seeks to go into more detail concerning the action, is, in the terms of section 3.1 and chapter 7, a change in the level of abstraction. It often helps, in order to prevent confusion, to give an answer at the same level of abstraction at which the question was asked, and likewise make a shift to another level of abstraction transparent during the course of a discussion. Chapter 7 will come back to this point once the levels of abstraction for the different kinds of judgment will have been made available.

The above considerations show that the assignment of the role of immediate object is subject to the required level of detail of the model. There is no need to consider a specific *kind* of object rigidly as the immediate object. Certainly, normally there is no X such that the proposition ‘ I bring about X , as a means to bring about the movement of my hand’ is true. We normally just move our hand and there is no intermediate event we can arbitrarily bring about *such that* it follows that our hand moves. Accordingly, it would neither make sense to plan our action on the basis of bringing about such an X , nor to explain our past action involving hand movements on the basis of an intermediate X . However, in a scientific context – under a change of the level of abstraction – there is certainly such an X to be tracked if we mean by X an observable (nervous, or biochemical) signal that precedes the contraction of our hand muscles. Even pragmatically such a change in the model might occur, e.g. when people have to learn to control artificial limbs by means of computer-brain-interfaces. On the other hand, given that the mechanism between agent and distal object is very tight and a task is repetitively performed, the agent might treat the distal as the immediate object, thereby simplifying the model. In our situation, it might appear to Alice that she can influence the bell directly, only to be frustrated by an eventual failing mechanism, which forces a change in her way of modelling this type of situation.

Keeping in mind this context-sensitivity, we can recapitulate that the solution to Alice’s causal problem ‘How can I cause Bob’s reaction?’ involves the selection of an adequate immediate object. Such an object is adequate if it can be considered directly accessible by the agent and if it is connected to the desired final state by a chain of correlated observables (D - R - V - B). Whereas, in the informational context, we are interested in backwards correlations (e.g. what observing R tells us about the prior D), in the causal context we are interested in forward correlations (e.g. what will happen to the bell, and finally, concerning Bob’s reaction, given that we now manipulate D). Obviously, the issue of context sensitivity also applies to the second part of the modelling decision, which concerns the numbers of observables to be taken into account. That is where V might drop out of the model as gratuitous.

The scheme that describes the creation of an agent’s causal model resembles a two-stage process that will be revisited in the subsequent section, covering the process of *becoming informed*. A causal agent that faces a situation with several variables must build a model at a level of abstraction that is appropriate to context and purpose. The agent can perform an intervention on one of her immediate objects, and she will select the intervention that is most likely to result in a beneficial outcome. Whereas complete control of the direct object is *assumed*, affecting the indirect (distal) causal object, or the event involving the indirect object, is dependent on the correct workings

of the mechanism that connects direct and indirect object. The details of the mechanism are subject to various empirical constraints, and the agent's knowledge about the functioning of the mechanism might reduce to just a belief in the dependence of *B* (indirect object) on *A* (direct object).

A similar thing happens in the semantic information theory (see next section), where we take apart the physical and the semantic aspect of becoming informed, in order to address both aspects separately from each other. At some point, describing the flow of information by the constraints of a physical channel has to stop, in order to shift to a semantic perspective. If this shift of perspective doesn't happen, a concern of regress similar to that afflicting the agency-account arises for the semantic information theory: tracking the flow of information of a perception further and further inside the agent's brain, we will eventually encounter the *observable reaction* of the agent (who is, in such a context, considered as an *object* of investigation), and no semanticisation of information ever seems to happen. This informational variant of the regress problem is discussed, for example, by Dennett (1991), page 49. Similarly, the solution to circumvent the circularity and regress objection against the agency-account of causation requires a shift of perspective from a mechanistic point of view, or at least a view relating observable regularities, towards a view that relates a free decision to act with an observable event resulting from the decision. This would be my suggestion for how to interpret the relation between the 'free agent' and the *A*.

The immediate action, which is identical to the 'means' (= '*A*') in the agency-formula, has the property of immediacy by virtue of the fact that, from the point of view (or *within the model*) of the agent who evaluates the causal claim, there is nothing in between the decision to act and the immediate object, or immediate event, *A*.³¹ To catch up the most obvious example of such an immediate action, consider again that (usually) we cannot bring about an *X*, such that *then* our arm rises. Instead, we bring about the rising of our arm, without even being able to further explain that ability.

Two additional explanations and qualifications might be helpful. First, it seems that the immediate action *A* should always be identified with voluntary bodily movement. However, in the story about Alice's desire to inform Bob about her presence at the door, the immediate object initially considered was the button, not the hand that operates the button. A change in the model by incorporating the hand was required only after asking a further question that was not tractable within the old model. Fixing the model at the right level of detail depends on context and purpose, which in the context of explaining causal judgements in turn depends on what counts as a satisfactory explanation. It seems that the level of satisfaction of an explanation depends both on constraints from the considered situation and on factors we can decide, just like the selection of what the immediate object is. The notion of a *constraining affordance* is of help here. The limbs of our body over which we have voluntary control make them a natural choice of immediate object, but in most contexts the direct control will be extended to at least the range of things we can reach with them. There is little sense in assigning a degree of probability of our successfully depressing a button, just like we would not assign a probability to the movement of our hand towards the button – except if special circumstances obtain.³² By contrast, what we expect to happen once we have depressed the button is outside of our control. We have done what we can and now have to rely on the chain of unfolding, merely observable, events to work in our favour.

³¹ As it is usual practice in the literature on causation, I abstract from the differences between objects and events if they are considered as causal relata.

³² For example, someone might want to prevent us from reaching the button.

The second qualification is just the reminder that, in my interpretation of the agency-formula, analysing the causal relation by splitting it into immediate action and correlated observables is meant as a solution only to the *conceptual* problem of causation. We can now explain what we mean by the judgment 'A causes B'. The metaphysical problem of causation ontologically understood, i.e. the question: 'what is the X, the Humean *secret connection*, between the A and the B, such that they seem to be necessarily correlated?' is left unanswered. The agency concept *requires* regularities rather than explains them. Likewise, epistemological problems, such as how to proceed in order to disclose the causal structure of an unknown system, are not addressed. To indicate the problem, consider that Woodward (2003b) explicitly forbids, in his definition of interventions, any side effects of intervening on A in a suspected 'A-causes-B-structure', in particular side effects influencing B. That is because, by probing the structure by means of the interventions, we want to know whether any causal influence reaches B *via* A, not through the intervention itself. In the agency-formula, we have a similar, but less precisely formulated idea, reflected by the fact that the agent assigns A as the object of immediate action, while B is her *indirect* object, to which a probabilistic value needs to be assigned. The agent might be right or wrong in the attribution of the roles of cause and effect to the considered variables, but the analysis in terms of agency, conceptually understood, is pitched at the right level no matter whether the agent is right or wrong, since it is the agent's *judgment* that we want to explain in non-causal terms. The agent that performs the action (even if only hypothetically) is *identical* to the agent who judges the causal connection.³³ Under general assumptions, we can assume that if an agent mistakenly believes she brings about B via its direct object A, while for all other observers she brings about B directly, then it is reasonable to assume she will find out eventually and change her model accordingly.³⁴ Still, it must be granted that the agency-approach alone is much too crude for deriving a methodology for causal analysis, given the epistemological problems that the Objective Interventionist approaches try to solve. Each of the two manipulationist approaches serves a different purpose, and one must choose the adequate approach depending on whether one is more interested in questions concerning the causal structure of specific systems, or interested in the analysis of the concept of causation and its acquisition as such.

³³ This is the crucial difference between the agency-approach and the objective interventionist approaches. The license to judge that the causal influence comes *from* A is due to the fact that A is *the intervening agent's* direct object.

³⁴ Interestingly, on page 193, Menzies and Price (1993) cite the fact that it is a virtue of agency-theories to account for the fact that an agent can err in his judgement. But this remark of theirs suggests a *conceptual* reading of their solution, rather than clarifying the 'metaphysics' of causation. (I suspect, however, that Price would not concur with my assessment)

5.2 A parallel analysis – Becoming informed

Dretske (1981) outlines how a message that contains a contingent matter of fact can be communicated from a sender to a recipient, such that it can subsequently underpin a belief state associated with the recipient of the message.³⁵ Contingent matters of fact are understood as matters of fact, which, for all we know, could have turned out differently, which is why we can only learn about them via a corresponding signal reaching us from where the fact occurs, but not by other means, e.g. by means of a deduction from information we already possess.³⁶ For an explanation of how a signal, or message, is transferred through time and space, Dretske makes use of the concept of a Shannon information channel (Shannon (1949)). In Dretske's reception of that concept, any physical structure can serve the purpose of being a substrate for an information channel as long as it features the following structural constraint: the catalogue of alternatives at the receiving side, of which one is realised after the transfer, must allow a faithful mapping to one of the alternatives at the sending end of the channel, which corresponds to the specific event that is to be communicated. The condition ensures that an observer can infer what happened at the source based on the received information at the receiving end of the channel. This condition will be called henceforth the *non-equivocation condition*.

Examples of designed channels are telephone lines: what the listener hears corresponds to what the speaker at the other end has said. The way in which animals leave traces (footprints, broken twigs, etc.) when they move about can count as a natural channel, and a tracker can extract the information about the whereabouts and conditions of the animals that caused the traces. With the concept of the channel the flow of information, the passing of the information from one relay station to next, is covered. The subsequent step that completes the process of becoming informed is the extraction of information. In the phone call considered before, the voice is first turned into an electric signal and is then passed through various different relay stations, until it is finally understood by a listener, rather than passed further along.

In Dretske's information-theoretic model, the semantic quality that information acquires by the second stage of the process (the 'understanding') is covered by the concept of 'digitalisation' of information.³⁷ Digital information and analogue information are defined relative to each other. Below are their definitions and a few examples to illustrate the point (quotations in this section are from Dretske (1981)).

Analogue and digital information: 'A signal (structure, event, state) carries the information that s is F in digital form if and only if the signal carries no additional information about s , no information that is not already nested in s 's being F . If the signal does carry additional information about s , information that is not nested in s 's being F , then the signal carries this information in analogue form.' (p. 137)

The definition makes use of nested information: 'The information that t is G is nested in s 's being $F = s$'s being F carries the information that t is G .' (p. 71)

³⁵ To avoid confusion, I use the contrastive pairings of 'sender' and 'recipient', when agents communicating via a channel are concerned, and 'source' and 'receiver', understood as correlated events that are connected by a channel.

³⁶ Dretske does not use the denotation 'contingent matters of fact' literally, but it is adequate for the kinds of fact that his theory describes; therefore, I am going to use this denotation henceforth.

³⁷ This is not to be confused with the ordinary meaning of that word, as the subsequent definitions show.

Dretske's examples of analogue and digital include a picture that contains, among other objects and persons, the depiction of a woman. Since there is more in the picture than just the woman, the picture represents the fact that she is part of it in *analogue* form. A *digital* form of this piece of information is the proposition: 'The picture contains the depiction of a woman'. This proposition abstracts from all the other details of the picture, and once it has replaced the picture as a container of the information, the rest of the information included in the picture is unrecoverable. The proposition does allow us to *infer* more from it, e.g. that a human being is in the picture, but this piece of information is only *nested* in the proposition (the former follows logically from the latter) and therefore does not undermine the fact that its form is digital. Another example of nestedness is the information that a geometrical figure is a square, which contains, as nested information, the fact that the figure is also a rectangle. Generally speaking, logically weaker pieces of information are nested in logically stronger ones.

Yet another way to understand the distinction is to consider digital and analogue gauges. Looking at a hand of an analogue gauge enables the conversion of the information into a representation by a digital number, and the closer you look at the hand, the more digits of precision the digital representation allows. But given we are left with only the digital number, the level of precision is fixed, and the analogue information is irreversibly lost. This *loss* of information during digitalisation contrasts with the *preservation* of information that we associate with the Shannon channel, and it is a decisive step to explain cognition. As Dretske puts it, 'a cognitive system is not one that renders a faithful reproduction of its input in its output. Quite the reverse. If a system is to display genuine cognitive properties, it must assign a sameness of output to differences of input. In this respect, a genuine cognitive system must represent a loss of information between its input and its output.' (p. 183)

We pick up the example situation of Alice visiting Bob once more, and the protagonists are this time sender and recipient of transmitted information. Analysing the example, I will focus on the conceptual problem of becoming informed, which I define as follows: how can an agent explain the judgment 'I have become informed of the contingent matter of fact F '? Answer: by stating 'I have perceived the signal r , which tells me that F has happened at the source of the physical channel that transmitted the information about F .' In other words, my becoming informed can be conceptually equated with the perception of a signal and the transmission of information through a physical channel. The motivation for this equation will be explained in the rest of this section.

As an illustration of the concepts, I will reuse the example from section 5.1 with Alice, the visitor, who pays a visit to Bob. Alice arrives at Bob's door. At his door she depresses the button of the doorbell. The bell rings, and Bob gets informed that someone stands in front of his door. Turning to the conceptual problem, we could now prompt Bob to account for his belief state. Two things seem to require an explanation: his belief concerning the perception of the doorbell and the belief *mediated* by the perception, which implies the presence of someone at the door. The proposed solution uses two relations. One relation is between a catalogue of observable events at the source and a catalogue of observable events at the receiver e.g. 'button is depressed/is not depressed', and 'bell ringing/not ringing'. The relation allows a one-to-one-mapping, and is therefore a function, from the receiver-events to the source-events. The second relation is between a catalogue of events at the receiver and a subsequent semantic structure, e.g. a proposition such as 'I hear the doorbell ringing' (stemming immediately from the perception), or 'Someone is at the door' (via inference from the previous proposition). Making use of this bipartite scheme of becoming informed in a specific situation implies the identification of a direct object (henceforth also called 'the object of

the digitalisation') and an indirect object (connected to the direct object via an information channel). The object of digitalisation is the received information signal, which is given in analogue form. On its grounds the semantic proposition is formed by digitalisation, e.g. from the ringing doorbell we can form the semantic proposition: 'I hear the doorbell ringing'.

Application of the scheme to the example of Alice and Bob yields that the bell is Bob's obvious choice as object of digitalisation, since it is his object of perception. Perception is an act that we do not further analyse under normal circumstances; we can therefore form a digital structure, like a belief or a proposition, on the basis of the object of perception without further ado. It is, of course, correct to say that the sonic waves travelling through the air from the bell to Bob's ear, and likewise the vibration of his eardrums, are part of the same physical information channel that the ringing bell belongs to. But, whereas – in ordinary contexts – it makes sense to say 'I heard the doorbell, therefore someone has depressed the button', it does not make much sense to say 'My eardrums are vibrating; therefore someone has depressed the button'. Accordingly, the explanation of our belief state does not have to go into further detail at this point – the object of perception is the agent's best choice as object of digitalisation, and these two kinds of object will match in a normal context of explanation. The further matters of fact are then *inferred* on the basis of the agent's knowledge about the information channel, and with that second step the explanation of the belief state 'Someone is at the door' is complete.

This scheme of explanation takes for granted both that our sense organs work the way they are supposed to, and that the button-bell mechanism works the way it is supposed to, viz. such that the non-equivocation condition of an information channel is satisfied. In a default context, the assumptions are reasonable and, correspondingly, the explanation as to why we feel informed should be satisfactory. It needs to be emphasized that a satisfactory answer depends on the type of question asked: of course it is a perfectly reasonable assumption that there is an information channel, not only between button and bell, but also between bell and eardrums, eardrums and *X* inside our skull, and it might be an interesting scientific question to determine *X*. But the relation of digitalisation is *not* to be described by a channel, i.e. coupled systems that transport relevant differences through time and space. Rather, it is a relation between an observable and a semantic structure. The selection of the object of digitalisation therefore marks the end of viewing the process of informing by means of a channel and the entrance into viewing it under non-observable, semantic categories. This change of views, however, corresponds to a change of context: at some point there is the expectation that a different conceptual scheme must be employed in order to render a satisfactory explanation; the physical scheme has to transition into a mental scheme, in order to explain agent-behaviour in an adequate way.

The way we applied the template of 'becoming informed' to the example is not meant to be rigid, but can vary according to context and purpose of the required explanation. Given that our sense organs work the way they do, it is a constraining affordance to choose the object of perception as the object of digitalisation, but in explaining a judgement it is still an agent's *decision* to bipartition the process in a way tailored to a specific situation and the combination of context and purpose. For example, it would indeed be an awkward application of the concept of 'hearing', if one applied it to the vibration of one's own eardrums, but it is conceivable that one could learn to refer to such kind of event in more detail, such that inference to external distal events becomes possible on the basis of such a referencing. But before it would make sense to apply the perceptual scheme in this way, one would expect a strong shift to an alternative context-purpose combination, compared to a more ordinary situation. The same holds true if the more distal event, the depressing of the

button, *D*, receives the status of the direct object of digitalisation. *R*, the doorbell's ringing, is much more likely to be identified as the direct object, and knowing *D* via *R* is a matter of causal inference. Given we are not deaf and our ears are not plugged, we have quasi-direct auditory access to a doorbell. If it rings, we will hear it. At least we do not have to take into account whether the air will transmit the sonic waves in a specific situation – it will do so in the usual circumstances. Door buttons are much more likely to fail to bring about the desired effect of *R*, and in this situation (at the latest) we learn that we do not in fact hear the button itself. Again, the identification of direct and indirect object is not a rigid scheme across different situations, but in normal contexts the choice naturally arises from the way our perception works. For someone it might make sense to think of a context-purpose combination where the mechanism connecting bell and button is so infallible that a simplification of the model does make sense. Then the two collapse into one object, and one hears the button, therefore someone is at the door.

6 Thermodynamics of acting and information processing

The preceding paragraph is an attempt to defend the claim that Peter Menzies' and Huw Price's version of agency-causation can be given a reading that does not lead to a vicious circularity or vicious regress on the conceptual level. However, this was merely a defence by logical means. We have assumed that all the concepts are already in place and have meanings that are readily exploitable for a logical discussion. The import of agency-causation will be increased significantly if it could be made plausible how the concept of agency can arise in nature. A conceptual account merely requires making explicit the grounds of a causal judgment, by analysing the causal relation into its elements of a direct action and a subsequent, correlated second event. But the normative strength of this account would be reinforced if it were at least plausible that a judgment of causation is genuinely grounded in a judgment of agency, rather than just allowing a truth-preserving substitution of an agency-statement for a causal statement. The question of concept acquisition is a not so convincing aspect in Price and Menzies (1993), relying on the idea of 'ostension'. I will argue in this section that a teleo-semantic account is more plausible than an account based on ostension. The teleo-semantic account that I am suggesting turns on both the necessity of acting, necessary for the very existence of the agent, and the sense of identity of actor and processor of information about the contingent environment.³⁸

Next to the question of concept acquisition and its import for the normative strength of the account, the agency-formula has at least two further questionable elements, besides the still to be dealt with anthropomorphism. It contains a reference to a *free* agent, a qualification that needs to be made further sense of. And the formula appears to be applicable to only a very confined range of examples, namely binary relations between two observables. But causal judgments seem to require the accommodation of much more complex judgments than binary relations.

Finally, there is more reason for dealing with Dretske's semantic information theory and the agency-theory of causation in a parallel fashion than merely exhibiting their curious similarities. In the following subsection, I will elaborate on the idea of mechanisms as information channels and vice versa, identifying one with the other. This has been foreshadowed already in the preceding section by the fact that we used the same example (the doorbell example) for both kinds of things. A logical next step is to look at the semantic aspects of both perspectives, viz. the informational perspective (looking at physical differences that allow inferring what has happened at an earlier time) and the causal perspective (considering differences that make some difference at a later time). If the semantically enabled agent that processes incoming information is also a causal agent that intervenes on the basis of the information, a new composite concept can be created, dubbed 'causation by information'. Again it is a desideratum to further justify this synthesized concept by looking at natural constraints on causation. It will be seen that actions are indeed informationally constrained like the new concept would have suggested. To summarize the points above, here are the four motivations to look at the natural origins and physical constraints of agency:

- We need an explanation of *causation by information* and its inherent *referentiality* to past events. This phenomenon seems to allow causally inert facts of the past to become causally efficacious and also seems to contradict Markovian ('memory-less') pictures of causation

³⁸ A teleo-semantic account similar to mine, although not directly grounded in thermodynamics, can be found in Dretske (1988); also see chapter 12 in Dretske (2000).

- We would like to have a more plausible account than ostension of the *concept acquisition of agency*. The idea of the section will be that if we can find the non-causal criteria for identifying an action then these could be the criteria to explain the acquisition of the concept of causation non-causally. This would corroborate the claim of the agency-approach that intuitions about actions actually underlie our intuitions concerning causation.
- We want to understand what the criterion is for being considered a *free agent*. Similar to the problem of causation by information, we have in the agency-formula a concept that seems difficult to square with Markovian models of physical reality, and this fact calls for a further explanation.
- We need further information about actions, in order to accommodate the agency-account to cases beyond the abstract, binary relation between two observables. I will try to achieve this by looking at the properties of natural actions.

I will discuss the four motivations in the following sections in the order given above, starting with *causation by information*.

The argument that I am trying to make at this point goes as follows: We have seen that agency can solve the asymmetry problem of causation; viz. from the first-person perspective the judgment of manipulation of a *B* via a direct action *A* is possible and well-grounded, since it does not require an objective truth-maker besides the subjectively felt control of *B* via *A*. In cases of causation by information we are able to judge whether an action has been taken by *another* agent. If the assumption is true that this judgment hinges on information processing, and in turn information processing is physically grounded, then by causation by information no discontinuity is introduced into a conceptual framework that assumes a causally inert past. What appears as a semanticisation is actually dependent on the physical processing of information that dispenses with any references to past events. The seeming lacuna that has been introduced by causation by information would have been closed. As for the question of concept acquisition, actions are also teleological; they tend to bring about a beneficial situation. If the actions in causation by information, whose constraints are examined in the present section, are of the same kind as the actions which we hypothetically perform while judging cases of efficient causation, then causes in general, by virtue of identity with natural actions, inherit the constraints of actions thus construed. The constraints that the following subsection will yield are properties typically ascribed to causation: locality, asymmetry, and regularity.³⁹ Before the identity of agency and causation can be established, the physics of information flow and information processing and their bearing on the notion of action will have to be examined.

³⁹ For a similar list, capturing the same idea of identifying typical properties of causation, see Menzies (1996), or Psillos (2002), page 6, 7.

6.1 Causation by information

Given the conceptual clarifications of the sections 5.1 and 5.2, we can now turn to the question of what happens if the information processor is also the acting agent, and the information received constrains or triggers its action. This would tell us something about the relation between informational and causal processes. The doorbell example considered in the previous sections shows that the same chain of events can be seen as established by a causal mechanism, or alternatively by an information channel. Other considerations of sections 5.1 and 5.2 suggested that the notions of mechanism and information channel both allow for different ways of modelling a chain of events, depending on the level of detail required by the explanation. Conspicuous events like the vibration of a doorbell can be considered, under an appropriate alternative level of abstraction, as just one station within a – potentially continuous – spatio-temporal causal process (or process of information flow). To pick them out as distinct events seems to be a constraining affordance, dependent on the required level of details. These considerations apply to both kinds of channels, which suggests that mechanisms and information channels are physically the same kinds of things. On the one hand, the availability of information about a contingent fact at another place requires a causal mechanism to convey the information. On the other hand, an effect occurring at some place often⁴⁰ allows abduction to the cause that brought the effect about via a causal mechanism, such that information about the occurrence of the cause is preserved.

Furthermore, both the informational and the causal view featured operations with unobservable relata, which, respectively, mark an 'exiting from' and 'entering into' the channel view on the events, i.e. the view that picks out correlations between observable relata. Having already identified mechanisms with information channels, it now makes sense to ask whether the informational and causal pictures can be even more tightly linked, by tying up these two semantic operations, digitalisation and immediate action, which, in the doorbell example, are at opposite ends of the chain of events. This linking leads one to the concept of *causation by information*, to be elaborated in the following paragraphs. Causation by information requires both the extraction of information from a signal and the subsequent acting on the basis of the information. Otherwise it was not the information contained *in* the signal, but rather other properties *of* the signal that were causally efficacious. Some examples will highlight the difference.

6.1.1 Intuitions concerning mechanistically and informationally driven processes

A typical example one can look at in order to study the phenomenon concerns the sunflower. We can construe the movement of the sunflower's blossom following the trajectory of the sun as caused mechanistically by impinging sunlight. In this reading the relation between sunlight and movement of the blossom is like the relation between the button and ringing of the doorbell: the ringing is brought about by the button's depression in the sense of Aristotelian efficient causation. According to this reading, the sunflower, like the doorbell, does not process information about prior events and reacts to them appropriately, but is rather forced to move according to a chain of efficient causes. The alternative reading is that the sunflower processes the information about the whereabouts of the sun based on the light that impinges on the flower. It *extracts* the information and *then* reacts appropriately. According to the mechanistic intuition, the sunflower is merely a part of a causal

⁴⁰ If the mechanism conflates two kinds of causes into one effect, the condition of non-equivocation of an information channel is not satisfied. So, to be precise, mechanisms that are also information channels with respect to a set of possible causes must satisfy this structural condition.

mechanism. According to the informational intuition, it is an agent that processes information and acts both on the basis of this information and according to its vital purposes. A similar typical case which divides intuitions over whether a process is information-driven or driven by efficient (or mechanistic⁴¹) causation is the case of the thermostat. Some would say that it is a semantic device (it processes information about the ambient temperature in order to regulate it); some would say it is a simple causal device with no semanticisation (for a discussion, see Bogdan (1988)). We can call it the 'semantic stance' (in analogy to Dennett's 'intentional stance' in Dennett (1987)) if we impute the capacity of semanticisation to some system in order to describe its behaviour.

There are certain criteria that make a difference to our intuition of whether we see efficient or information-constrained causation. For example, it would not be, to say the least, very straightforward to attribute the semantic stance to a billiard ball when we see it moving after the impact of the cue ball. As the analysis in section 5.2 has shown, and as 6.2.4 will confirm, information flow requires physical channels with relay stations that are physically affected by prior parts of the channel, and that pass that physical influence on to the next station. At the final stage, there is a kind of epistemic interaction with a signal. But a billiard ball seems to have no way (as far as we can tell) to physically represent 'information' about another approaching ball, such that it can *then* react based on that information by moving spontaneously. That is because the only observable way in which the billiard ball is affected by the event of the approaching cue ball, and therefore the only way in which the ball could represent physically the appropriate information, is its own movement. But this is its reaction *already*. The direct contiguity of cause and effect featured by this example is not, as one might think, a necessary criterion for why we are sometimes reluctant to impute the semantic stance, as the following example will show.

Imagining a long series of dominoes we do not naturally think that the last domino falling 'extracts the information' that the first domino in the series had been tipped over, on the basis of the 'signal' that the last but one domino is falling against it. In this example, there is not at any station a process of *referencing back* to past events, but only concatenations of Markovian pairs of events: once the proximate cause is instantiated, the previous events in the causal chain, including the distal cause, are irrelevant. For the last domino falling, the length of the chain is irrelevant.

Contrast this with a letter that is written in Australia, to be sent to Europe. It refers to some contingent matter of fact,⁴² say the death of one of the addressee's relatives, which can be seen to have caused the writing of the letter. The letter is then passed from one messenger to the next, which can be seen as intermediate relay stations, until it arrives at the final recipient. In explaining any behavioural reaction the reader of the letter might show (we can assume she will be emotionally affected), the intermediate stages *including* the most proximate cause seem to be irrelevant. The crucial difference is the event *referred to* in the letter, and the referential relation between letter and event seems to enable the causal power of the letter, not any of the intrinsic physical properties of the communication medium. This, as opposed to what happens to the series of dominos, is an example of causation by information.

⁴¹ I use the terms efficient, mechanistic or Markovian causation interchangeably when contrasting them with causation by information. All three expressions are adequate to highlight the aspect of memoryless causation, for which distal causes have no further effect once the proximate cause is set.

⁴² Another remark concerning terminology: When I refer to '(causally relevant) past matters of fact', I do not want to imply that these kinds of causal relata are not events. So, I do not use 'matters of fact' as a contrastive term in contrast to 'events'. I refer to these relata by 'matter of fact' instead of 'event', since those are events whose existence is preserved as facts by being informationally represented and thereby have the potential of becoming causally efficacious.

One can look at another pair of examples, which both feature the same causal agent (a chemical agent in this case). This shows even more strikingly the connection between semanticisation of information and what we consider to be a thereby caused *action* on the side of the effect. Consider the two relations:

- (1) The toxic gas caused lung injuries as it spread in the building.
- (2) The toxic gas caused people to flee from the building.

The first process does not involve semanticisation. But the second does, if we assume, for the sake of simplicity, that the people involved are fully aware of the threat, and could therefore account for their rushing out of the building. They understand that a threat to their health and safety is present and decide the best option to take is leaving the building as quickly as possible. Along with the semanticisation in (2) comes the feature of action, which (1) lacks. What has semanticisation to do with action as realizing one of several possible options? For sure, whereas one cannot say that it would be an option *not* to get injured given one inhales the gas, one can well say that, although contrary to one's interest, it would be an option to stay in the building, rather than fleeing, thus becoming exposed. But attributing the ability to ponder alternatives requires the imputation of semantic capacities in order to make sense of the alternatives in the first place. The default way we can expect the scenario to unfold is that the spreading news about a toxic substance causes flight reactions – hence the causal connection in the judgment (2). But the semanticisation of that information by some involved agent 'cuts through' the connections of this mechanism (i.e. the *spreading-news-mechanism*), via which the toxic gas exerts its remote influence. By virtue of the semanticisation by digitalising the information, room is made for an alternative decision. No such cutting-through occurs if the gas is inhaled, because here the agent does not process the gas *as a signal* carrying information about *another* event. In short, if an agent, due to the lack of semantic capacities, cannot make sense of the situation it is in, then it cannot choose between different alternative actions either, because these alternatives would be meaningless for the agent. Therefore, imputing the capacity of choosing between alternative options implies imputing the semantic stance.

To recapitulate the result so far, we have found that information is merely the carrier of some difference-making factor that was realised earlier. But information and past matters of fact do not simply stand in a relation of proximate and distal cause, although there is nothing wrong in saying that the past matter of fact caused the information referring to it. The simple relation of proximate and distal cause is instantiated in the series of dominos. The last but one, say, tips over the last one. The first domino of the series was the distal cause of the last one's falling, and also a cause of the last but one's falling. But if we assume that somebody's hand had intervened on the first domino, such that the series of falls was started, we do not consider the first domino as a 'highlighted' cause in any sense. The hand could have brought about the series of fallings at any number in the series. Causation by information works differently. Here, the way the signal takes makes not much of a difference, as long as the signal carries, by virtue of its structure, information about the past matter of fact. The people rushing out of the building filled with intoxicated air might have been informed via very different pathways, e.g. seeing the effect of the gas on someone who has inhaled it, being warned by others, hearing an evacuation alert signal, etc.

Causation by information raises questions. It seems as if the causally inert past – and also counterfactual facts, via *mis-* or *dis-*information – becomes causally efficacious via information. That is a conceptual, and maybe an even ontological problem. Moreover, an epistemological problem we

have at this stage is that even if we count information, rather than its physical carrier, as a cause of action, we still have to explain why, by means of *observations*, we can tell apart efficient causation and causation by information. Why does the interaction with a physical signal differ from efficient causation, if we cannot tell whether it was information contained *in* the signal, or the physical properties themselves that cause the subsequent physical behaviour? Sure, we normally do not observe light rays, or sonic waves, or pheromones, but the cause object, which makes a remote difference only via these means. But even if the physical carriers of information are conspicuous, it seems we can tell the difference between the two types of causation. It is possible to 'drill down' to arbitrary levels of detail to track the physicality with both efficient causation and causation by information, but for the latter case we are running the risk of completely missing the story of causal relevancy. In causation by information, the object of the referentiality, the actual explanans, is dropped at some point. It needs to be underscored that this is a problem of judgment from the third-person perspective, because the first-person perspective (dealt with in sections 5.1 and 5.2) with its operations of semanticisation and direct action, does not allow such a drill-down, since the objects of semanticisation and direct action are immediately and unanalyzably connected to the subject.

When the agency-formula is applied to examples of efficient causation (for now: as it is intuitively understood in contrast to causation by information), we do not take the meaning of the free 'action' at face value, a fact that can easily be shown by applying the scheme to a chain of efficient causes. We need the notion of action to easily identify the cause-object as the source of the causal influence. But with the dropping of the agency-scheme we lose the action. This can be exemplified by a causal chain consisting of a billiard cue and two billiard balls, directly and indirectly affected by the striking of the cue. The causal chain thus can be represented as: A cue approaches a cue ball (event *A*), strikes the cue ball such that it rolls toward the black ball (event *B*), which in turn brings about the movement of the black ball (event *C*). Or: *A* causes *B*, *B* causes *C*. To prevent concerns that the two *B*'s might not be identical, and thus not give rise to a causal chain, I will frame the event concerning the cue ball more carefully as moving on its actual trajectory for some distance in between the final position of the cue and the initial position of the black ball. Although the crucial part of *B*'s causal influence on *C* is its impingement on the black ball (let us call this *B**), the rolling toward that very place is a necessary precondition of reaching that position, and therefore a cause of the latter. By means of causal transitivity, *B* causes *B**, and *B** causes *C*, therefore *B* causes *C*. Now, *B* is clearly an effect of *A*. We now apply the agency-formula: *A* causes *B*, iff *A*, brought about by a free agent, is an adequate means of bringing about *B*. That seems correct. Does *B* really cause *C*, too? Yes, since, according to the agency-account, if we bring about *B* by any means, of which the proper use of the cue would be one, then *C* would follow. *B* clearly has a dual role, as cause and as effect. As the role of effect shows, *B*'s movement cannot be spontaneous, although this is what its having the role of cause with respect to *C* depends on.

We can do the same test with what would be judged a causal chain comprised of both causes and actions, *independently* of the agency-approach. Take a pair of amoebas, of which one, by virtue of its smaller size, qualifies as possible prey of the other. The appearance of this smaller amoeba in the range of perception of the bigger one (event *A*) leads to a movement of the latter towards the former, and eventually to the small amoeba being devoured (event *B*). Whatever follows causally from event *B*, e.g. the small amoeba being digested, can be designated by '*C*'. We first apply the agency-formula again: Setting *A*, say, by positioning the small amoeba within range of the big one, leads, in a regular fashion, to the bigger amoeba's reaction of approaching the smaller

one. Bringing about B, perhaps in the fashion of remote controlling the behaviour of the amoeba, similar to experiments that have been performed on cockroaches (Holzer and Shimoyama (1998)), would lead to the process of digestion, as if the amoeba had acted on her own. The causal link between the processes has again been verified by this procedure. Now we drop the agency-scheme – do we get back, like with the billiard balls, the causes from the actions? Not this time, since the amoeba’s action of seeking and gobbling the prey remains an action according to our judgment. The only difference from our mode of evaluation is that, in the actual course of events, the action is the amoeba’s *own* action, not the experimenter’s actual action of invention, or the theorizer’s hypothetical action, in order to mathematically evaluate a causal claim.

Thus it seems that we cannot drop some scheme of agency in the amoeba-type examples in order to get efficient causation. Even if actions are seen as an effect of a prior event, they seem to retain some aspect of spontaneity. We could, modestly, content ourselves in observing that there are some instances where we stick to the agency-scheme, if the cause object is somehow similar to ourselves, in the sense that it semanticises information and directly acts given a supposed goal. But if it were possible to determine some objective, structural criteria on what that judgment hinges, then it might be the case that these criteria are what underlie or own first-person perspectival notions when we impose the scheme of agency. The first- and the third-person perspective would mutually sustain each other.

We have thus two questions that require an answer, the question of referentiality and the question of the free action. How can the seeming reference back in time to a past matter of fact be explained in causation by information? And – the epistemic aspect of the question – how could we determine the extension of the phenomenon as external observers? Secondly, might we be able to explain how the agent perceives its actions as ‘free’? Both phenomena seem to contradict feedforward models of causal determination like Markov chains. The answer to the second question depends of course on whether one believes that the two actions, the action of the free agent that is also a theorizer on causation, and the natural action that can be observed by another agent, are identical. When revisiting causation by information in section 6.2.6, I will try to give an answer to the question of referentiality that, while negating that a single Markovian model has a sufficient expressiveness to depict causation by information, asserts that higher order Markovian models can be given that explain the phenomenon. Regarding the question of action, I will argue for that there is some plausibility that the two kinds of actions are the same, and that therefore the mutual sustainment of first- and the third-person view on actions can be made plausible. This would have quite a remarkable consequence. It would assert that the contradiction between first- and third-person perspectives does not need to be resolved, and that the objective third-person view would be left completely undisturbed by subjective considerations, while the latter could claim truth even if they seemingly contradict an objective view. Thus, the explanation suggested by this thesis would espouse a dual-aspect approach to causation.

Causation by information is the key phenomenon that will guide the train of thoughts of large parts of this chapter. Looking for an adequate path to approaching ways to elucidate the phenomenon, I will first delve into the thermodynamics of spontaneous and non-spontaneous processes (section 6.2.1 and 6.2.2), on the basis of which a thermodynamic account of actions can be developed (section 6.2.3). The role of a paradigmatic model of a causal agent will be played by Maxwell’s demon (section 6.2.4). On the basis of that, I will try to demonstrate that actions by necessity need to be informationally constrained, such that the matter of fact that produced the information signal can be ascribed causal responsibility. And since the kind of intentional actions in

question require information that refers to *contingent* matters of fact, by necessity an information channel is required, with the endpoint of which the epistemic agent can interact. The link between agency and information will require a computational model that can be ascribed to the agent. From this model (section 6.2.5) it will follow that actions appear as unconstrained to the owner of the model, which will explain how ‘free actions’ can be embedded into a web-of-causes-and-effects-view of the world. I will then show that the properties conventionally ascribed to causation (locality, regularity, asymmetry (section 6.3)) are attributes of agency as well, which gives further evidence for the link between agency and causation. On the basis of these properties, I will characterise a level of abstraction for *efficient causation*, which can be defined *a posteriori* as causation without identity-relations. At that point, the reconstruction of causation is advanced to such a stage that the limited scope of binary judgements, which the agency scheme affords, can be extended so that additional structural aspects of causation can be taken into account, and it can even be outlined – this part will remain very sketchy, though – how other approaches to causation, which do not rely on agency, can be accommodated.

6.2 Natural agency and its properties

As the examples considered in the previous section indicate, there is a tendency to consider semanticisation of information as a concomitant indicator of actions. How strong is this link? And is it of any use for the overall argument in favour of the agency-approach to causation, given the fact that semanticisation is as unobservable as action itself? From the first-person perspective, both semanticisation and action do not require objective truth makers. For me, the agent endowed with a first-person perspective, a message can have a meaning even in case it does not carry any truthful information. Analogously, if I find out that I can control a light via a button with which I can directly interact, I have verified a causal connection from my perspective. In both cases, my conclusions might be incorrect, but what I have established suffices to explain further judgements, viz. feeling better informed than before the reading of the letter, or feeling in control of an observable object. The preceding subsection, with its concept of causation by information, raised at least two further concerns. First, the agency-formula prescribes a concept of a 'free agent'. To establish the first-person impression of *control* of, say, a light via a switch, the corresponding *free action* indeed seems to be a necessary ingredient of the agency-formula. But now the putative action is one the judging agent does not perform itself, but it is ascribed to the agent in its role as object of the judgment, and its action is *caused* by information, viz. it is *not* free. Secondly, causation by information seems to contradict the causal Markov condition, which screens off more distal causes from the effect, whilst now the proximate causes seem to be irrelevant as long as they perform their function of relaying information about the distal cause, the matter of fact which is the actual cause of the observable behaviour. The dilemma of free yet caused actions also offer an opportunity, namely to learn more about actions. The notion of free action of the preceding section is useful in judgments from the first-person perspective. But it has remained unclear how the notion can be integrated into a naturalistic view of the world (understood in a Quinean sense). In particular, it is not straightforward to see how actions can be not only causes but also *effects*, since this is their causal role in causation by information. By virtue of the extensional identity of *B* in a causal chain $A \rightarrow B \rightarrow C$, where *B* has a double appearance as effect (' $\rightarrow B$ ') and as cause (' $B \rightarrow$ '), we can hope to learn more about *B* as *cause* if we bring to light the constraints it is subject to as *effect*. This is information we have been lacking so far.

By following this strategy, we will have to assume a perspective that does not make use of any first-person considerations. We therefore have to look into a theory that is devoid of that perspective. Among the possible theories for a physical underpinning of actions, the thermodynamics of anti-entropic processes seems a good candidate, since they identify processes that require an explanation from those that do not. Hence, they allow a causal interpretation. Anti-entropic, or non-spontaneous, processes feature both the asymmetry and the locality that are properties of every-day examples of causation. From thermodynamics and causation I will then synthesize thermodynamic agency. In section 6.2.4 I am going to reuse a model scenario taken from thermodynamics, Maxwell's demon, to show the link between information and causation from a physical perspective. The actions of Maxwell's demon can be taken as a paradigmatic model of acting with certain properties. Via some considerations concerning the ancillary notion of *centralisation* and an internal computational model of an epistemic agent, I will revisit causation by information and give an explanation of this phenomenon on the basis of the new concepts. It is then possible to identify the kind of processes that are at the origin of our experience of acting, an account that goes beyond the notion of ostension. We can then judge actions from both the first- and the third-person perspective, and we also know that both concepts are linked as much as they

can be linked across the divide of the first- and third-person perspective. From a third-person perspective, an observed action will turn out to be an efficient cause, and whichever the constraints of actions are, must be constraints of causes, too.

6.2.1 Thermodynamics, spontaneous processes, and entropy

It has been acknowledged for a long time that there is some connection between thermodynamics and causation, and that the causal arrow, the arrow of time, and the thermodynamic arrow of the increase of entropy are somehow linked (Reichenbach (1956), Lewis (1979)). At this stage we will exploit another conceptual pairing between causation and thermodynamics that suggests itself. In contrast to what seems to be a more common view, I will not interpret the temporal succession from cause to effect as a course aligned to the gradient of entropy increase, but instead will focus on *anti-entropic* processes, by interpreting them as those kinds of processes which are in need of explanation. The emphasis is thus put on the *relation* between two thermodynamic processes that have a certain structure.

In thermodynamics, there is a notion of spontaneous and non-spontaneous processes. A spontaneous process, considering a system as a whole, is one where the entropy of the system increases. Non-spontaneous processes, on the other hand, cannot be observed in an isolated system apart from small-scale fluctuations, since such an observation would contradict the second law of thermodynamics. If we do observe such a process, we would have to consider a second process of greater magnitude that *is* entropic and therefore spontaneous, and which could, by virtue of its greater magnitude, account for the decrease of entropy in the system we first observed.

Somehow, we are able to judge whether a process is spontaneous or not without having studied the theory of thermodynamics. For example, we can tell immediately whether a film is shown in correct or in reverse order. Interestingly, a reverse playback would be considered improbable, if, say, we saw an egg putting itself back together and jumping onto the table and into an eggcup. Nevertheless, we could intellectually square this with a deterministic trajectory of all single atoms involved, and the reversal in time of these single trajectories does not seem to have a bearing on the question of causation. Causal insufficiency of the observable factors cannot be the cause of the perceived improbability, for if the normal order of playback covers the total of a causally sufficient system (for example, when a camera's perspective captures all parts of an *isolated* system), then its reversal should also be causally sufficient. Since the impression of improbability remains in scenarios which are causally sufficient, a Kantian account (Kant (1787)), which reconstructs the temporal order on the basis of causal asymmetry, seems to be implausible. If we try to ground the temporal direction on thermodynamics instead, there would still be the problem of explaining why we think temporally alongside the thermodynamic arrow of entropy increase, and why a breach of this way of thinking gives rise to impressions such as those that tell us that something about the picture of reverse playback must be wrong. I want to bracket that question, though, and again refer to my assumption of an arrow of time independent of causation, such that causal questions are only prospective. Thus, I can focus on the remaining question of how the causal and the thermodynamic arrow are related.

So, I will in this section accept the apparent fact that we can tell phenomenologically which processes happen naturally and which require further explanations, and relate these intuitive judgments to thermodynamic notions of spontaneous and non-spontaneous processes. If we focus on specific changes of parts of a bigger system, we can draw the connection between improbable developments of events and causes needed to explain the improbability. The most general regularity

that can be observed in the universe is its increase of entropy. Causal changes within a system depend on drawing boundaries of events, and the roles of agents and patients of these changes turns on whether the events are judged as happening spontaneously or non-spontaneously.

The idea can be illustrated with some examples. After putting a glass of hot water on a table in a room at room temperature, we would expect the water's temperature to settle at the ambient temperature, which would correspond to a state of thermal equilibrium of the combined system comprised of the glass of water and (the rest of) the room. If we allow some more time to elapse, we would expect the water to evaporate completely from the glass and the ambient humidity to rise accordingly. In contrast to this scenario, if the water's temperature decreases below the level of ambient temperature, then we would judge that some specific cause must have acted on the water to account for this unexpected development of events.

Switching to a mechanical example, we know that if a windmill starts moving although no wind is blowing, there must be another process somewhere else, connected to the shaft or the blades, that drives the movement.

Generally, any isolated system's entropy is allowed only to stay at a constant level, or to rise, but not to decrease. An adiabatic, sealed container filled with oxygen at constant temperature is an example of a system of constant entropy, while the combustion of some additional fuel in such a container is an example of a process of entropy increase. Decrease of entropy can only be observed in *parts* of an isolated system. If the system's boundaries have not yet been clearly defined, and a decrease of entropy is in fact observed, then the boundaries have to be expanded or redrawn such that another process of entropy increase of higher magnitude can account for the observed decrease of entropy.

6.2.2 The connection between thermodynamics and causation

When it comes to the connection between physics and causation, thermodynamic processes are a good candidate for conceptualising causal relata, and they also provide the explanatory means for why agency assumes conceptual priority over efficient causation.

There is an explanatory asymmetry involved in the relation between the non-spontaneous process and the spontaneous process that drives the former. The spontaneous process happens by itself, without further need for an explanation, apart from a possible need for explaining the exact time at which an event has started to unfold. For example, when water starts evaporating from a previously sealed container, the fact that the seal was removed counts as such an explanation. But the process of evaporation itself happens for statistical reasons. If all the molecules reversed their course and strove to get back into the cask, that would be an observation that calls for further explanation. Similarly, if a meteor in free space continues to move along its set trajectory, we do not ask for a cause of this continuity, although we do see its movement as a 'causal process' (Salmon (1994)), *if contrasted* with the movement of a shadow. But if the meteor changed its course, this would prompt us to look for a cause. It is this narrower sense according to which I understand the relation of causation in this chapter. The asymmetry problem of causation thus construed consists in the non-spontaneous process requiring the spontaneous one, but not vice versa.

Likewise due to the second law of thermodynamics, the examination of any object under investigation requires in thermodynamics its division into system and surroundings. What is said in this section is, of course, about thermodynamics, but since what follows in the subsequent sections is more general yet still in line with thermodynamics, I will for the sake of simplifying the language refer to the surroundings by the more general name 'environment' throughout the rest of the thesis.

The notion of environment is important because often systems are not isolated, but constantly exchanging matter and energy with their environment. This is how changes of decreasing entropy become possible in the system at the cost of their environment, by exchange of heat. The fact that we have to draw boundaries around a considered thermodynamic event is also relevant because later on an aspect of commonsense examples of causation, its *locality*, will be linked to thermodynamic boundaries.

Non-spontaneous processes, if observed in isolation, are *surprising*. In this context, other observations from studies of causality come to mind, like Hall's default/deviant distinction, which are system behaviours that are considered normal and surprising respectively. They will be discussed further in section 7.1.4. (Hall (2007) gives Christopher Hitchcock partial credit for this conceptual distinction.) If we consider thermodynamic processes as special cases of events, then the non-spontaneous processes count as the events that require explanations. A one-to-one correspondence between effects and non-spontaneous processes on the one hand, and the effects' causes and the spontaneous processes that drive the non-spontaneous ones on the other hand, suggests itself, but is not strictly possible to adhere to. Certainly there are cases where this holds true, which is encapsulated in some sayings of the older wisdom of causality, like 'effects cannot be greater than their causes', or Leibniz' interpretation of '*causa aequat effectum*' (Leibniz (1695)). While converting one form of energy into another, as a fuel-driven electric generator does when it converts chemical energy into electrical energy, we get, as the effect of the process, at most the amount of energy we put into the system, the cause of the process. In reality it will be even less because no generator runs with complete efficiency.

There are cases in which causes merely trigger big effects, like when the trigger of a bomb is activated. For these kinds of cases, however, a more fine-grained explanation can be given that re-establishes the order of *causa aequat effectum*. In the bomb example, we can expect that, whatever the exact mechanism is that sparks off the explosion, there is some threshold that needs to be overcome, e.g. depressing a button. The process that enables the spontaneous process, the 'big effect', the bomb's explosion, to unfold, is non-spontaneous and requires an explanation, like the fact that someone depressed the button by exerting a force on it. The application of the terms of proximate and final cause, for the triggering mechanisms itself, and the action that activates the triggering mechanism, respectively, suggests itself for these configurations. Previously used examples, the removal of a seal from a cask, or pushing a button to turn on a light, have the same structure and thus feature non-spontaneous processes as proximal effects.

Other examples have a more straightforward structure, from a thermodynamic viewpoint. When a cart moves uphill, someone must be pushing it. The roles of cause and effect in 'A is a possible cause of B' for these kinds of cases can be reformulated as: 'A is a process that is capable of driving a non-spontaneous process B'. This approach takes into account the question of what kind of entities the causal relata are: We first identify B as an event that is of such a kind that it requires an explanation. B cannot happen without something else happening first. As the explanans we then identify an appropriate A. The question of spontaneity of A does not have to be taken into account explicitly, due to A's role of being an explanans, not an explanandum. Therefore, addressing the question whether A itself requires an explanation is not required to evaluate the judgment concerning the causal connection between A and B. To abstract from the question of spontaneity of A reflects a commonsensical judgment concerning the dual role of an event A in a causal judgment 'A causes B' – *although at this point we have to be very careful with not conflating the terms of spontaneity in the action-theoretic and in thermodynamic sense*. The role of A as the cause of B

suggests the spontaneity – in the action-theoretic sense of spontaneity – of *A* occurring to bring about *B*. But there is no contradiction in assuming that *A* is embedded in a larger nexus of causes and effects, and that it happens itself due to another cause *A**, preceding *A*. Spontaneity, in both senses, is a relative notion. In the thermodynamic sense it depends on how one divides an observation in space-time into system and environment.

If *B* is explicitly understood as a spontaneous process in the thermodynamic sense, then the explanatory need connected to its occurrence is due to a necessary triggering factor. This can be the activation of a trigger to make a bomb explode, or the supply of activation energy of a spontaneous chemical reaction. In these cases, ‘*A* causes *B*’ translates into ‘*A* is a process that triggers the spontaneous process *B*, which would not have happened at that time without the triggering event occurring’.

6.2.3 Thermodynamics and actions

I have indicated the connection between thermodynamics and causation. What is missing is the connection between thermodynamics and agency. This link is established by the thermodynamics of steady states (see Bertalanffy (1969)). These are specific kinds of homeostasis that feature a constant input of free energy into a system. Steady states are to be contrasted from equilibrium states, which can be observed when an isolated system has reached the state of maximum entropy. For systems that are connected to their environment by diathermic containment, the equilibrium state is reached when the temperature of the system has reached the level of temperature of the environment. If the walls of the container are permeable for particles, in addition to the thermal gradient the gradient of diffusion has to be zero for all kinds of particles that can pass through the walls. By contrast, a steady state is a state that is *not* characterised by maximum entropy. Organisms live in these kinds of states, and therefore have to continually procure free energy to maintain their own structure and prevent the macrostate of maximum entropy, which means death. The steady state is a rare and anti-entropic state, but it is also a functional state, since only from this state can the organism perform its function to sustain itself. These properties of the state, the functionality and the improbability, are logically related to each other. In mechanical machines the same relations can be observed. One cannot create a watch just by shaking up its parts, because the likely outcome, a state of high entropy, is also one of the many dysfunctional states.

Agency enters the picture with the *interest* of maintaining the functional state far from thermodynamic equilibrium. Considering causes in efficient causation, we might speak, at least in the English language, of agents in a wider sense, like ‘chemical agents’. But in genuine cases of agency there is a relation of *identity* between the acting agent and the agent that benefits from the action, and this identity needs to span the time from the decision to act until the time of consumption of the action’s utility, which means, in its basic form, the maintenance of the steady state. Achieving this vital goal requires continual interventions. Animals maintain their functional state of low entropy by foraging for sources of free energy, or Gibbs energy. It is mostly contained in the form of chemical energy in plants or other animals, and ultimately has its sources in the electromagnetic energy of the sun (Atkins (2007)). By the process of oxidation ATP is created from carbohydrates. ATP can be considered the fuel of the cell, and by this container of Gibbs energy both the anti-entropic (and non-spontaneous) process of continuously re-establishing the functional state, as well as the externally observable behaviour of the body, is driven (see also Weber (2014)).

But bringing about non-spontaneous processes by exploiting sources of free energy is not by itself a criterion of the kind of agency we need in order to apply the agency-formula. There are other

non-spontaneous processes driven by sources of free energy, like certain kinds of physical pattern formation, with a dynamical equilibrium that ceases to exist once the input of energy into the system ceases. The link to agency is supplied only by considering the flow of information. For sure, it makes sense to say 'Eat this (*A*), such that your hunger is satisfied (*B*)', and to conclude 'Eating this edible object causes the hunger to be satisfied'. But this judgment cannot be easily applied from a third-person view. It will not help with the question of the semantic stance, and therefore will not be a solution to the problem of causation by information. To solve the problem, I will make an assumption whose validity will be demonstrated in the following section about Maxwell's demon. The assumption to be made is that actions are events that are caused by information, and therefore require an acting agent that is also an information processor. Since the second assumption, expressed earlier in this section, implies that an action's immediate effect is a non-spontaneous process, we need to show that the flow of information is continuous with physics, since non-spontaneous processes need to be driven by spontaneous physical processes in order to comply with the second law of thermodynamics. The result has been anticipated already in the analogy between information flow and causation, and will be shown to be a necessary result of the second law of thermodynamics in the section covering Maxwell's demon as a paradigmatic model of acting. Information that triggers the 'right' action to bring about a non-spontaneous process needs to be processed as well (first perceiving, or measuring, its physical signal; then calculating the optimal action). That process, as it is shown by the thermodynamics of information processing, is entropic.

Actions of the kind required by the agency-formula occur spontaneously, but, as causation by information shows, there are also *caused* actions, a seeming contradiction. As the notion of efficient cause shows, this seeming contradiction does not have to be a fatal conceptual problem. The thermodynamic framework, which promised to provide more information on the constraints on causation, might also endow us with a resolution of this conundrum. The commonsensical intuition of causation by information implies that information is the cause of action. We can replace, according to the argument, the idea of action as realisation of several possibilities by the idea of action as the result of both receiving and processing information *and* the acquisition of a token of utility. The optimisation of the choices concerning the possible actions to take in a situation constrained by information yields the regularities that a causal judgment requires, while still allowing for the intuition of seeing an action. Given a certain contingent situation and the belief-desire configuration of the agent, there is one optimal way of behaving, i.e. the rational choice. For a judgment on type level, the observable reaction of an agent that has processed the contingent information about a given situation will be the rational choice according to its internal model. More on the internal model will be said in section 6.2.5. The constraint concerning the token of utility is necessary to judge whether an action has been taken or not, since the processing of information alone, while yielding a set of possible choices to take, does not favour any single one of these choices over the others, and therefore would fail to give rise to the regularities that our judgement requires.

Acquisition of a token of utility, on the other hand, introduces further problems, notably to judge when observed behaviour of another agent is 'adequate' for reaching a goal (Dretske (1981), chapter 8; also Dretske (1988), chapter 5). In order to do this (implying the intentional stance), we have to have a natural notion of utility. As will be explained in the subsequent section, we will posit entropy reduction and acquisition of free energy as the natural utility, with all goals on higher levels building on this basic notion of utility.

Assuming the notion of utility is correct, and expanding entropy reduction on information as well, I propose the following interpretation of agency within the thermodynamic framework: 'An action *A* brings about *B*' means 'A is a process that is capable of driving a non-spontaneous *B*', with the following additional constraints: *A* is an observable event attributable to a structure that benefits from the event *B*. This entity is called agent and it can be assumed that it maintains its identity between the time of doing *A* and benefiting from *B*. What the benefit consists in has to be explained further according to context. Since the physicality of the concepts is at stake in this section, suffice it to say here that the utility can consist in consuming a token of free energy, or the reduction of entropy of the structure that we identify as the agent. The utility can also be informational, such as increasing the knowledge (or decreasing the uncertainty) about a contingent matter of fact, but even in this case the token which represents the flow of feedback back to the agent must have a physical basis. The entropy, for the cases where the token of benefit is information, is to be interpreted as physically embodied entropy *relative* to another situation. The reduction of this *referential entropy* can be interpreted as knowledge about a contingent matter of fact that might be of vital interest *later* (for a similar conceptual distinction like that between physical and informational entropy reduction, see the distinction of 'vital' and 'active goals' in Bogdan (1988)).

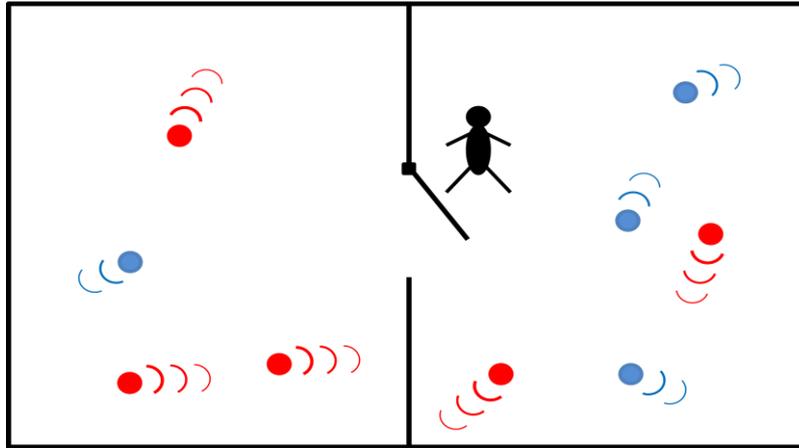
6.2.4 Maxwell's demon as paradigmatic causal agent

Energy cannot be created from nothing. Correspondingly, a so-called *perpetual motion machine of the first kind*, which claims to do just that, is impossible. The insight that entropy puts a constraint on the production of energy, too, came later. The second law of thermodynamics precludes the creation of a *perpetual motion machine of the second kind*. Such a machine could extract energy from the environment, e.g. in the form of thermal energy, and transform it into work, which could in turn be used for all other kinds of purposes. Energy would not be created from nothing, but – almost as good as that – it could be reused indefinitely. However, perpetual motion machines of the second kind are nowadays considered as impossible as those of the first kind.

A process that is thermodynamically equivalent to the extraction of thermal energy from the environment and its conversion into work is the creation of a temperature gradient within a system that is currently in the state of thermal equilibrium. According to the second law of thermodynamics, such a process is non-spontaneous and requires a second, spontaneous process to drive it. To be precise, the lack of such a second process driving the first one would contradict Clausius' formulation (see Atkins (2007)) of the second law:

Heat does not pass from a body of low temperature to a body of high temperature without an accompanying change elsewhere.

Yet James Maxwell devised a schematical setup which, although far from being an operable perpetual motion machine of the second kind, was meant to show that a temperature gradient could be created without expenditure of work that must be put into the system to drive the gradient's build-up. The gradient is achieved by the action of an intelligent being, Maxwell's demon, which performs opening- and closing-operations on a trapdoor through which particles can pass from one half of the container to the other. The thought experiment has been developed over several stages by other authors, and its gist has turned out to revolve around how to model the 'demon', who is the central actor in the setup.



Graphical scheme of Maxwell's demon

Since the system is said to be in thermal equilibrium at the beginning of the process, the gas molecules filling the container will have a Gaussian distribution around a mean value of kinetic energy corresponding to the temperature. No energy is to be inserted from outside into the system in order to create the temperature gradient, so the only way to proceed is to sort the particles such that the particles at a speed lower than the mean end up in one half of the container, the faster particles in the other half. The trapdoor indicated in the graphical scheme is meant to achieve this sorting, if the demon manages to open it in just the right time to let through the particles of the right speed that comply with the intended gradient build-up, and block the particles otherwise. The sorting, so the argument of the thought experiment goes, does not involve expenditure of energy, since the trapdoor's operation is frictionless, and the demon's decisions are *mental* or *intelligent* operations not contributing to the balance of physical entropy. After the sorting of particles has been performed by the demon, the temperature gradient between the two compartments of the container can be exploited for extracting work, e.g. by means of a turbine. In terms of thermodynamics, the demon's actions have decreased the entropy of the system while holding its internal energy constant, which is tantamount to increasing the system's free energy.

However, we have made the assumption that the trapdoor is not a source of friction and therefore not a source of increase of entropy. But we need a rise of entropy somewhere within the container or in another system that exchanges heat with or performs work on the container, in order to explain the apparent decrease of entropy. *Prima facie*, there does not seem to be such a source apart from the trapdoor and the demon. With the trapdoor taken out of the equation, authors writing on that subject have concentrated on the demon and its 'mental' operations; basically the decisions when to make the trapdoor move, as a possible source of entropy in order to reconcile what happens with the second law.

We have to consider the demon's decisions as *systematically* correct. It is beyond a lucky coincidence if the demon manages to separate a significant portion of molecules of different speed. But then the decisions have to be based on contingent information – the whereabouts of the particles. Either the state of every particle is measured once and their trajectories are computed, or the measurement is done locally, in the proximity of the opening in the separating wall, such that it can be decided for a single incoming particle whether the door is opened or not. Historically, it was thought that the conundrum of entropy could be resolved by the entropy balance of the measurement process. Indeed it seems as if the measurement implies some kind of coupling

between two physical systems, but the subsequent mental operation of taking a decision can still be seen as an event taking place in a non-physical space, not contributing to the increase of entropy. A typical setup for performing a measurement uses directed light which is scattered by interacting with a particle (see the discussion in Wiener (1954)). The difference this interaction makes for the background radiation is the criterion for measuring the particle, and, according to this account, the resulting diffusion of light would have restored the entropy balance. But it was later shown (by Bennett (1982)) that there does exist a way of measuring a particle which does not increase any entropy in the measurement device whatsoever. This prompted the search for another source of increase of entropy, and the answer was found in the resetting of the register that holds the measurement results.

To understand what the result means, it is useful to look at another formulation of the second law of thermodynamics, Kelvin's formulation (Atkins (2007)):

No cyclic process is possible in which heat is taken from a hot source and converted completely into work.

We can illustrate the idea of the balance of a complete thermodynamic cycle by looking at one of the stock examples, the heating of a glass of water in a room at normal room temperature. If the heating up passes beyond the room temperature, we know that there is something besides the glass of water and the room that needs to be taken into account to explain the phenomenon. It might be that an electric heater, driven by a battery, has been at work in the water. To comply with the first law of thermodynamics, the conservation of energy, the battery has to 'pay the energy-bill'. But since we are dealing not only with an increase in overall (thermal) energy of the total system comprising the glass of water and the rest of the room, but also a temperature *gradient* between glass and environment, which can be exploited for performing work and therefore represents free energy, we also have to have the battery 'pay an entropy-bill'. This it does by running from the initial state (fully charged) to the final stage (depleted), the state of chemical equilibrium. The gain of free energy in the form of the temperature gradient cannot be higher than its loss by depleting the battery. This can be verified by recharging the battery by exploiting the free energy consisting of the temperature gradient brought about by the working of the battery. By such a process the battery would be set back towards its initial state, the state before it started to heat up the water after the heater had been turned on. However, the input of work to recharge the battery would be higher than the work performed by the turbine between glass and environment. To put it differently, the temperature gradient cannot fully recharge the battery in any real setup of this experiment.

Turning back to Maxwell's demon, we are not dealing with an increase of energy of the considered total system, but only with an (expected) increase of entropy, since we know that the building-up of the temperature gradient is non-spontaneous. There is no energy-bill, but an entropy bill to pay. If Bennett's solution to the problem of Maxwell's demon is believed to be correct, the process does not require input of work up until the creation of some physical representation of the measurement result. But if we consider the time when the totality of operations on the trap-door has been executed, we have a measurement device with a storage-part that is in another state than at the time of the first of all the operations on the trapdoor. Similarly to the battery, which requires work to be recharged, we need a reset-operation to put the storage device back to, say, *state of all registers = zero*. However, Landauer's principle (see Maroney (2009)) predicts a higher input of work necessary to do that than extractable by the turbine:

There are no possible physical implementations of the resetting operation that can do better than to reset a bit to zero converting less than $(kT \ln 2)$ of work into heat

The resetting of a bit that previously held one of two possible states, and therefore represented 1 bit of information, corresponds to a compression of logical space. It was Landauer's assumption that for every compression of logical state there is a corresponding compression of physical space, for which standard calculations show that the expenditure of work amounts to $(kT \ln 2)$. Regarding the entropy change of the system (the particles in the canister), the set of possible states for a particle that has been sorted already has been exactly reduced by one half, since it can only reside in one half of the container instead of possibly occupying all of it. The free energy gained per particle is likewise $(kT \ln 2)$, such that the net gain amounts to zero.

The resetting of the register, the completion of the thermodynamic cycle for the whole system that is causally sufficient for the phenomenon, thus reconciles the thought experiment with the second law. It seems that the final solution of the conundrum is closely tied to the tacit and unanimous *premise* of the analyses of Maxwell's thought experiment: Interventions that bring about improbable results (such as the building up of a temperature gradient without expenditure of work) need to be informationally constrained. If information refers to contingent matters of fact, then a measurement needs to be done that is subject to physical constraints. If this constraint could be circumvented, it seems as if agents like Maxwell's demon would be possible. But due to Landauer's principle, which does not rely on the second law of thermodynamics, whose validity is at stake, intelligent actions of a purposive kind are not a viable way to circumvent the second law, but merely a particular class of entropic processes able to drive non-spontaneous ones.

This result explains why nature has not, to our knowledge, come up with an organism that can extract ambient thermal energy from its environment in order to drive its metabolism. Such an extraction counts as an anti-entropy process, which needs to be driven by an entropic process of higher magnitude. But an organism whose body parts comprised agents of the kind of Maxwell's demon could 'let in' faster particles in order to build up an internal hot reservoir. In relation to the ambient temperature, with which the organism could be thought of as in thermal equilibrium, the hot reservoir would count as a system that contains Gibbs energy, and the organism could live off this energy and could renew it, apparently *ad libitum*. The thermodynamic considerations that take into account the whole of the causally sufficient system explain why we do not observe such an entity anywhere in nature.

What needs to be taken into account from the analysis is that it puts important constraints on both information processing *and* acting. Maxwell originally wanted to show that the second law of thermodynamics has only statistical relevance and is not a law of nature on a par with laws that are true by necessity. That seems true even after the analysis by Bennett, Landauer and other physicists and philosophers who dealt with the problem. After all, the demon could bring about the sorting of the particles by opening and closing the trapdoor at random and happen to be just extremely lucky in succeeding with the task. But what has been shown is that one cannot beat the second law in a systematic way by *intelligent actions*, since these depend on external contingencies and therefore require measurements, whose result needs to be stored in a register. This register needs to be reset in order to complete a thermodynamic cycle, which comes at a cost of entropy increase. This has an import for the notion of information in this thesis, i.e. information interpreted as a referential structure coupled to an external matter of fact, such that it can later serve as a basis for a target-oriented action. Information in this sense needs to be physically represented –

otherwise a loophole for a perpetual motion machine of the second kind would be created. For the study of causation by information we can note that no discontinuity regarding physical laws is to be expected by caused 'intelligent actions'. The difference between causation by information and efficient causation is therefore unlikely to lie in a realm outside of ordinary physics.

Sorting procedures of the kind considered in Maxwell's scenario consist of binary decisions resulting in physical interventions. Since sorting thus understood is such a basic form of acting, and since no assumptions were made concerning further particular characteristics of the demon, or specific ways in which the demon was to implement the requirements of his task, it is to be expected that the two results – the necessity of measuring contingent information for purposive actions, and the necessity of representing that information physically – can be generalised. Under general assumptions, information must be gathered in order to act, otherwise actions cannot be expected to satisfy the second of our criteria for discerning an action from other kinds of events: it was asserted in section 6.2.3 that actions must also be beneficial, and the null-hypothesis is that, in our world of competition for limited resources, decisions to act that are not informed, not constrained by contingent information, do not regularly result in a beneficial outcome. This refers both to competing with an entropic nature and to competing against other agents within an entropic environment. (See Wiener (1954) for the description of both kinds of competitive environments, and Werner (1991) on the relation between information and game theoretic strategies.) One can therefore say that Maxwell's demon is a paradigmatic toy model of acting, and that informed decisions do not come for free. Actions require information, and information about contingent matters of fact, like the whereabouts of hot and cool particles in Maxwell's setup, needs a physical underpinning with a cost of entropy.

From these considerations, the following concatenation of events is to be expected in causation by information:

- A) Generation of the matter of fact which is causally relevant
- B) A physical signal, which is coupled to that matter of fact, interacts with the boundary of the agent
- C) Physical realisation of a computation whose input is the signal and whose output is the optimal action according to the agent's causal model
- D) Observable action of the agent
- E) Situation which can be valued in terms of beneficiality for the agent⁴³

In causal regularities that involve actions, there is always a causally relevant matter of fact realised at some point in time. This fact generates causally a signal that carries information about the matter of fact. If it interacts with an agent that can extract the information from the signal, and for which the matter of fact is in some way relevant, this agent can be expected to react to the signal with an observable reaction. By default, a rational, or at least a fit-for-purpose reaction will follow in most cases. In other words, the scheme of acting will give rise to a regularity, which can underpin a causal

⁴³ Section 6.2.6 will deal with the criteria for distinguishing a reaction to information from an effect triggered by properties of the information-carrying signal. These criteria will turn on the different structural features which the relay stations of information-channels on the one hand, and effective causes in a causal chain on the other hand, have. This will enable us to recognize causation by information. But, as the previous example of the letter informing about a relative's death shows, not all effects of information are actions. But information is a necessary ingredient of an action, as this section has shown. The intentional stance implies the semantic stance, but not vice versa.

judgment. Hence, caused actions are no contradiction according to this scheme. In its final step, a beneficial situation is realised that has some connection to the acting agent.

Maxwell's demon is a paradigmatic agent since it complies with all the steps of the scheme, including the final step, if one adds to the standard analysis of Maxwell's scenario the assumption that the demon is the beneficiary of the free energy constituted by the temperature gradient. By this extension of the standard setup the demon's actions become purposive, or teleological. However, due to the second law of thermodynamics, in a closed universe, sorting procedures are not sufficient to feed functional structures similar to Maxwell's demon in the long run, in order to reset them to a state from which a thermodynamic cycle has started. The thermodynamic cycle considered in the thought experiment did not even imply additional expenditure of physical work that might be needed to restore parts of the agent that could have become dysfunctional. Therefore we know that additional free energy will be needed, and irreversibly wasted, in order to keep a structure like the demon functional. In that respect Maxwell's demon is of course an idealised agent. Natural actions that result in accessing free energy often involve getting hold of some other organism, which contains free energy in chemical form, or involve exposing oneself to directed sunlight, rather than creating the free energy by a sorting procedure. But, as mentioned above, these kinds of actions will be likewise informationally constrained.

Step C of the above scheme involves a computation. This point is quite straightforward. We do not want to explain an action as an effect coerced by a preceding cause. On the other hand, causation by information needs to accommodate actions that give rise to regularities. Both aspects are best squared with each other under the assumption that there is an optimal action for a given informational constraint, which will be the action most likely observed in the regularity. I will therefore model the relation between a signal that carries the contingent information for the agent and the corresponding best action by a function. A function requires a computational model, such that the output for the input can be computed. I will outline the computational model in the next section. Once the model has been described, I will apply it to Maxwell's demon during the revisiting of causation by information. Then it will be seen that the demon is, very much like the sunflower, an example that can be described in two ways: according to a causal reading, and according to an informational reading.

6.2.5 Internal computational model

In this section, I will not devise a full-fledged account of the reasoning processes that an agent undergoes before committing to an action. The rest of the thesis will cover different aspects of what underlies the judgment that a certain intervention is considered appropriate by an agent in a contingent situation. Here, I will only give the desiderata of a computational model, and the constraints that are necessary to distinguish efficient causation from causation by information.

Starting with the desiderata, the model is supposed to contain criteria that enable an observer to classify an action of another agent from the third-person perspective, viz. without invoking some notion of similarity on which such a classification would hinge. For example, *my* judgment of having observed an action is not supposed to involve a change of perspective that exploits my own experience of having performed a similar action in the past. Secondly, the model is supposed to preserve the conceptual frugality of the approach exercised so far. Thirdly, the account of action derived from the model should include actions as causal effects and should therefore not exclude the possibility that actions give rise to observable regularities, while at the same time contrasting actions as effects from those effects which are due to efficient causation. Finally, it

should be possible to recognize actions as such, in spite of the previous points made. This last desideratum is more than a trivial remark. One could read the account of action that results from the setup of Maxwell's demon, the internal model of computation, and the account of centralisation in section 6.2.7, jointly as a technical definition of the term 'action' that also works for some examples of causation by information. But one could also make a stronger claim: if the criteria expounded in the respective sections *constitute* the concept of action, then the action we hypothetically or actually perform when judging causation (the concept of action employed in the agency-formula) would fully comply with these criteria – the concepts would be *identical*. But then we would have found a second way of identifying causes and actions. Not only are actions those kinds of events we hypothetically bring about by assuming a first-person perspective with direct control of some observable event, which needs the (first-person) concepts of direct action and spontaneity, and possibly semanticisation, in those cases when the action is triggered by information. Alternatively, we could judge action as a result of a computational process, which in turn can be read in different ways. A feed-forward interpretation gives a model of efficient causation. A feed-backward reading of the same process could explain the referencing back to an inert past, and could therefore explain semanticisation. But then actions cease to be a subclass of causes, since there would be in principle no way of telling apart causes and actions, and therefore no domain of causes that is not also occupied by actions. Then, if further constraints on actions are found, these would be informative for the concept of cause as well, since every time we apply the scheme of cause to explain an observed situation we actually apply the scheme of action.

Part of the explanatory story involving caused action is the physically embodied information signal, with which the agent has to interact. The signal is a product of a causal process. Via the signal, the past matter of fact turns out to be causally efficacious at a later time. But there is no objective, local property of the signal that can be just read off it in order to get the past matter of fact, since the signal does not 'contain' information about the fact in any literal sense. The signal can have a very different structure than the matter of fact. Therefore, the agent must be configured such that it can extract the matter of fact from the signal. The answer to the question of referentiality therefore lies both in the structure of the information channel and in the computational structure of the agent that processes the signal, and there can be both a failure of extraction and a channel failure. The agent might not be able to interpret the symbol token in cases of semantic information. For example, she might not understand a letter written in a foreign language. Or the agent might not know a law of nature to decode environmental information, like the correct mapping of the number of tree rings to the age when it was cut. Mis- and disinformation are examples of channel failures, as are breaches of the non-equivocation condition of the channel. False perception and false abductions from effects to causes are cases in which it is difficult to say whom the failure in mapping is to be attributed to.⁴⁴

If the inference from all signals to their causes is successful, the agent is informed about the contingent situation in which it is situated. This is when the second stage of the calculation from the signal to the optimal action begins. Although it is conceivable that a non-composite function from the signal directly to the optimal action is computed in some situations, this is unlikely to happen in the general case, since one causally relevant matter of fact can be informationally conveyed via

⁴⁴ Of course, if there is no default channel behaviour, or no normative prescription of how a signal is to be read, then there is no point in making the distinction between the two kinds of failures.

different pathways, e.g. it can be communicated in different languages, or in different styles of expression. For combinatorial reasons, it thus makes sense to distinguish an inference from the signal to its most probable cause, and a subsequent calculation of the best action given the contingently constrained situation.

We can now contrast a scheme of judging efficient causation with a scheme of judging causation by information. We first consider the chain of events from the end of section 6.2.4 again. This was a chain of events that we would expect with examples of causation by information, but here we interpret the events as connected by efficient causation:

(A: Generation of the matter of fact which is causally relevant) causes (B: physical signal interacting with the boundary of the agent) causes (C: physical realisation of a computation whose input is the signal and whose output is the optimal action according to the agent's causal model) causes (D: observable action) causes (E: beneficial situation).

We can think of this concatenation of observable events as seen through the eyes of an external observer. The observer can impose an interpretation that wants to see efficient causation at work. This can be verified, among other ways, by applying the agency formula. For example, given the agent is configured as an information processor, *setting B*, i.e. exposing an agent to a signal that is similar to a signal which would be brought about by *A*, we expect that a computational process is triggered in the agent. Similarly, we can perform 'set-operations' at other stages of the concatenation of events. By this means at every stage the setting of the predecessor yields the successive event. There is no referencing 'back in time'.

To model causation by information that preserves the judgment that the contingent fact was responsible for the observable action, the model of efficient causation needs to be extended by a *type level* consideration. As far as *D* is concerned, not only *C* as an individual computational process, but also *A* and *E* have an influence, *but as types, not as tokens*. *C* is the agent's *individual* way of extracting from the signal *B* the information that *A* has happened, and *E* is the beneficial outcome of the best action *D* in the situation that is characterised by *A*. The beneficiality of *E* consists in a feedback that flows back to the agent and has an impact on *C*. However, our previous, feed-forward scheme did not allow depicting this feedback. The feedback acts like a force that brings about the selection of all those agents in a theoretical population that act such that *E* is achieved, on the basis of interacting with signals, of which *B* is merely one example.⁴⁵ Alternatively, one can think of a single, adaptive agent that is punished by negative feedback and adapts its computational model on the basis of that learning experience. In both cases, we have to consider *types* of situations, not their tokens. This is how and why the intermediate stages, although they are all necessary parts of explaining how the agent learned about *A* and from which point in time the agent could possibly act adequately to *A*, nevertheless drop out as explanatorily irrelevant to account for why the agent has done *D* rather than an alternative *D**.

The following assumptions are made implicitly: the agent has the capacity to infer *A* from *B* or is structured in such a way that *A* becomes part of his information base, and it has the causal background knowledge to predict the effect of *A* on *E* with and without its doing *D*. That is, the agent has at its disposal a model according to which it can determine whether its utility, revolving around the final state *E*, would be increased or decreased by committing to a specific option of acting (of which not acting is one), in the context determined by *A*, or even in collaboration with *A*. Note that

⁴⁵ Section 6.2.6 will demonstrate all of this by means of the setup of Maxwell's demon.

evaluating the same concatenation of events according to the first, or the second scheme, *do not contradict each other*, and are therefore alternative ways of evaluation. Picking one scheme in favour of the second depends on the kind of explanation one seeks.

From a third-person perspective, the occurrence of the immediate trigger of the observable action explains why the action needs to happen, but it does not explain why the agent is configured such that *this* action rather than *another* has occurred. The third-person observer, while observing the action of the agent, needs to understand that the agent's calculation, depending on its internal configuration, is an idiosyncrasy that would *not* give rise to a general rule, if the reference classes are fixed in an unbiased way. That is, if the agent is connected to an information channel, then the signal the agent interacts with is not the event the agent is responsive to, but the event at the *source* of the channel is. This latter event has therefore also been dubbed the 'causally relevant matter of fact'. The agent's action, if successful, reinforces the correct response to increase the chance of future successful interaction in similar situations. The external observer therefore understands the transmission medium as accidental; an alternative channel would have brought about the same observable behaviour. The fact that the causally relevant fact, instantiated in the source event, was mediated by *that* signal, is as contingent as additional, causally irrelevant properties that the source event may have had, say, the exact shape of a lamp that emits light and thereby provides information about an emergency situation.

I assert that the model satisfies the four desiderata above. It satisfies the first and the second one, since only correlations between observable relata are involved. There is a relation of efficient causation (it can be assumed that the external observer has the concept at its disposal) between all the pairs *A* and *B*, *B* and *C*, *A* and *C*, etc., on the token level, viz. for a considered individual agent. There is a correlation on type level between *A* and *C*. And there is an observable feedback from *E* to *C* in the context given by *A*, either by selection, or by adaptation. Given that the fourth desideratum is also satisfied, then the third desideratum is satisfied, since an action is the result of a computational process that computes a function value, which corresponds to an optimal action in the situation. In other words, given that the action is indeed classified as an action, then the model assures that a regularity arises in repetitive scenarios.

6.2.6 Causation by information revisited

We can now return to the problem of explaining causation by information. The problem consisted in elucidating why an action can be caused while still retaining the aspect of spontaneity that is needed to classify the observable behaviour as an action. The proposed computational model solves this problem by not letting the agent interact directly with the cause that determines its behaviour, but instead with its signal. Several different signals, via different channels, can result in the activation of the same rule from the causally relevant property of the channel source to the observable behaviour, therefore the intermediate steps that concern the channel drop out of the scheme as explanatorily irrelevant. In this sense the event at the channel source is highlighted in a unique way unlike other distal causes that might be highlighted in some specific contexts. Neither from the external perspective of an observer that is not identical with the agent, nor for the agent itself, reflecting on its own action, is there a direct contact with the cause. This solution comes at the cost of referentiality, which is *prima facie* difficult to account for in causal contexts. However, the reference of the signal to its cause has been taken into account as abduction from effect to its efficient cause, and the abduction process has in turn been explained by the feedback stemming

from the final, beneficial state, which alters the configuration of the agent in the situation constrained by the distal cause.

The minimal account of an action suffices for distinguishing causation by information from efficient causation, by switching to the appropriate scheme of actions guided by internal computations, if we have reason to believe that the simple scheme of concatenated, efficient causation is explanatorily unsatisfying. The scheme of efficient causation, however, always delivers a correct account. This seems to create a paradox, since we use the two schemes to distinguish the two types of causation. This paradox will be resolved in section 6.2.9. For now it suffices to assert that the computational model delivers an account of action that is sufficient to explain causation by information. According to both schemes, a contingent situation comprises some causally relevant matter of fact. This fact generates further facts via causal mechanisms, of which some possibly interact with the boundary of an agent. For example, an animal that qualifies as possible prey for a predator disseminates odorant molecules of which some happen to interact with the olfactory sense of the predator. The predator then takes action, which is triggered by the reception of the signal, but causally explained by the relevant matter of fact, i.e. the presence of the prey. Finally, the result of the immediate action is a further event or state that is of some utility for the agent. The predator catches its prey and consumes it.

The paradigmatic models for actions, the operations of Maxwell's demon on the trapdoor, can now also serve as a test case for causation by information, and likewise as its paradigmatic case. All that needs to be done is the conceptual mapping to the aforementioned scheme. In Maxwell's setup, the contingent situation is the initial condition of the canister, which is in thermal equilibrium. The demon has a measurement device at its disposal, which serves as the signal correlated with the relevant fact, the thermal value and the movement's direction of a particle approaching the opening between the halves of the canister. The demon commits an appropriate immediate action, i.e. closing or opening the trapdoor. After a while, consecutively engaging in this activity, the final beneficial state is established, if we add to the standard story the assumption that the agent can make use of the free energy contained in the temperature difference.

The part crucial to explaining causation by information is the rule that determines the operations on the trapdoor. In order to define the target state, we assume that a turbine, which can convert the free energy of the canister into work, is installed in such a way that it is required that the fast (hot) particles are in the left side of the canister. The assumption concerning the direction of the turbine's installation is, of course, one that does not entail a loss of generality. Since it is a contingent matter of fact whether a cool or hot particle is approaching the trapdoor, an agent (or mechanism, for that matter) needs to be informed of this fact via a physical channel. But although for any specific implementation of Maxwell's agent there will be a rule from a channel state to the best action, this rule is explanatorily irrelevant.

For example, the register of the measurement device might show state 'zero' for the result of the last measurement, which, by virtue of its wiring, might be correlated with 'fast particle approaching from right side / slow particle approaching from left side'. Then the rule the demon abides by will be: 'If register state = zero, then open trapdoor'. But the wiring up could be according to exactly the reverse mapping: 'If register state = one, then open trapdoor', with a corresponding complementary rule. The rule from the measurement result to the action is not explanatorily relevant for why the beneficial outcome is successfully brought about, since the best action is determined by the relevant matter of fact, not by how it is internally (within the agent boundary) or externally (as part of the information channel) represented. In other words the mapping from the

signal to the action is irrelevant as long as the rule 'If fast particle approach from right side / slow particle approaching from left side, then open trapdoor' is implemented.⁴⁶

A *particular* example of the demon will only yield an explanation according to the scheme of efficient causation. This can be verified in a simple Markovian model like a Bayesian network with conditional probabilities. The demon opens the trapdoor within a significant time interval because his register shows state 'one'. The register shows state 'one' because a fast particle is approaching from the right side. By transitivity, the demon opens the trapdoor because a fast particle is approaching from the right side. The latter is the distal cause of the action, but not in any highlighted sense as causation by information requires, i.e. in a sense of *back-referencing* at the place where the decision is taken. To explain the referentiality of causation by information, one has to look beyond a particular implementation of a demon and look at how the wiring between system and measurement device, and between measurement device and trapdoor, have been fixed. One possible way to do this is to think of a model that works by a randomized initial wiring and a subsequent process of simulated natural selection in a population of demons. Any demon whose wiring does not implement the relevant rule 'If fast particle approach from right side / slow particle approaching from left side, then open trapdoor', will not benefit from the resulting state of disposable free energy and will consequently be purged from the considered population. How the *internal* wiring around the register of the measurement device will be carried out is irrelevant. If we now consider a set of Bayesian networks that represent models of a set of demons that took part in the outlined game of evolution, the rules revolving around the register will average out, if we add the plausible assumption that there are no further symmetry-breakers that favour one variant of internal wiring over the other. Then the only significant conditional probability remaining will be the rule relevant to the agent's 'survival', or, equivalently, the explanatorily relevant rule, which is the rule that relates the particle state to the intervention, by-passing the register state. The phenomenon of referencing back in time is explained by changing the perspective from modelling the individual to modelling a set of individuals, or from modelling a situation to modelling a type of situation. Under this new perspective the register state drops out as causally irrelevant. In contrast to the previous perspective, that looked at a single agent, the process will appear informational rather than mechanistic: the register-state is merely an indicator for the relevant particle state and therefore only an exemplified implementation of a range of possible informational pathways, and we know that it could have also been the inverse register state that could carry that information. That is how we could best explain the phenomenon of back-referencing from a third-person perspective.⁴⁷

The account of causation by information is meant to explain the judgment of discerning caused actions and coerced effects, even if the actions are brought about *regularly*. Since the concept of efficient causation is used to explain the latter of the two types of causation, we need an account of how the two concepts of action are related to each other. One of the two concepts of action concerns the notion of 'bringing about' in the agency-formula, which connects the 'free agent' and the immediate object of its action. The second notion of action concerns action seen as

⁴⁶ Notice that this is the explanatorily relevant rule relative to the way the turbine is installed.

⁴⁷ An alternative to averaging over the members of a population of demons, the same consideration can be made for a single demon with an adaptive internal wiring, if we give the demon the chance of running the experiment several times and then decide on an internal wiring.

effect in causation by information. This difference between the two concepts, in terms of the primary intension, has been stressed in the example of the contrastive pairs concerning the causal chain of cue and two billiard balls on the one side, and the amoebas (see section 6.1) on the other side. The next sections try to tackle the obvious question whether the two kinds of actions might be extensionally identical while remaining distinct on the grounds of the subjective-objective divide, viz. they are two aspects of the same thing in the sense of the double-aspect theory. The operation of setting a value of a variable in models of causation based on random variables (e.g. in Objective Interventionism) would have found its semantics by the subjective notion of immediate action. Whether an observed effect is an action or a coerced effect depends on whether we find a structure that has the capacities of information processing, which implies having been selected or designed such that on the basis of interacting with a *B* a rule connecting a prior *A* (rather than *B* itself) to an immediate action *D*, such that a final state *E* follows, is implemented by the structure, potentially in a regular fashion. Since the first-person observer is itself such a structure, whose judgemental capacities have been pruned by these relations, we would have an argument for the assumption that what underlies our judgments concerning causes is indeed a judgment concerning actions.

I think of that project as another instantiation of an approach which Price calls ‘naturalized Kantianism about causation’ (Price (2007), p. 255).⁴⁸ Although Price’s account is reductive in that he tries to reduce both the temporal asymmetry of causation (causes precede their effects) and the means-end-asymmetry (causes are means to their effects, but not vice-versa) to a causal perspectivalism, whereas my approach relies on the direction of time as a (non-causally) given, I still completely concur with Price’s interpretation of perspectivalism and its relation to science: ‘[T]here is a tendency to think that perspectivity is incompatible with good science, in the sense that science always aims for the perspective-free standpoint, the view from nowhere. In my view, it is important to see that science itself might challenge this philosophical conception of science.’ Price (2007), p 253)

My strategy for making the case for an identity of the two actions will be as follows. First I look at the constraints on acquiring the concept of agency, subsequently I will address the question of how the main obstacle, the aspect of spontaneity of an action, can be couched into a natural account of concept acquisition (something that has been done only in a limited sense when judging actions from the third-person perspective), and finally, we must show that the account of action arising in that context fits the main purpose to which we put it, i.e. making sense of the difference between correlation and causation. To anticipate the result, I think that the point that causal judgments are necessarily grounded in judgments about action cannot be made decisively, but it can be shown that it is a plausible possibility.

6.2.7 Concept acquisition

For biological agents, acting is necessary. In order not to deteriorate, the agent needs to get hold of tokens of free energy. But in order to do that, it must build a structure that is able to process

⁴⁸ What comes closest to my understanding of the conceptual problem of causation (the explanation of a causal judgment in non-causal terms) in Kant (1787) is his exposition concerning the ‘schemata of the pure conceptions of the understanding’, where he writes: ‘The schema of cause and of the causality of a thing is the real which, when posited, is always followed by something else. It consists, therefore, in the succession of the manifold, in so far as that succession is subjected to a rule.’ This mirrors Menzies’ and Price’s account, except that they allow for a probabilistic reading of the causal rule.

information, because unconstrained actions are not adequate for getting hold of a token of free energy. But this sub-structure of the agent is again subject to deterioration, so its existence depends on this very token of free energy, such that the anti-entropic process of maintaining the unlikely, functional structure can be performed. This mutual interdependence clearly defines a structure for which a criterion of identity *through time* can be applied. At the time when the contingency of a situation has been fixed, the potential best outcome is fixed for the agent in that situation, and the addressee of the eventual flow of feedback *is the same agent*.⁴⁹ One can call this phenomenon (which efficient causation that does not count as causation by information lacks) ‘centralisation’ (echoing Bertalanffy’s notion in chapter 3 of Bertalanffy (1969)), in the sense that the information processing structure and the effector, e.g. a limb, serve the purpose of the same entity, which is necessary to close the feedback cycle on which the interdependence turns. That is what a thermodynamic-cybernetic picture of acting has so far shown us.

Although this feedback cycle applies to biological organisms, it is plausible that any other structure capable of formulating causal judgments that is not constructed and maintained by an intelligent designer, and therefore an extrinsic cause, is subject to a similar cycle, since the capacity of formulating causal judgments requires a degree of organisation that makes the existence of such a structure very unlikely. For structures of such a kind, acting is therefore necessary, and the necessity of acting is mirrored by the evolutionary pressure to acquire a *concept* of acting, in the above limited sense of being able to exert direct control on an object. Such a concept is needed to model situations asynchronously, when the agent is not actually present in the respective situations. Explicit planning as well as communicating the plan to others, also anticipating the course of action of rival players from the environment, are all capacities unlocked by an explicit concept of action.

The paradigmatic model of action, consisting of a measurement by coupling an internal to an external structure (the measurement, or perception), and a computation that feeds into committing to an action, also affords the explanation of how higher-order goals can arise from the basic thermodynamic necessities of harnessing free energy that serves to maintain that very process. In the paradigmatic model of Maxwell’s demon, the measurement process is an operation that reduces the informational entropy of the agent, which can be defined as the degree of uncertainty of the agent relative to the contingent situation outside of its boundaries. This is a necessary *prerequisite* of performing an adequate action, whose final result is – in the paradigmatic case – the reduction of physical entropy of the agent by re-establishing its functional state. But in many examples of causation, reduction of physical entropy is not palpably an agent’s goal, as the example of the next paragraph will show. However, a large class of additional scenarios can be interpreted as higher-order goals ultimately grounded in that low-level goal, such as when a process of informational entropy reduction serves the goals of enabling the agent to perform a more immediately relevant informational entropy reduction. To illustrate this case, an agent might want to screen the trustworthiness of an informant before deciding to question her about the actual matter of fact at stake. Likewise concerning these higher-order problems, we find that in some contexts the question

⁴⁹ For example, if I sit in my office and someone calls my name from the hallway, I understand that I am the addressee of the other person’s calling. I initiate a whole-body reaction of getting up and leaving my room, in expectance of another subsequent situation that concerns me as a person. Clearly, a single entity is involved across these interactions. Contrast this with a situation where my body is coerced to move, due to an extrinsic force. Here, my body is not addressed indirectly, and its reaction is not triggered via some notion of personal identity. Instead, it is addressed as a physical object whose identity over time is irrelevant for the causal transaction in question.

whether an immediate redemption of an informational asset is the most rational option is less straightforward to answer.

For example, a commander of an army might hesitate to issue an attack command and instead decide to collect some more pieces of information about the possible weaknesses of the enemy's army. Similarly, an academic might feel he has to read yet another paper on a subject before he feels confident enough to elaborate his own ideas. The hesitation can become pathological if there is no end to the procrastination. Collecting more and more intelligence about a task to be performed can lead to pathological information greed, similar to the greed of acquiring more and more money by continually reinvesting one's returns, in order to be protected from future contingencies, rather than committing to its consumption at some time.

Obviously, describing these constraints and affordances of concept acquisition is far from delivering an account of how it actually happens, not to mention a reductive one. A difficulty seems to lie in the fact that I need the concepts of (logical) dependence and temporal order, which one might take as intimately connected to causation, so that a non-question-begging account cannot just presuppose them. Another problem seems to be that the conceptual framework established so far is insufficient to make a clear case for the priority of the acquisition of the concept of agency over that of efficient causation. The computational model outlined in section 6.2.5 highlights a specific distal cause, and the agent acts in the correspondingly constrained situation such that the target state is brought about, but it is not easy to identify in this model a criterion that enables the agent to distinguish an efficient distal cause from a cause that works via informationally constraining the agent. Tantamount to this is the agent's problem of recognizing its action *as* an action rather than an extrinsically coerced behaviour. One of the possible solutions, the assumption that the agent can vary its behaviour in a similarly constrained situation, just passes the problem to the question of how to distinguish a degree of freedom in its own action and variation on both the sides of cause and effect in probabilistic efficient causation. However, this problem might be due to an overcautious and inappropriate alike treatment of actions and events. There is a long tradition of theorising over a direct epistemic access of an agent to its own actions (see Bayne (2011) for a discussion), but it is difficult to say whether application of this epistemic power requires the prior grasp of the *concept* of action according to such accounts. But in the thermodynamics of actions there is a further source of asymmetry between actions and merely observed events, including extrinsically coerced movement of one's own body. It concerns the distinction of spontaneous and non-spontaneous events, which is potentially reducible to statistical, rather than causal, notions. If I observe myself moving in accord with an external force, my movement will appear as a spontaneous process, while moving against the external force must be a non-spontaneous process – interpretable as an action.⁵⁰ This account, however wanting, seems still more adequate than an account of concept acquisition based on ostension, as in Menzies and Price (1993).

⁵⁰ Unfortunately, the spontaneity of actions in the action-theoretic sense and the spontaneity of processes in the thermodynamic sense seem to work in diametrically different directions in these cases. Moving uphill, if only the body and the environment are observed, is *not* a spontaneous process in the thermodynamic sense. But if the internal processes within the body could be observed, if we expand the perspective, then the process of moving uphill evolves spontaneously. The concept acquisition of action depends on the fact that I do not observe my own internal processes, therefore I am not able to apply the scheme of efficient causation to myself, which would yield thermodynamically spontaneous processes everywhere (which becomes interpretable as the nexus of causes and effects, if we draw the boundaries accordingly). The thermodynamically *non-spontaneous* processes thus appear as *spontaneous* actions.

At any rate, I think that a proper account of concept acquisition, of which it would be questionable anyway whether it could be plausible in the context of a philosophical story-telling, does not have to be delivered to comply with the current task, which is to show that the acquisition of the concept of causation – in whichever way it exactly happens – can be expected to arise in the thermodynamic scenario outlined in the previous sections. The decisive order of logical dependence lies in the concept application in a judgment, and the acquisition of the concept does not necessarily have to reflect this order.

6.2.8 Free action

The computational aspect of the paradigmatic model of 6.2.4 concerns the implementation of the rule mapping the causally relevant fact to the action that brings about the desired target state. Of some relevance to the question of concept acquisition is the representation of the result of the computation, since this is what determines which of the possible actions will be committed to. In this context, it is a reasonable assumption that, given that the outcome of the computation is what determines the agent's decision to act in a certain way, that outcome cannot be again represented in an internal model of the agent. If the result of the computation were an internal observable, then something would have to be done after the observation, since the observation itself would not be the agent's decision. Rather than that, the agent would need *another* rule connecting the observation with the decision to act, something similar to reasoning: 'If the result of the computation of the best possible way of acting is observed, then obey that result and commit to the corresponding action.' But this rule would be tantamount to another mapping, the mapping from a recommendation to the positive or negative decision to abide by the recommendation, and therefore another function would have to be calculated, with the iterated question of what happens if the result of computing *that* function is observed. Therefore, the immediate trigger of the decision to act cannot be an internal observable of the agent's model.

Next to the external, causally relevant factor as such, the computational model also allows taking weight factors of decision-relevant criteria, and also motives of higher order, into account. For example, if someone in a pub suddenly punches me in the face I might immediately strike back out of reflex. Judging with hindsight, this would still leave the possibility open to construe my doings as an action based on a triggering factor stemming from my environment. But I might also take some time to think through my situation, pondering the pros and cons of what to do, and could then still decide to take revenge. Although the cause of my action is still the aggressor's previous action, which has this time gone through a complex intermediate process of computation and whose primary cause has thereby been enriched by further motives of second-order, one would also judge that my action, in this more complex second scenario, has acquired more of a quality of a 'free action'.

An extreme case is the situation that prompts the agent to perform causal experiments concerning a possible connection between a putative cause and its effect, like a switch and a light bulb. In performing an individual action, the agent is not directly constrained, and the immediate trigger, which makes the agent 'arbitrarily' toggle the switch, does not appear as a causal predecessor in the agent's model, when the agent recollects the episode of having operated the switch and formulates a causal story to accommodate that recollection. This account of the opacity of immediate triggers in unconstrained situations leaves the thermodynamic model of agency intact, and it accommodates the free action as its extreme case. With that, a case has been made

concerning the identification of thermodynamic action and the free action of the epistemic agent that judges causation according to the agency-formula.

6.2.9 Concept identification

At this point it might be worth recapitulating the stage of the argument once more. In examples that involve effects triggered by information rather than efficient causes, an external observer classifies the effect as something that resembles an action by another agent. At that stage we have assumed that the external observer possesses all concepts required to make the conceptual distinction. In order to describe causation by information more systematically, Maxwell's demon has been found as a paradigmatic case. But the thermodynamic scenario that embeds the demon is one the external observer is also subject to, and it is a scenario that makes it likely for a concept of action to arise according to the paradigmatic model derived from the demon's actions. It is therefore a plausible assumption that the first-person concept of action used in the agency-formula uses the same kind of model that must be imposed on the agent whose action is judged in cases of causation by information. In contrast to the judgment concerning causation by information, when the epistemic agent observes another causal agent, the 'free action' in the agency-formula is an action hypothetically performed by the epistemic agent prompted to utter a causal judgment. The causal agent is thus identical with the epistemic agent.

The main plausibility-problem of that story is the spontaneity of the agency-formula's 'free action' that arises from the nexus of causes and effects, but for resolving this puzzle there is now a solution given by the account of the observables of the internal computational model (section 6.2.5). The phenomena of spontaneity and back-referencing, which are directly given by the first-person perspective, can be explained alternatively by a third-person perspective. Still, there is a further obstacle to reach the next stage of the argument. Although the concept of action involving a free agent can be embedded into a framework in which actions are coerced, the question of logical dependence between actions and causes has been left open, viz. we have still the open question whether the agency-formula makes the right assertion of explaining *all* causal judgments as agency-judgments. But we have not only embedded actions into a framework within which they appear as coerced, but we have also allowed for an alternative view within the same framework where coerced actions are not different from other kinds of coerced effects.

The pivotal step in the argument can also be represented as such:

Agent A (the observer) has been considered at a time when it already has all concepts at its disposal, in particular CAUSE and ACTION, leaving open their dependence-relationship. Agent A observes Agent B (the acting agent) and judges the latter's behaviour as action. Agent A sets up a model of how, from within a closed nexus of causes and effects, a causally relevant fact can become efficacious through information processing in Agent B. This model entails the opacity of the immediate causal trigger for Agent B's doings. Therefore, the action taken by Agent B appears to be free from B's perspective. It follows that the agency-formula can be employed by B, so that B now has a means of deriving efficient causation from action. But Agent A's concept application, on which the explanatory model for B's behaviour depends, is constrained in the same way as B's concept application. So B could apply the same reasoning judging A's behaviour, so that in A's conceptual

scheme, efficient causation is derived from action, independently from what ontologically underlies efficient causation.

Thus the efficient causation that might be a necessary ingredient to conceptualise non-spontaneous processes depends on the agent's concept of agency. But what the application of the concept in the end effects is an account of the origin of its concept. It can be debated what follows from this conceptual loop (rather than from a flat circle) – whether the concept refers to an illusion or is something rather substantial, but at least it follows that it is not inconsistent.

In the model of Maxwell's demon there is a difference between considering populations of demons, which yield the explanation of back-referencing of causation by information, and considering an individual demon, for which the aspect of being an action is explained away since its doing equates the workings of a mechanism in which every part of the chain of causes and effects is determined by its immediate predecessor. Applying the agency-formula from the subjective, first-person perspective consecutively to a chain of events $a \rightarrow b \rightarrow c \rightarrow \dots$ etc., yields that the same kinds of events can be causes and effects, by virtue of extensional identity. That much had been clear before, but now that the concept of action has been supplemented by its natural constraints, we have finally identified a further source of information of how to characterize an action, such that the agency-formula might serve further purposes besides evaluating binary causal claims. These further characteristics are the thermodynamic properties of events that are embedded in relations of entropic order, such that an event of higher entropic order causes an event of lower order. Actions inherit these properties from thermodynamic events, and in turn causes inherit these properties from actions. The most conspicuous of these properties are locality, asymmetry, and regularity (see next section).

The work done by the preceding sections is almost sufficient to answer the objection of anthropomorphism raised against agency-theories. First, my suggestion of a limited interpretation of the agency-approach as a theory that addresses only the conceptual, not the ontological aspect of causation, makes the difficulties of this task easier, since I thus do not have to find an ontologically committed account of agency, which then allows a kind of extension to cases that do not seem to involve genuine agency. In addition to that, the thermodynamic scheme, which explains where the immediacy and spontaneity of actions come from, seems to deliver a sensible semantics of the setting of the value of a variable – a concept the Objective Interventionists require – while the connection between the observables A and B can be explained as a probabilistic dependence.

This philosophical position embraces the duality of perspectives and does not see any further use in analysing the experience of agency as far as the *first-person perspective* is concerned. The case is similar to what Chalmers (1996) says with respect to conscious experience in general:

'Indeed, as far as central processing is concerned, it simply finds itself in a location in this space. The system is able to make distinctions, and it knows it is able to make distinctions, but it has no idea how it does it. We would expect after a while that it could come to label the various locations it is thrown into—"red," "green," and the like—and that it would be able to know just which state it is in at a given time. But when asked just how it knows, there is nothing it can say, over and above "I just know, directly." If one asks it, "What is the difference between these states?" it has no answer to give beyond "They're just different," or "This is one of those," or "This one is red, and that one is green." When pressed as to what that means, the system has nothing left to say but "They're just different, qualitatively." What else could it say?'

In that regard the analogy between experience of agency and experience of colour seems correct, although, as I have made clear in section 6.2.7, I am sceptical whether this analogy covers the acquisition of the concept as well, unless it is underpinned by a distinction between spontaneous and non-spontaneous processes.

6.3 The properties of causation: asymmetry, locality, regularity

In the preceding section a case was made for the claim that an agent can reasonably explain a causal judgment on the basis of the concept of action. An effect *B* can be seen as correlated with *A*, and an observed correlation between *A* and *B* continues to hold when *A* has been brought about by means of a direct action, which would render two merely spuriously correlated events *A* and *B* independent. A problem of the approach of grounding the concept of cause on manipulation is the question of how to proceed from there to make more informative statements about causation. Mellor clearly states the problem thus:

Causation's means-end connotation is even more basic than its evidential and explanatory connotations, being to my mind the very core of the concept: causation is essentially the feature of the world that gives ends means. [...] This may not however tell us much about causation. For to bring about a means in order to bring about an end is just to cause the means in order to cause the end. This makes it look as if we need to invoke causation to say what it is to be a means to an end. If we did, the means-end connotation would be as useless as the connotations [...] that effects are 'produced by' or 'derived from' their causes, expressions whose meaning here obviously derives from that of 'caused by'. (Mellor (1995), p. 80)

Spohn (2001), on p. 8, makes a very similar remark. What I find particularly problematic in this context is that we learn little more about the nature of the *relata* by merely acknowledging that they stand in a means-end-relationship. A second problem is that it seems difficult to say more about specific causal structures involving more than the action and its effect, so that from a set of binary relations a more complex model, e.g. a causal network with multiple nodes, could be constructed.

Theories that start from or at least acknowledge a relation of manipulability, wherever we see a causal relation, incorporate the insight that there is importance of that notion to causation in a different way. Not many philosophers develop a theory of causation completely *from* agency and its perspectivism (Huw Price is one of the few), but some at least make some effort of integrating the phenomenon of agency into a wider, coherent theory. In order to present my own approach to the question of how to progress from here in contrast to how other authors proceed regarding the question, I will present a little survey of treating this question.

Of course, for some philosophers, like Wolfgang Spohn, there is no deeper issue to the question. While acknowledging that agency is a particular instance of causation, many writers seem to content themselves with believing that agency *uses* causation, but that this is all there needs to be said about it. I think this cursory way of dealing with that question is a mistake; not only because by looking at agency we can learn more about causation, but also because we can learn more about the relation between the first- and third-person perspective in general.

Mellor seems to be taking agency seriously. His quote from Mellor (1995) suggests that his analysis of causation in terms of a means-end-relationship is a reductive one, similar to Price's, given the fact that he sees this relation at the 'very core of the concept' and its 'essential feature'. In contrast to 'production' and 'derivation', causes seen as means to ends do not depend on causation in a derived way. Rather than that, causes and means to an end are in some sense notions on a par with each other. But it seems that the agency inherent in the means-end-relation has no *logical precedence* over causation in his account, which is already indicated by his treating agency as one of the 'connotations' of causation. In his discussion of the means-end-relation, he outlines his version of expected utility maximisation, but in doing that is careful to distinguish expected valuation, with subjective credence and subjective utility, on the one hand, and mean utility, with objective chance

and objective utility, on the other hand. Mellor's concern is a possible outranking of expected utility by the dominance principle, which would give the wrong result in standard (Non-Newcombian) problems of decision theory. Mellor's example is a patient facing the question whether to take medicine or not. The dominance principle prescribes not taking it, since no matter whether recovery takes place or not, the patient would be worse off conjoining this event with the intake of medicine, which would be unpleasant and incur a further cost. The application of the dominance principle is easily rejected in these kinds of problems, since both the evaluation of the expected utility according to evidential probabilities (i.e. looking at all the possible combinations of the probabilities and utilities) and the evaluation according common-sense causation (the intake of the medicine influences the likelihood of the recovery) accord in favour of taking the medicine. Mellor does not discuss Newcomb's problem directly in this context, but his idea of distinguishing expected valuation from mean utility clearly serves the purpose of providing a criterion for being able to tell when evidential probabilities should *not* determine an agent's decision when the dominance principle recommends the opposite. In Newcomb's problem, two-boxing is the rational choice, in accordance with dominance, if an independent account of the causal connections tells the agent that his decision will not influence the past, even if the rules of the game contradict this causal fact. But if the cause's role as a chance-raiser is all an agent has to go by in order to judge a causal connection, her choice in Newcomb's problem, since it makes a difference to the expected utility, will classify the chance-raising action as cause. Since Mellor has, by virtue of his philosophy of time, an independent account of when an event objectively raises the chance of another event, he can discharge this conceptual confusion by attributing the error to the agent's failure to adopt the 'real chance' as his most rational choice of credence, rather than the evidential probability. An evaluation of Newcomb's problem according to the real chances, unlike the evidential probabilities, is in accordance with the dominance principle, since causes must precede their effects, according to Mellor. It follows that Mellor's account is not a reductive analysis, and accordingly, his theory of causation is not synthetically built up from the notion of the means-end-relationship.

Price's philosophy of causation (Price (1992), Price and Menzies (1993), Price (2007), Price and Weslake (2009)) is a consequential attempt to explain causation from agency, and is meant to be a reductive analysis of causation. As a consequence, Price abides by the recommendations of evidential rather than causal decision theory, even in Newcomb's-problem-like scenarios, i.e. scenarios where the two variants of decision theory drift apart. In a plausible interpretation of evidential decision theory in this context, the logic of this choice leads to backwards causation, and that in turn opens an interesting but hard to oversee conglomerate of questions concerning the direction of time, of causation, and of thermodynamic state transitions. Price embraces the possibility that causation allows for an atemporal interpretation, with 'effects of its interventions showing up in various directions, throughout the manifold' (Price (2007), p. 282), given that there might be creatures with a sufficiently all-encompassing perspective on space-time. The conceptual distinction between 'options', 'knowables' and 'fixtures' (in *ibid.*) in an agent's deliberation allows to represent structure beyond the binary case of the cause-effect-relation. Price also seems to endorse the formalism of the objectivist branch of interventionism, which offers a calculus for these multi-variable structures. However, he disagrees with their notion of intervention, which he calls a 'Trojan Horse against objectivist approaches' (*ibid.*). Unlike my interpretation of agency, Price is interested primarily in the metaphysical consequences of a perspectivalist interpretation of manipulations. One of the aforementioned, hard to oversee consequences of negating a reality of directed time independent of causation is the correct interpretation of Price's distinction between 'options' and

'knowables'. If something is an option, then the agent can either realise the option or refrain from realising it. Unlike a 'knowable', the agent cannot tell whether the option will get realised or not *before* it will actually have been realised. Since causation hinges on deliberation, and deliberation hinges on the distinction between option and knowable, even for a godlike agent it is required to introduce an additional personal, asymmetric time. The consequential next step indicated by Price is the denial of causation in an objective sense, if this minimal constraint on a perspective in any sense is lifted.

Judea Pearl is an ardent critic of evidential decision theory (see Pearl (2000), chapter 6). His theory of causation, belonging to the interventionist branch, is not a reductive theory. In particular, he does not reduce causal knowledge to evidential probabilities (next to Pearl (2000), see also Pearl (2001) on this point). Although his view is that causation can be defined by means of the 'set-operation' (see section 4.2.5.1), which is a formalisation of his notion of intervention, this mathematical operation has to be seen in context with evidential probabilities underpinned by a causal graph. In this graph, all arrows converging into a single node are always interpreted as *mechanisms*. It is unclear in his exposition in (Pearl (2000)) where causal background knowledge, which seems to provide the basic causal structure in a lot of considered cases, originally comes from, and what might bootstrap the applicability of his method.

The question whether evidential probabilities and the notion of manipulation is sufficient for delivering an account of causation, which would qualify as a reductive account, arises for all other theorists concerned with probabilistic theories of causation, too, given that the importance of the notion of manipulation to underpin the probabilistic approach is acknowledged. This is usually the case, since no contemporary theory claims that causation can be derived from probabilities alone. The question of reduction is therefore focused on the relation between manipulation and causation. Cartwright (2007) holds that the methods of causal inference and the devising of effective causal strategies have to be seen in conjunction with the right causal metaphysics, so her theory is not reductive in that respect, either. Among the other theorists, besides Judea Pearl, working with Bayesian networks, Spirtes et al. are agnostic about the metaphysics of causation. They are also not interested in providing a definition of causation, or conceptualising it in a sense that goes beyond the constraints needed for causal inference. So for them the question is left open. Wolfgang Spohn's theory might be considered reductive. He thinks Bayesian networks exhaust the idea of *causal dependence*, but his theory of causation has to be considered in conjunction with his theory of ranking functions, which formalise our notion of belief. If his theory can be considered reductive, it will be reductive with respect to ranking functions. In any case, there is, explicitly, no reduction to or development from the notion of action or intervention, although, interestingly, he treats action variables as conceptually distinct from observational variables (Spohn (2012)), similar to Price's distinction between 'knowables' and 'options'. Williamson's theory in (Williamson (2005)) interprets causal Bayesian networks as a product of observation and prior causal beliefs of an agent. The question of reduction in this theory appears to be tied to the quest into where the prior beliefs of an agent stem from. However, since Williamson has also focused on investigations into causal mechanisms, it is unlikely that his theory is reductive; at least it will not be considered a reduction to probabilities and interventions.

With respect to reduction of causation to intervention, Woodward's theory is an interesting case. He explicitly states that his theory is not 'reductive' as he understands the words, since the notion of an intervention is itself a causal notion rather than an independent primitive notion (Woodward (2003b), p. 27). He goes on saying that seeking a reductive account leads one into a

subjectivistic or anthropomorphic conceptualisation of manipulation. In contrast to such projects, exemplified by Price's theory of causation, his project attempts to 'elucidate the concept of causation by tracing its interconnections with or locating it in a "circle" of interrelated concepts' (ibid.). Accordingly, Woodward defines interventions in terms of causation (see again section 4.2.5.1), while causation is defined in terms of manipulability (Woodward (2003b), p. 45). Like Spirtes et al. and like Pearl, Woodward espouses the causal Markov condition, but unlike them regards this condition as derived from manipulability. Thus he doesn't presuppose this condition as the former authors do, but sees it as a corollary of his notion of manipulability.

In the remainder of this section I will outline my own answer to the question of how to get a more informative account of causation based on the concept of agency. The result of the preceding section has yielded the extensional, ternary identity of subjectively free actions, triggered actions and coerced events, which can be effects or causes, depending on the context. Since the case for the priority of actions in an agent's judgment over causes has been made, I can now say that causes inherit the properties of the natural actions. After the properties will have been assessed in this section, it will be the subject of much of the remaining thesis to draw conclusions from them to the way we form causal judgments and to the standards of validity we assign to causal models. The properties in question correspond to what is often referred to as 'connotations' (Mellor (1995), or 'platitudes' (Menzies (1996)) about causation.

6.3.1 Asymmetry

To start with the most obvious connotation, the causal relation is most often considered asymmetrical. Intuitively, a series of interventions renders two observables not related as cause and effect statistically *independent*, even if a correlation has been observed before. This seems to hold true without additional assumptions, which has already been stated as the virtue of the manipulationist approach. An approach that explicitly addresses the conceptual level of the problem can content itself with accepting the asymmetry as a given fact of the agent's first-person judgment. The usual approach of conceptual analysis (Margolis (2014)) matches cases intuitively judged against the result of applying the theory to be tested.⁵¹ My approach analyses causal intuition merely up to a certain point, from which agency-intuition takes over. I have subsequently analysed the origin of the agency-intuition, where the thermodynamics of spontaneous processes plays a role. It seems that the problem of asymmetry of time, thermodynamics and causation arises after all, but only via the problem of grounding thermodynamics, which this theory brackets. For the agent that sees the necessity to intervene because some process would not happen without its doing, the question where *that* intuition comes from must be left open given the limited means of the analysis used in the section covering the thermodynamics of causation.

It is therefore quite clear that my approach is guilty of what Price (1992) calls 'conceptual buck-passing'. It explains the asymmetry of causation by appealing to the asymmetry of agency whose orientation is as problematic as that of causation itself. But it is of a benign kind and thus one of those instances of conceptual buck-passing which, according to Price, can allow rendering problems more tractable than before. This is in fact how I think about causation itself: if it can be cashed out in terms of agency, as the objectified version of the latter concept, then causation might

⁵¹ Obviously, the literature on causality and causation makes extensive use of that approach, no matter whether the theory is explicitly categorised as a piece of conceptual analysis or just proceeds in such a fashion that in the end it looks like conceptual analysis.

allow us to address the heavier metaphysical questions of time, rise of entropy, and regular patterns *within* the rise of entropy, in a more tractable way.

Asymmetry is a property of causation that presupposes that the agent can tell whether two events *A* and *B* are distinct. Otherwise there is no sense in speaking of bringing about *A* such that *B* follows, *in contrast* to doing *B* such that *A* follows. But this is a prerequisite of applying the agency-formula in the first place. There are certain setups that defy the idea of asymmetry. For example, one can think of a pair of buckets and a hose that connects the two buckets through a hole in their respective bottom section, such that if one fills one bucket, one also fills the other. An agent can target each of them as the immediate object of the action irrespective of the same result. But a symmetry-breaker would still be given by the fact that there is a dependence of the indirect event on the direct action, in the sense that, in the eye of the judging agent, the indirect event would not happen without the action. In practical contexts, the reason for believing in the dependence will often be a mechanism, but can also consist in a blunt belief of dependence no matter what the further circumstances are. However, the belief should imply at least a notion of dependence that stands in contrast to mere *logical* dependence. Pearl's often used example of 'making the grass wet causes the grass to be slippery' (in Pearl (2000) and elsewhere) is a very questionable example to make his readers acquainted with his invention of causal Bayesian networks, since the application of water brings about *both* changes in the properties of the grass with the same action, and the asymmetry is merely a logical one. Supposedly, applications of all kinds of liquids, including water, make the grass not only wet but also slippery, whereas slipperiness can also be brought about by other means. Therefore, there is an asymmetry in reasoning from one proposition to the other; but one can have doubts about whether this involves a genuine *causal* asymmetry. In the same questionable sense, one might want say that drawing a red ball from a ballot box 'causes' the drawing of a coloured ball, whereas the converse does not hold true. Notice that the difference between causal and explanatory reasoning might not be possible if one does not have a theory of the relata of causation as physical processes, according to a view similar to the one I have developed in this chapter.

6.3.2 Locality

Another point often observed but differently formulated is what I want to call 'locality of causation'. It is contrasted with Laplacean universal determinism, according to which a state of the universe determines the state of the universe that ensues temporally, rather than entailing that a local cause influences events in its surroundings. Law-based approaches to causation (e.g. David Armstrong) on the one hand, and approaches that treat objects as bearers of local capacities or dispositions (e.g. Steven Mumford, Nancy Cartwright) on the other hand, also reflect this dichotomy.

In the conceptual context, again in decisive contrast to the ontological context, the locality is, like asymmetry, a relatively uncontroversial property of causation, and this holds true for agency-approaches in particular.⁵² Locality captures the idea that the agent has a limited range of influence, and that the effect's occurrence is due to what has happened in that limited range. Whereas an explanation from a cosmic perspective of how things in the world unfold in time might favour a law-based account that does not attribute causal power to individual things, the agency-stance prohibits

⁵² This cannot be said of a more metaphysically committed reading of the agency-approach. For example, Hausman (1997), in his critique of Menzies and Price's agency-theory, represents this theory as maintaining that causation is a relation *extrinsic* to the primary aspects of the involved events, in virtue of having a 'secondary quality'. This reading conflicts with locality seen as an *intrinsic* quality of causes.

the assumption that actions are instantiations of a law, since it requires them to follow from a free decision to act, even if this view eventually turned out to be an illusion. Leibnizian pre-established harmony is a metaphysical theory of causation that takes into account the manifest but false judgment of an agent that its action can bring about a physical effect. Another instance of falsely attributed locality of causation is when we see a film or a computer simulation, where there are no objects that make a local difference to what we see in the ensuing frame. But when I act in reality, my decision to intervene results from my belief that my intervention is called for, since if I refrain from doing what is in my local sphere of influence, the intended event would not occur.

Locality is a crucial feature of causation according to the way I have construed the concept, which is why I juxtapose it next to asymmetry and regularity and assign it equal importance. Going back to the spontaneous, non-spontaneous distinction, there was a worry of an underlying, hidden circularity, if we ground the concept of agency on that dichotomy and then derive causation from agency. But there are clearly processes that do not require a causal explanation. For example, if water evaporates from a glass into its surroundings, or when a temperature gradient of gas in a container gradually levels out. The reverse of this process is non-spontaneous, but as the second law of thermodynamics shows, if such processes are observed there are other processes elsewhere that restore the balance of entropy. So, looking for causation requires the partitioning of space, and therefore causes are necessarily local. Importantly, if this outlook is sound, there is a prospect of reducing causation to non-causal concepts, viz. to statistics, with no hidden, underlying circularity involved. Of course, this would only be true if the direction of time can be grounded non-causally. We have relied on the direction of time by implicitly taking for granted that the next most likely state to be assumed by a system at a given time, according to its statistics, is the state that lies in its future, not its past.

6.3.3 Regularity

In the conceptual context, the property of regularity can be defined as follows: whenever we judge that '*a* causes *b*', there is a corresponding '*A* causes *B*'-judgment. The first proposition involves *a* and *b* as event tokens, while the second involves the corresponding types *A* and *B*. A situation involving *a* and *b* might seem non-repeatable, but from the fact that we made a judgment about the causal connection between the two relata it follows that this situation would give rise to a regularity, if instantiated repeatedly. Thus, a judgment of the form '*a* causes *b*' depends on '*A* causes *B*', but not vice versa. Since my approach abstracts from the epistemic question of how the agent learns about a type level causal claim, this is not supposed to preclude a reversed dependence as far as the causal epistemology is concerned.

The argument for this claim goes as follows: The causal claim '*A* causes *B*' hinges on the concept of action. But the application of the concept of action hinges on an event that takes place in the thermodynamic context as outlined in sections 6.2.4 and 6.2.5. An action is always constrained by information about the situation that embeds the agent. According to information and its use in the agent's model, different hypothetical situations are contrasted, which include acting in different ways. Whether the agent commits to an action depends on the result of such a computation. Evaluating the result of its own action, the agent is again dependent on its internal model. It follows that the agent is never in direct contact with the event (the causally relevant fact) that triggers the

action causally. It is rather configured to react to a causal type.⁵³ This scheme applies to causal judgments concerning an observed past, but also concerning the planning of a future situation.

The second argument for a causal judgment mediated by types, rather than pertaining directly to observed tokens, takes into account that the agent is limited in its computational resources to represent the rule from the causally relevant fact, in conjunction with the range of possible direct actions, to the targeted situation. The agent is limited computationally because it is limited physically, and information, according to section 6.2.4, needs to be represented physically. So the causal model, the basis of the computation of the best action, must involve causal types.

The first argument is further corroborated by the fact that the inference from the signal to the matter of fact is itself a rule, and therefore concerns types. Every signal depends on the bandwidth of the information channel via which it is produced, and the event at the sending end of the channel will therefore always appear as an abstract representation of a possibly more concrete specification, in other words it will be given to the agent on type level. The second argument can also be made further plausible by reminding oneself that actions are always informed actions. So far, the necessity to measure contingent information has been highlighted, but evidently this information needs to feed into the model that makes sense of the measured data. But, as the thought experiment of Maxwell's demon shows, information needs to be physically represented throughout a process between an input (the measurement) and an output (an observable action). This puts a limiting constraint on how an agent can produce a causal judgment. Relying on types given by causally relevant properties, which individual events merely instantiate, is the most obvious solution to the problem of how to make maximum use of an agent's computational resources.

The causal types are best interpreted as causally relevant properties, and individual causally relevant causes are seen as instantiations of classes that bear those properties. No assumption concerning the reality of these properties is made (as opposed to a metaphysical reading of these properties, as in Ehring (2009)), since they are just instrumental concepts that inform the agent's internal model. An account based on causally relevant properties can make sense of what the task of causal prediction (of which planning an adequate intervention is a subclass) consists in, and how the agent is enabled to cope with this task. For example, a prediction to be made might concern the most likely ramifications of what will happen if a certain observed object is thrown into a window. Let us assume the possible final states are coarse-grained in such a way that a bivalent result: 'window breaks' / 'does not break', is expected, following a vigorous throw of the object, targeted right at the centre of the window. Then a successful prediction will depend on the right classification of the object in question, and this classification depends on the causal relations the object engages in with the observer, *prior* to the causal interaction in question. E.g., a gold bar, a marble plate, a stone, a billiard ball, etc. are all possible objects that qualify for breaking the glass of the window, whereas a sponge, a paper dart, a snowball, etc. do not qualify for that outcome. The combination of causally relevant properties, in this case the fragility of the glass plate, and the solidity of the stone thrown into it, determines the outcome of the causal interaction, and the class of similar causal interactions determines the causally relevant property of which the individual stone observed in a specific situation is an instance. But if a prediction is to be made, this property has not been observed via the causal effect in question. So the correct classification *prior* to the causal interaction

⁵³ This is true in particular, but not only, when the agent's decision is based on an inference from an observed signal. For example, if a flashing light tells me to engage in security procedures, since this signal indicates super-critical pressure in a vessel, then it is clear that the *type* of super-critical pressure causes my action, because presence of the type is all the information the flashing light can deliver.

must depend on other causal interactions. Given the object is indeed a stone, the agent needs to be informed via channels that attribute colour, shape, size, the circumstance of where the object was found, possibly weight, etc., to the object. Once the classification has been performed, access is granted to information concerning further properties of the 'stone-hood'. Among others, this set includes properties that are causally relevant to the current situation, which is in the considered case its density and solidity.

This scheme suggests a distinct temporal order for this kind of causal inference: being exposed to indicator variables – classification – prediction. The first two steps of this sequence enable an agent to classify an observed object, which unlocks access to further properties of this object-class, of which some are potentially causally relevant in the situation. It is interesting that the agent is thus exposed to some information about the object via an information channel that is distinct from the causal channel that concerns the content of the causal prediction. For example, perceiving the stone via its visible features is a causal process distinct from the possible future causal interaction when the stone is thrown. Catching up on an earlier example, becoming informed of toxic gas spreading in a building is realised via a channel different from the causal channel to which one would become exposed if no action, like leaving the building quickly, is taken.

If the interpretation of types as causally relevant properties, which are grounded in an agent's model, is correct, then we would also have an agent-dependent account of reference classes. The similarity of causal situations within one reference class turns on the fact that the same causally relevant properties are instantiated in the situation. Moreover, the relata of a causal relation are now further specified. Next to knowing that the relevant distinctions of causal 'variables' are indeed connected to localisable events (from locality), we now know the relata are instantiated properties. On the other hand, the downside of the approach is that we need an account for how token level causal claims are to be evaluated. This will be done in section 7.1.1.

The three properties, asymmetry, locality, and regularity, describe the relation of causation. These properties were inherited from the concept of naturalised action, with which causes were subsequently identified. They describe how agency looks like from the third-person perspective, but from that perspective there is no difference between actions and causes. The concept of causation is therefore analysed a second time. The causal relation is a local, asymmetric, relation of two correlated observables. It is the 'X' in the term 'correlation plus X', if we exclude notions like 'direct action', 'free action', 'semantic information', etc., which all belong to the subjective category, in order to describe what causation is.

The grounds of the asymmetry are still the agent-manipulations understood as in section 5.1, a concept whose natural origins are explained in chapter 6. The concept thus allows for an agent-centred and an objective reading, in the sense of a dual-aspect theory. The approach so far consisted in selecting agency as a promising account for analysing causation, and the subsequent naturalisation of the concept of agency, such that its thermodynamic properties can be revealed. Now these results have been used to constrain the causal relations and its relata, so that more can be said about valid causal models that consist of more complex structures than merely assertions involving two variables. In the following sections, I will make use of that information to underpin some causal judgments. The properties that constrain the causal relation and its relata are those that are, in other accounts of causation, often directly inferred from typical causal examples. Therefore, the relations that the analysis in this section has come up with do not come as a surprise. However, the virtue of the analysis consists in the fact that now the origins of these properties can

be explained further, since without such an analysis the expectation that enlisting properties can inform our quest for a better understanding of causation begs the question.

7 Causal judgments in LoA1, LoA2, and LoA3

Associated with the different stages of the analysis of agency and causation across the sections of chapters 5 and 6 were the different perspectives on agency. These can now be translated into the concepts that the philosophy of information recommends. Thus each perspective can be mapped to a corresponding level of abstraction (as introduced in 3.1). There is a conceptual distinction between intuiting causation on the basis of an agency-experience and explaining that causal judgment linguistically on the basis of the concept of agency. But I will assume that the contents of the two will not be divergent. Both forming the judgment and explaining the judgment are considered as occurrences that are detached from the actual course of events. The judgment is either prospective, or retrospective. With this precaution, we clearly stay at the conceptual level of causal analysis. Also, it is thus ensured that the informational account of causal judgments, which the method of abstraction (henceforth: 'MoA') promises, will not try to achieve more than it can deliver. In particular, this means that if a concept of free action is applied prospectively or retrospectively, the action does not really have to be free to underpin the causal judgment to which the concept of free action is applied.

Causes are evaluated as hypothetically performed actions, but actions have been explained as events that have a natural origin in the physics of non-equilibrium thermodynamics, which I have not analysed further and whose theory might require causal notions after all. My method does not ultimately decide on which of the two, cause or action, has the *ontological* prevalence.⁵⁴ But in order to underpin judgments, actions have the advantage of exploiting the first-person perspective, which yields a non-circular account, whereas analysis based on objective causes, like James Woodward's theory, seem to run into difficulties at some point. The order of the three levels of abstraction, which reflects the course of the argument of my thesis, depends critically on the first-person perspective. This dovetails with the fact that the identity of acting and judging agent posited by LoA1 is not explained further, whereas LoA2 merely approximates diachronic identity of the (observed) agent by the idea of centralisation. Markovian models like Bayesian networks, which belong to LoA3, cannot make sense of identity, except via the (extra-logical) interpretation of their variables.

From the point of view of my argument, the problem with those variants of materialism that do not posit subjects as distinctive kinds of entities is the assumption that a subject can 'model itself' like it can model all the other objects it observes (and which are observable by other subjects, too). This leads to some problematic results like the account by Dennett (1991) of how Mary, the colour-blind scientist, can anticipate her first experience of colours because she has knowledge about it by reading books about the neurophysiology of colour experiences. According to Dennett, that would enable her to judge whether certain prepared objects are truthfully coloured, because she would be able to watch her own reaction to the exposition and compare that to the expected reaction according to her disposition. Although it is conceivable that some bodily reaction of one's own body can be observed, it is by no means guaranteed that it is possible to the degree required by Dennett's argument against the reality of qualia, since two faculties of the subject – the reaction and its observation – would operate at the same time. On the side of actions, this corresponds to completely objectifying one's acting body, as if one is one's own puppeteer. But our body is steered

⁵⁴ Since a thermodynamic picture like the one outlined in chapter 6 has both actions and causes depend on drawing boundaries, it is plausible that there is no ontological order of causal concepts to begin with; at least if we read causation as a binary relation, which is a conceptual choice. In that case, such ontological questions arise at the level of regularities, not at the level of causation.

differently from how we steer an external puppet, or a virtual avatar. These are controlled *causally*, our body is controlled *directly*. LoA1 reflects this crucial difference.

LoA1, in its simple form, consists of just two variables, which become observables (in the MoA-sense) by interpreting them as an action-event and effect-event respectively. The degree to which the variables can be strictly typed depends on the context. For example, strict types can be assigned for events in a strict thermodynamic context, which is the paradigmatic context on which the whole theory of agency-causation ultimately rests. The behaviour of the direct-action-variable is governed by the assumption of any value from the range of possible values (determined by feasibility), following a free decision of the agent. The corresponding predicate is true simply when the event that putatively happened according to the judgment of the agent in fact happened as represented in the judgment. The behaviour of the effect is governed by a relation of dependence on the action, in the sense of chance-raising. The values the effect-variable can assume again depend on the specific context and the variable can be typed to varying degrees. On the conceptual level, there is no actual causal connection in an individual situation (as section 7.1.1 will show), therefore the truth of the causal relation between the two variables can only be established on type level by the agent. However, the agent can simply assign truth to a token level causal claim *bona fide*. In this case confidence is expressed that the action was a real difference-maker, and that the corresponding counterfactual ‘had the action not been committed, the effect would not have happened’ (for qualifications see 7.1.2 and 7.1.3).

If the agent judges its own action as caused by information rather than free, there are two ways to modify the simple variant of LoA1, i.e. the LoA for modelling a free action. Either a further event will be taken into account, and represented by its variable. This variable will stand for the fact that triggered the action. In addition to that variable, an explicit signal that carried the information about the past matter of fact for the agent can be represented. Corresponding behaviours of dependence can be defined between the causally relevant event, the signal, and the action. A judgment that explicitly represents these new variables could be formulated such as: ‘I did C because I learned that A on the basis of observing B, in order to bring about D.’ If one wants to avoid the teleological aspect of C’s dependence on D (which my proposed account only requires in order to explain the conceptual scheme of the agent), one could also formulate it as follows: ‘I did C because I learned that A on the basis of observing B, such that D followed from C.’ The judgment preserves the aspect of freedom from coercion stemming from the simple, two-variable case, in so far as the agent reconstructs the experienced event such that its action is suggested by considerations of optimising behaviour, while allowing that the time, at which the action is committed to, could have been chosen otherwise. The same holds true for the choice of the context from which utilities are chosen (for example, one could choose to commit to an action that makes use of a piece of information needed for that action, or one could choose to commit to an action that serves the improvement of the causal model in order to increase the chance of successful action at a *later* time; see also the remarks made in section 6.2.3 concerning vital and active goals, and 6.2.7).

LoA2 does not allow judging actions in a theory-free fashion as LoA1 does. The judging agent needs evidence that underpins the judgment that the agent judged upon does in fact commit to an action. Often, the observable action is part of a regular sequence of events. The quasi-spontaneity of the action has to be squared with the property of being dependent on other regular events. LoA1 faces a similar problem in the case of informationally triggered actions, where the seeming contradiction can be cashed out by resorting to a counterfactual evaluation that entails that every actual action could also have happened at other times, or could have been eschewed if the agent

had wanted that. In LoA2, the same purpose is served by identifying the variables belonging to information channels on the one hand, and the ones belonging to causal mechanisms on the other hand. This informs a behaviour (in the MoA-sense) between the causally relevant event that precipitates all signals that serve as information carriers. This flow of information is a causal process. But a variable's role as a signal requires the causal structure to be devoid of equivocation; otherwise no inference from the signal back to its cause would be possible. The channel structure, in conjunction with the computational capacity of performing the inference from the signal to the cause, enables the judged upon agent to react appropriately. Unlike the judgment at LoA1, judging the appropriateness of the action is more significantly tied to the flow of feedback (physical or informational) to the judged upon agent, otherwise it will be hard to see in most contexts how another agent's action can be discerned from an event that is not an action, unless the kind of behaviour is already well-known to be an action by the judging agent. In particular, this requires the judged upon agent to be semantically enabled. For example, one could think about a situation where certain test subjects are *told* to press a button when they see a flashing light, without giving them an idea which further events are set off by pressing the button. They might abide by the announced rule anyway and we would then interpret their doings as actions, even in case we are not able to identify any sort of beneficial feedback. To suppose that the test subjects have understood the rule and that they might have some motivation to abide by the rule would in such a case be sufficient for telling apart an action and a coerced bodily movement.

LoA3 enables the representation of causal structures, as long as one does not ask what a causal connection between two elements of such a structure *means* (an answer of which is given by switching to LoA1). The variables can again be typed to varying degrees of strictness, depending on context, and they stand for observable (in the default meaning of the word) events, which is the interpretation that turns them into observables (in the MoA-sense). At this level of abstraction, unlike LoA1, none of the events has a highlighted role to play. The number of variables is unbounded. A default assumption that governs all variables concerns their dependence on their respective predecessors according to an order. Variables can reside at a position parallel to each other in that hierarchy, and can have independent influence on their successors. This interpretation corresponds to the connotation of locality, whereas the ordering corresponds to asymmetry. All variables are interpreted as instantiations of causally relevant properties, which corresponds to the connotation of regularity. The aforementioned dependence on their ancestors determines the behaviours of the observables of this LoA. The structure of these causal networks depends on the assumed constraint-level (see section 4.2) of the causal system.

The next step which the method of abstraction recommends is the specification of the relations between the levels of abstraction. LoA3 can be supplanted by LoA2, if there is reason to impute the capacity of information processing and of being able to implement the result of the computation. Since this also implies the attribution of causal knowledge, or at least some causal model, replacing LoA3 by LoA2 results in a higher-order problem of causation. The causal agent judged upon might even react in a pre-emptive move since its causal model involves representing yet another agent that might interfere with its purposes, in which case the judging agent needs to model a causal problem of third order. The opposite procedure, replacing a LoA2-based causal explanation by a LoA3-based explanation, increases the number of variables. It loses the causally relevant matter of fact as the actually 'interesting' causal explanation, but allows asking for a detailed mechanistic story of what enables the adequate reaction. Alternatively, one might ask for further details of how an agent is informed of the causally relevant matter of fact, in which case one

can further fine-grain the model of the information channel, but still reside in LoA2, since the relay stations of the information channel are, in accord with this level of abstraction, not interpreted as efficient causes.

A plausible way of mapping the ideas incorporated by the above definitions of LoAs to the ideas underlying usual modelling strategies like causal Bayesian networks can be outlined as follows. LoA3 provides an interpretation of vertices (also called 'nodes') of a network as instantiated causally relevant properties, which entails that they will give rise to regularities. The ordering that results from the direction of the arrows corresponds to the connotation of asymmetry, whereas the fact that vertices of the same level of the ordering can be instantiated parallelly corresponds to locality. If the semantics of a node imply more than being an instance of a property that is causally relevant to an effect explicitly modelled in the graph, then these properties of the cause-object can be considered as accidental. If a cause-object has several effects, different causally relevant properties can play a role in bringing about this effect. That would be a configuration where we would expect modularity to govern the associated probabilities, since, if independent mechanisms entail modularity, it is a fortiori true that different types of mechanisms entail this constraint. Performing a manipulation of a variable from a network amounts, according to Pearl (2000), to lifting it from its old influences and putting it under the influence of a manipulation variable. The corresponding to switching from LoA3 to LoA1, where the target of the intervention is considered as a direct action that is not subject to a probabilistic measure, whereas anything that follows from the direct action, qua being indirect, is subject to such a measure. Spohn (2001) and Price (2007) make similar recommendations concerning evidence and action variables.

What has been explained narratively in section 5.2 can now be formalised more precisely. A direct action does not have to be an infallible measure of successfully bringing about an effect. It is not necessarily identical to bodily movement, either. But it is not assigned a rate of failure via a probability measure, as it is the case for indirectly brought about effects. If a failure occurs, this might prompt a change of level of abstraction. If an event, previously classified as a direct action, is now thought of as an indirectly brought about event, then this is a qualitative change in the status of this variable in the model. This is not quite the case in Objective Interventionism. A default assumption of Objective Interventionism is that what appears to be a direct cause according to one model can be an indirect cause according to a more detailed model. Therefore, drilling down into a more fine-grained causal model from a more coarse-grained model does not feature this qualitative change, since the presence of additional and previously unrepresented intermediate causes is the null-hypothesis of the models of Objective Interventionism. Similarly, in LoA1 any indirect effect following a direct action allows for insertion of additional variables that screen off the indirect effect from the direct action, while such an insertion is explicitly forbidden for the connection between the agent's decision and the direct action. If my account is correct in this regard, then there is no such thing as the 'intervention variable' from Objective Interventionism. An arrow in a causal model always allows for further fine-graining of the model, and therefore allows for a causal influence that fails to bring about the effect, including even the intervention variable, which is something LoA1 absolutely forbids.

7.1 Objectification - Switching from LoA1 to LoA3

The agent-perspective requires that the relation of 'direct action' is unanalysable (action as an unconstrained realisation of several possibilities), whereas the correlation (which is made asymmetric by *A*'s being the target of direct action) can be explained further on the basis of mathematical, therefore non-causal, terms. Since direct action is not analysable further, the semantics of the causal relation (understood as a binary relation) can be said to be grounded in direct appeal to everybody's experience as an agent. If I commit an action by bringing about *A*, such that *B* follows, then another agent, who watches me perform this action, should come to the same conclusion (using LoA2) as I do. As far the causal connection between *A* and *B* is concerned, this connection should persist even if we switch to a perspective that does not feature a free and direct action, and it should even persist if we perform a further step away from LoA2 and stop construing the observed behaviour as an action altogether (moving to LoA3). The causal judgment should survive these changes of perspective. Similar reasoning concerns the question of what it is that makes me think that my intervention would be adequate in the first place, before I start testing causal connections by intervening on a suspected cause in the system. The role of this truth-maker is played by the causally relevant properties, and the way they return a verdict depends on the structure of the internal model.

Since causal judgments of the form '*A* causes *B*' are, on the basis of this account, to be understood on type level and in a probabilistic sense, an agent can make the judgment '*A* causes *C*', on the basis of his judgments '*A* causes *B*' and '*B* causes *C*', if the *B*s are successfully identified, viz. the *B* which is dependent on *A* is the same causally relevant property as the *B* directly brought about in order to bring about *C*). But this means that, for *C*, it is irrelevant how *B* has occurred, since it is only relevant that the *B* *does* occur. In $A \rightarrow B \rightarrow C$, the *B* has a dual role: it is on the one hand dependent on *A* if we evaluate the $A \rightarrow B$ part, and on the other hand it occurs spontaneously if we evaluate the $B \rightarrow C$ part. But the agent's contradictory way of evaluating *B* does not instil a contradiction into the real events. It is rather the case that the seeming contradiction stems from a priori constraints from the level of abstraction used. Similarly, no contradiction arises if we allow for a revision of judgment concerning an event *B* that can either be considered as *another* agent's free action or follow (according to a causal rule) depending on a prior, yet to be observed, event *A*. This can apply even to judgments concerning our own actions, previously considered free and now considered determined in the light of new evidence.

The procedure shows that longer causal chains of causation can be synthesized by consecutively applying LoA1. If my action brings about *B* via bringing about *A*, and if I subsequently bring about *B* by means of a direct action, such that *C* follows from *B*, then I can infer that *C* follows causally from *A* via *B*. Agency is literally 'taken out of the equation' by extensional identity of the event of direct action and an event that is dependent on an action. But this procedure seems to require causal transitivity, which some authors, e.g. Christopher Hitchcock and James Woodward, argue against.

As an example, if we have three or more billiard balls and a sequence of one ball impacting on the next one in the sequence, then it is unproblematic to expect causal transitivity to hold in the example as far as our intuition is concerned. An example where such a conclusion is problematic is reported in Cartwright (2007): A dog bites the right thumb of a terrorist who was about to make a bomb explode by activating a trigger. Now that the right thumb is incapacitated, the terrorist uses the left thumb instead, and the bomb goes off. The transitive reading of the example has it that the

dog's bite causes the terrorist to use his left thumb to activate the bomb's trigger, and using the left thumb to activate the trigger would be the cause of the bomb's explosion.

Of course, causal transitivity fails in this example, and this is *prima facie* all the more the problem of an account that explicitly posits causal relata to be observables. The intermediate event (depressing the button with the second-choice thumb) is clearly only one physical observable, such that the causal chain should line up exactly as needed for causal transitivity. Therefore, transitivity should apply to the bomb-example no less than to the billiard ball example.

However, for each element in the causal chain we have to evaluate both relata of the relation individually, and this individual evaluation is done on the basis of the agency-formula. If one event is in fact the effect of the preceding relation, and would also be, brought about by any means, a proper intervention for bringing about the subsequent one, then we have the causal alignment needed for transitivity. This underscores that, although observable phenomena always play a role in causation, in the sense of a necessary criterion, they are not identical with causal relata, because, for the agency-account, *types*, or more precisely causally relevant properties, are relevant. There are empirical findings that corroborate the assumption that this is how our judgments actually proceeds (see 'inference over perception' in Sloman (2005)).

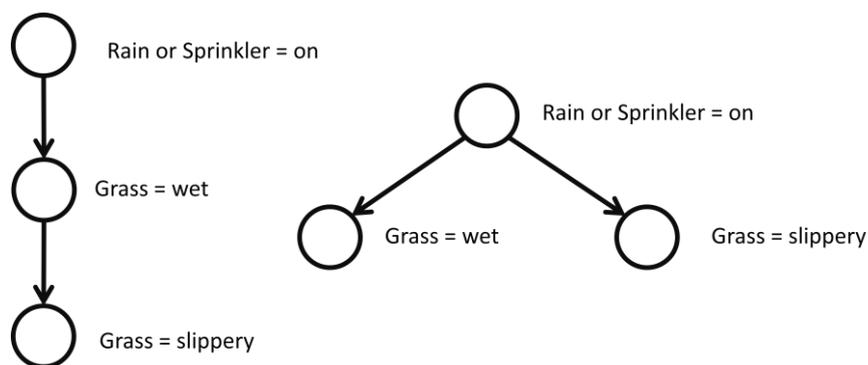
In the dog-bite example, one and the same observable represents two events in the agent's model: changing fingers in order to do whatever one has intended to do, and pressing the button no matter what finger is used, which are two events that don't line up for transitivity. The corresponding agency-consideration is: is making the dog bite any one of the fingers an appropriate measure to prevent the terrorist from pressing the button? No. Is making the dog bite the right finger a measure to make the terrorist switch the fingers? Yes. (This means the dog's bite causes merely a finger-switch. And the subsequent question is: Is switching fingers, given I intend to press the button anyway by any means, a measure to prevent the bomb from exploding? No (Result: *B* in its role as effect and *B** seen as cause don't align). By contrast, the billiard ball example passes the corresponding test.

Non-alignment of types of events is sufficient to negate causal transitivity in examples of token level causation such as the dog-bite example, and it shows that we reason correctly via models at type level. But this is a negative finding only. To establish token level transitivity requires the demonstration that the particular intermediate event brought about by the original cause was a genuine difference-maker with respect to the final effect in the transitive causal chain (see next section). Moreover, as the golf-example from the next section shows, the reasoning from types to tokens in the case of genuine token level causation that features unlikely chains of events is not a straightforward process of judgment. Positively, one can say that transitivity on type level is true. If *A* is a chance-raiser of *B*, and *B* is a chance-raiser of *C*, then *A* is a chance-raiser of *C* – unless there are further causal pathways via which *A* lowers *C*'s chance, in which case only the partial effect of *A* on *C* is positive (Woodward's definition of a 'partial cause' in Woodward (2003b) can be applied here).

The origins of the causal models lie in thermodynamics mediated by a thermodynamically formed concept of agency. As the name suggests, thermodynamically informed causal models are dynamical models. They have to work in conjunction with other internal models, like those that model logical relationships. There are several examples for the fact that such cooperation is necessary. Reasoning from the perceived properties of an object to the causal properties of the class to which this object belongs (see section 6.3.3) is one case. Another example is alignment in causal transitivity. If an event *A* causes an intermediate event *B*, then *B* will rarely be mentioned as an instantiated property causally relevant to *C*. For example, I enter my child's cluttered room and

stumble across a toy, which causes my falling to the floor. My falling to the floor causes the child to wake up. In this case, it is a specific property of my fall that does the causal work – its loudness. Probably any other loud noise would have had the same effect. Seeing my fall as an instance of that causal property is a non-causal inference.

The problem of token level transitivity shows that extensional identity of events alone leaves a chance of conflating the properties of being an effect and being the realisation of a causally relevant property into a single, but compound causally relevant event. There is a similar problem that arises when one intervention entails one observable effect, which is the realisation of different causally relevant properties. Pearl’s example of a causal Bayesian network with rain and sprinkler as two causes of wet grass, an example I have already criticised in section 6.3.1, has this structure. The wet grass allegedly *causes* the grass to be slippery as well. Pearl’s interpretation, besides confusing causal and explanatory reasoning, seems to stem from imposing the causal Markov condition as a necessary criterion for a correct causal model. The alternative model, the one the agency-approach in my interpretation would suggest⁵⁵, creates a v-Structure, where Pearl’s model features a causal chain (given one models the rain-variable and the sprinkler-variable as a compound variable for the sake of simplifying the case):



On the left Pearl’s version, on the right my version of the correct depiction of the causal relations of this three-variable case

The v-structure is due to the agent’s action of turning the sprinkler on or seeing that it rains over the patch of grass. This has the two effects of wet and slippery grass, which are not independent of each other, such that observing the wetness of the grass would be an indicator of the slipperiness, over and above what the presence of the cause indicates about the occurrence of the effects (the causes might fail to bring about the effect). As far as the concept of agency is concerned, there is nothing to be said against constructing the causal model of that situation in the above fashion.⁵⁶

Summing up these ideas, the account of objectification of agency (i.e. the account of how causation becomes an objective relation while being derived from the concept of agency) can be expounded as follows: In a typical causal model (causal graphs, Bayesian networks, structural

⁵⁵ Other interpretations might reiterate the fallacy of confusing causal with explanatory reasoning by judging that making the grass wet is an appropriate means of making the grass slippery as well. The fallacy consists in the fact that the measure taken by an agent is more adequately represented by saying that applying a layer of water on the grass makes it slippery (and it also makes it wet).

⁵⁶ Notice that Cartwright’s factory example (see section 4.2.4) also features a v-Structure forbidden by the causal Markov condition. But the structure is a different one, since product and by-product are extensionally distinct observables, while in the rain-sprinkler-grass example one and the same observable is an instantiation of different properties.

equations), every node has a causal predecessor, root nodes are subject to determination by error variables. The assumption of a value of a variable in such a model (Pearl's 'set-operation') is a mathematical notion. But if we apply this idea to a real world scenario and perform a causal judgment, then the agency-intuition needs to be applied as well. We cannot help resorting to the idea of agency and its uncaused events to attribute the role of cause to an event, even if the event is not connected to real agency in any way. I see this solution as a compromise between the objective and subjective aspects of causation. As far as anthropomorphism is concerned, it partly accepts the reproach of yielding an anthropomorphic concept of causation, while rebutting the simplistic conclusion of the objection, which denies that agency-causation cannot make sense of causation that does not evidently involve actual agents. I see the burden of proof lying with the objectivist camp, which has to show that causation does not impose our conceptual scheme as agents, especially in the light of the criticism of Russell (1913), Norton (2003), and others.

In the preceding section it was announced that the next due deliverable would be an account of token level causation. Such an account is required since token level judgments are probably equally relevant as type level judgements, so any conceptual account of causation cannot just ignore this class of judgments. Some more examples of applying the scheme provided by the three levels of abstraction follow in the subsequent sections, before discussing the causal Markov condition, which constrains an important class of causal systems and which is a further source of structural information about causal models.

7.1.1 Causal judgements concerning particular past events

The concept of causation that is given by the agency-formula affords causal judgements on type level. *A* causes *B* implies that *A* is a generic strategic means to bring about *B*. This can only be true if the connection between *A* and *B* is non-spurious and works at least in the sense of raising *B*'s chance compared to a default scenario. This kind of causal connection underpins propositions concerning *probabilistic* causation. So what if a particular event token, or, as a manner of speaking, a particular object, is assigned the status of a cause, like 'the cat's crossing the street caused the accident', which can be abbreviated to 'the cat caused the accident'? These statements have a certain implication, namely that

- I1 both events (the cat's crossing the street, and the accident occurring) did in fact happen,
- I2 and that the cat made a critical difference to the effect, such that had the cat-event not occurred, the accident would not have happened, at least not in the way it did happen.

The question is whether these implications survive scrutiny, and whether they do so across different concepts of causation, holding true in the agency-account in particular. 1. is quite uncontroversial; for a discussion supporting 1., see for example Mellor (1995), also Hitchcock (2002), who discusses 'cause' as a 'success verb', if used on token level. 2. Is supported by many counterfactual approaches to causation (Lewis (1973), Woodward (2003b)). To see that 2. also holds true, one must show that it would be meaningless to assign the cat the role of the cause in case the accident would have happened exactly as it did, even if we subtracted the cat from the event-description. Granted that both implications are true, however, they do not seem to follow easily from the agency-formula. First, it is difficult to think of any past token cause as being the result of an intervention. Is it plausible to think of the cat's crossing the street as the cat's own intervention on itself, or an intervention hypothetically added to the actual event specification? Even more difficult is the interpretation of the consequences of this supposedly direct action for the effect in question, since

the agency-formula only speaks of cause as an 'effective means', and the conservative reading of this term, which is also recommended by Price, implies that a cause merely raises the chance of its effect happening. Not considering here the question of the measure of probability at stake, and assuming that a measure for such a single token case can be found, the question would remain to which supposedly critical value the chance was raised by the cat's actions, in order to justify the causal blame.

These considerations show that the questions revolving around an appropriate interpretation and measure of *probability*, in order to address token level causation, are crucial. This section lays the foundation of an account of the meaning of chance-raising in the contexts concerning actual causation, and the account will be applied in the following section that deals with some well-known counter-examples against probabilistic theories. Another problem is the interpretation of causes as difference-makers in the context of agency, particularly worrying in cases of causal over-determination. That subject will likewise be treated in its own section. Both those kinds of problematic cases concern token level causation more than type level causation, which is why a successful defence of the agency-account depends on the model to be developed in this section. Another section of this chapter will shed some light on further problematic cases of token level causation, in the context of neural diagrams used to depict causal situations.

The problem of token causation is substantial for all probabilistic and all difference-making theories of causation. I will first expound the problem in its aspects specific to the agency-account, and in this context I will further specify, as usual, the challenges for the conceptual level of the problem.

We can approach the problem by thinking of an agent that intervenes on the state of illumination of a light bulb, via operating a button. The agent has observed a set of sequences of states with respect to these two observables. One pair of observations from this set consists of an interval where nothing happens (button remains in its default state of being undepressed, and the light is off), and a second interval during which the bulb either gets illuminated, or not. This specifies the agent's database of observations. In some cases, the agent has observed that the light turns on, although no change in the button's state has happened. In a second run, the agent can intervene on the button in the first of the two intervals of a sequence. The agent notices that after having performed the intervention of depressing the button in the course of the second run, the chance of the light bulb getting illuminated has increased compared to the first run, when it had been unable to intervene. If the chance of the light turning on is below 1 in the second run of the trial, there will have been cases for which the agent's intervention on the button was futile. Now imagine that the agent is asked, after the two trials, which of its individual interventions was successful, if we show the agent only the list of states of button and bulb for each individual sequence of the two runs, without further data. Since the agent knows that interventions can sometimes fail to bring about the effect, and on the other hand, the effect can happen irrespective of the button's state, such a question cannot be answered reasonably. Assigning truth values to individual events in that context would be tantamount to pretending to be able to predict an *individual* sequence of drawings of, say, red and blue balls from a ballot box full of equally distributed balls, rather than admitting that it is only the ratio of colours that can be predicted with some precision. Notice that in the considered scenario the agent would still be justified in asserting a type level causal connection between button and light bulb, and this would be true even if the agent is convinced that the type level causal connection depends on individual token level instances of a causal connection – it just happens to be impossible to tell the true cases apart from the false ones.

I neither want to delve further into a discussion of ontological priority of type over token level or vice versa, nor into an examination of how the information underlying the agent's judgments would have to be supplemented by further data, such that the above epistemic question could be settled. Instead, I want to focus on the conceptual question of what the agent's judgment could *mean*, given he does try to guess which ones of its interventions were successful, when a depression did in fact bring about an illumination, and frame this content in terms of agency.

The case of the depressed button that nevertheless might not be the responsible cause of the bulb's getting illuminated is analogous to how Woodward (2003b), on page 75, moots the problem of token level causation: 'It can be true that smoking causes lung cancer, that Jones smokes, and that Jones develops lung cancer, and yet false that Jones' smoking caused his lung cancer. (It was instead caused by his exposure to asbestos.)' The example chosen by Woodward clearly shows that, apart from special cases like causal over-determination, probabilistic accounts face a general challenge at this point. As do many other theories of causation, the agency-account relies on the concept of probability. The problem is that, as the example of button and light bulb shows, the concept of probability is adequate to help in forming type level judgments (via the frequency-interpretation of probability), and in situations where a prediction needs to be made (as in deciding whether the chance of the effect will increase if one intervenes). With respect to the computational model, which followed from the thermodynamic account of concept acquisition of action, both applications of the concept are necessary for a natural agent. The type level judgment is necessary to *create and adapt* an internal model based on experience (i.e. for inductive reasoning), and the prospective sense of probability comes into play in the *application* of the model, when in a concrete situation the agent needs to decide whether intervention is called for, or not. But the concept of probability is difficult to apply in retrospect, and that is a predicament for probabilistic accounts of causation as well. That is why some accounts suggest that the two claims 'A causes B' and 'a caused b' (see Woodward (2003b)) revolve around two different concepts of causation. It seems, however, the intuition of different concepts turns on an epistemological issue, rather than ontological constraint.

My suggestion of a solution to the problem is similar to Lewis' solution in asserting that it is primarily a counterfactual question. Also Pearl's solution takes that direction, which is even closer to mine since it also asserts that counterfactual questions involve a wider information base than prospective ones like predicting the effect of interventions or predicting a situation merely as an observer. Before formulating a generic proposition concerning a token level causal assessment by an agent, I will apply the ideas first to the aforementioned example involving the cat having caused an accident. The example is of an intermediate level of difficulty. It involves a type level probabilistic difference-maker (a cat crossing a street) of an accident. That makes the example more difficult than the deterministic examples which are considered in Woodward (2003b) in the section on token level causation. But it does not explicitly contain over-determining additional causes of an accident, and does not involve a difference-maker that, in default scenarios, *lowers* the chance of the effect, while still counting as the actual cause (as considered in the next section).

One thing to notice is that the only non-arbitrary values of probability for the accident happening after the subtraction of the cat from the scenario, given we do not opt for intervals of values, would be 0 or 1, unless we add some further background theory. As a first attempt to make sense of the token level claim, I will stick with the value of 0. This value-fixing, however, is justifiable only if we ask the question concerning the causal connection at the conceptual level, the question of what the basis of our judgment of the case is. Neither do we address the question in its

epistemological sense ('how could we verify the truth of the counterfactual conditional in the concrete case?'), nor do we address it in its metaphysical sense ('at the moment when the cat took action, was the complementary scenario fixed, such that the cat's action was a genuine deterministic cause of the accident? Or did an indeterministic factor remain *after* the cat contributed to the scenario specification, such that it remained merely a contributing (chance-raising) factor for the accident occurring?'). When we *judge* that the cat was the cause of the accident, we are assuming a chance-raising from 0 to 1, which is not necessarily a correct one.

It is worth discussing the case distinction that is indicated above in parentheses, concerning the metaphysical question of indeterminacy, and its possible bearing on the conceptual question. First, we exclude deterministic causation from our considerations on type level, for the ease of the discussion. *Viz.*, we assume that, *in general*, cats crossing a street do not necessarily produce accidents. Rather than that, there is merely a certain chance that they do. Then there are two further cases to consider.

- A) (epistemic indeterminacy) Either the cat's crossing the street is merely one of several factors determining how a situation unfolds, and the seeming indeterminacy even if we know about the cat's exact actions stems from our not knowing about all these other factors. These factors, however, will typically vary across different example situations of the type 'accident involving cats', and from this variance follows the epistemic indeterminacy, while the unfolding of events is actually completely determined.
- B) (metaphysical indeterminacy) It could also be the case that the cat is merely what Mellor (1995) calls an 'indeterministic cause'. Given a complete specification of the situation, and adding the cat to this specification, still a degree of indeterminacy remains. The cat has merely raised the chance of the accident happening.

These two variants can be modelled by encapsulating the 'other factors' in a single factor (which can be dubbed the 'second' factor) that works towards the outcome in conjunction with the cat. A), epistemic indeterminacy, can then be modelled by a randomization of the second factor *before* the cat takes action, such that the cat either makes a critical difference (chance is raised from 0 to 1), or it doesn't (chance remains at 0). Although it seems that the indeterminacy given by the fixing of the second factor (=conjunction of all other factors) contradicts the metaphysically deterministic scenario, this is not the case, since the randomization merely models the way we *arbitrarily select* an example situation from within a deterministic world, such that on type level a probabilistic scenario arises. On token level, however, the case becomes a deterministic difference-maker once the other factors of the scenario are fixed. B), on the other hand, is modelled by adding a genuinely indeterministic element by randomization *after* fixing the cat-event and all the other factors at their respective values. Now the cat is an indeterministic cause. This is how David Lewis and Humphreys (1989) model indeterministic causation.

Given I1. and I2., with I2. excluding the problem of over-determination for now, we see that the counterfactual is evaluated such that the cat did in fact make a critical difference to the accident which then did in fact take place. Evaluated according to A), the cat is judged to be an INUS-condition in a Mackiean sense, which necessarily had to bring about the accident given the other factors are in place. Evaluated according to B), the cat's action changed the chance of the accident happening, alongside the other causal factors, and *then* the remaining indeterminacy was fixed. Supposedly, without the cat's interference, nothing would have happened, since the chance in *that* scenario, after the fixation of its corresponding degree of indeterminacy, would have resulted in no

accident happening – at least this is what is that the judgment claims, otherwise the cat would not have been judged ‘the cause’ of the accident. But then it seems it is – assuming the stance of the judging agent – merely a matter of when to roll the die of the remaining indeterminacy, either before, or after fixation of all other relevant causal factors, including the cat. Whereas the question of metaphysical vs. epistemic indeterminacy is meaningful in general, it is meaningless in the context of judging a particular past event.⁵⁷ The question is whether there is a way the agent operates with its internal model and its levels of abstraction such that such a retrospective evaluation of events can be made plausible.

We do not have to stick with the considered example to notice the general issue at stake: ‘ $a:A$ caused $b:B$ ’ (with ‘ $a:A$ ’ meaning ‘ a as an instance of type A ’) connects particulars, and it does not allow for the kind of probabilistic qualification as ‘ A causes B ’. ‘ $a:A$ caused $b:B$ ’ is an absolute judgment. It asserts a ’s success in bringing about b , but not merely in terms of b ’s occurrence, since a ’s and b ’s joint occurrence could be spurious. For now, I assert that the relation implies the *assumption* that, without a , b would not have happened. At the core of the account must therefore lie the idea of a $a:A$ as a critical difference-maker, which places this account of token level causation in the conceptual range of the agency-account. But in the following it will be needed to spell out further qualifications of that idea. The qualification will revolve around how an agent makes use of its computational model to evaluate causal judgments.

First, the internal model, on whose calculations every decision to intervene depends, is for many pragmatic cases a probabilistic model. But the application of agent-probabilities (cf. Price (2007), regarding the discussion of stances of deliberation) and probabilities in general (Gibbs’ comment reported in von Weizsäcker (2006)) to past occurrences is conceptually problematic. Accordingly, the question of abduction (‘What was the cause of b ?’) is not a simple reversal in time of the prospective question of what will happen to b when we set a . Taking the question in the prospective sense, we are interested in the probability of b occurring, whereas the retrospective question takes, at some stage, the occurrence of a possible cause as granted, and instead asked for whether this cause was *critical* for the effect, rather than co-occurring with the effect in a spurious way. This grounds a difference in rationality constraints concerning token level causation. Under general assumptions (see von Weizsäcker (2006), Williamson (2009)), the subjective probability distribution should equal the objective probabilities. That is a constraint that enables the agent to assess type level causal connections, or predict token level instances of causation of future cases. This rationality constraint derives from survival pressure, since if the subjective probabilities deviate from the objective ones, then the agent will not be able to deploy effective strategies to reach its goals. But no survival pressure of that sort seems to apply to judging correctly what the cause of a specific observation was. In terms of the utility of solving such a question, two possibilities come to mind: assigning blame, and learning about a causal connection from a *single* instance. In assigning blame, we are not asking whether a putative cause event occurred, but whether it was the critical factor in bringing about the cause. Probabilities can corroborate one’s case, but they are insufficient

⁵⁷ Lewis would have rejected this interpretation, saying: ‘[T]he objection presupposes that the case must be of one kind or the other: either e definitely would have occurred without c , or it definitely would not have occurred without c . [...] But I reject the presupposition that there are two different ways the world could be, giving us one definite counterfactual or the other . . .’ (Lewis 1986, p. 180). I think that the difference between Lewis’ and my interpretation could be understood as trading on the conceptual-ontological distinction that I usually make in this dissertation.

to settle the matter (see section 7.1.2, 7.1.4). Similarly an agent cannot use single observations to increase one's causal knowledge in the form of incorporating a new causally relevant property into the causal model. That could not be grounded on a single instance, since any observed co-occurrence could be spurious. It is evident that the rationality constraints in judging past tokens of causal connections are therefore less well defined compared to type level judgments.

The softened rationality constraints in hindsight assessments correspond to a more flexible use of the level of abstraction of the internal model, which is why the corresponding scheme should be more sensitive to an agent's purpose and context of forming a token level causal judgment. Obviously, one has to be careful not to succumb to overly ad hoc adjustments. On the other hand, a certain flexibility in describing what an agent is actually doing during the formation of the judgment has to be granted as well. That said, in retrospective evaluations, the agent does not only often know about the occurrence of the effect *and* the cause, but can also supplement these observations by further assumptions of how other variables would have to be fixed. The agent can also twist the probability measure, by highlighting certain outcomes that are, in the type level context, considered unlikely. The subsequent sections will clarify these points further.

A possible formulation of a token level cause from an agency-stance is:

TLC: $a:A$ caused $b:B$ means that a particular a (of class A) was a difference-maker either for the fact that b (of class B) occurred, or for the specific way b occurred, or when b occurred, contrasting the real occurrence of b with a potential failure of b occurring without a , according to the internal model of the agent.

For a judgment, the effect's chance without a could have been high as well, but since there is a chance that the effect would not have happened, the potential outcome of the effect not happening in that counterfactual scenario is sufficient to ground the above judgment. If the causal structure, according to the internal model of the agent, is such that other causes would have necessitated the effect, and a had no influence on the specific way (including the exact time) b would have occurred, then a is not a token level cause of b . On the other hand, if we individuate the effect-event b , and the subtraction of a from the bundle of causes would result in a different individuation of an effect-event b^* , then a is deemed b 's cause, even if other causes from the considered bundle of causes would have necessitated the same type of effect. The point in time when the event starts is often particularly relevant for its individuation, as section 7.1.3 will show.

If, due to the presence of other sufficient causes, a scenario of the necessary occurrence of the effect is approximated, then a causal judgment that singles out one event as 'the cause' will appear more and more distorted and fail to capture the truth of the matter. But this is in accord with how token level judgments are sometimes made, e.g. in the context of blaming a, possibly preselected, cause. A token level judgment – ' a did in fact cause b ' – is a crisp concept, but causation might not work in this crisp way. Type level judgments allow for formulation of a varying degree of influence of a cause on an effect by the notion of probability, a notion that is problematic in retrospective scenarios. That is why token level judgments sometimes seem irrational and overstated.

' A causes B ' allows for qualifications, like A 's having a strong or not so strong influence on B , but ' a caused b ' only allows for the use of the term 'caused' in an absolute sense. That does not entail that we have two different concepts, but that we use the concepts in two very different ways. Attributing the status of 'the cause' in the face of the possibility of genuine indeterminism forces the

agent to commit to an assessment that seems more and more irrational, the more the effect's chance after setting the cause diverges from 1. The absolute sense seems to imply unwarranted determinism in a world that might be genuinely indeterministic (Hitchcock (2004) discusses this inconsistency and mentions its possible consequences for jurisdiction).⁵⁸

In sum, it seems that in token level contexts an agent does not operate in a betting scenario similar to when situations have to be assessed prospectively. The agent is also free to make additional assumptions that would not be made prospectively.⁵⁹ Taking all these preliminary considerations together, it is clear that a good account of token level causation must accommodate two conflicting intuitions – a certain flexibility on what counts as a token level cause, which is often unavoidably subject to interest, but which also takes into account the simple fact that intuitions are divided in some borderline cases. But on the other hand, an account cannot simply deny the intuition that whether something was a cause of an effect is also an objective issue, and not up to a completely arbitrary assignment of an agent. So any plausible account will have to strike a balance between both aspects successfully.

⁵⁸ This absolute sense 'forced upon' an epistemic agent prompted to make a judgment can be compared to trying to make sense of the statement 'the rose is red' in a black-and-white only world. Although sense can be made of that utterance, the corresponding proposition would be necessarily wrong.

⁵⁹ Notice a peculiarity of the question of token level causation. Although often addressing a problem conceptually abstracts from technical difficulties, it is in some contexts easier to solve the epistemic problem of elucidating via which means an agent came to a causal judgment than spelling out what the agent really means by the judgment. For example, see Woodward's example, cited above, of lung cancer, which has developed, not from smoking, but due to exposure to asbestos. One could adapt the example by imagining that asbestos, in all known populations, reduces the risk of developing lung cancer significantly, and by assuming that we do not know which sub-population Jones belongs to. Then, epistemically, the case allows for a pragmatic approach to the question: the smoking was the cause of its death, since the only known additional factor was a preventative of lung cancer. Still, one could assert that the latter was the cause of the disease. But it is hard to say – in probabilistic terms at least – what this judgment's content actually is.

7.1.2 Counterexamples against chance-raising

After expounding the general problem of how to make sense of a token level judgement grounded in an agency-oriented, type level account, more specific concerns can be addressed. These are illustrated by examples meant to be counter-examples against the idea of causes raising the probability of the effect. Although the concern is more specific for my theory, it must be clarified that all probabilistic theories, not just the type level theories, are affected by this concern. The issue consists in the problematic interpretation of ‘raising the chance’ of a singular event. Although the concept of ‘chance’, which typically (Williamson (2005)) implies the singularity of the case, is not inconsistent, the assessment of what the *value* of the chance is, poses severe problems, as will be seen in this section. Since the concept applies to singular events, a type level theory will have to make sense of this in a derived way. In my short overview of how to deal with the problem, I will first consider the singular probabilistic theories, then the type level theories.

There are several theories of singular causation that are also probabilistic. Hitchcock (2002) names: David Lewis, Patrick Suppes, Hugh Mellor as proponents of singular probabilistic causation, and he himself would have to be included in this list. Hitchcock (2004) states that there are two types of counterexamples against the theories of these authors:

- (1) causes that appear not to raise the probabilities of their effects;
- (2) and events that appear to raise the probabilities of other events, without causing those events

Since (2) concerns a phenomenon that corresponds to spurious correlations, I will not consider (2) at this point in much detail. (1) has attracted attention by some philosophers via an example that is originally due to Deborah Rosen. Hitchcock (ibid.) reports it thus: ‘A golfer lines up to drive her ball, but her swing is off and she badly slices the ball, sending it on a trajectory well to the right of the hole. Her slice decreases the probability that it will land in the cup for a hole-in-one. By chance, however, the ball bounces off a tree trunk at just the right angle to send it on a trajectory back toward the cup. As it happened, her slice did cause the ball to go into the cup, even though the slice lowered the probability of this outcome.’

This is how Hitchcock treats the case: ‘perhaps the relevant alternative is the one in which the golfer refrains from swinging altogether; relative to such an alternative, the slice actually increases the probability of a hole-in-one. My view [...] is that there is no objectively correct alternative for purposes of probability comparison. Rather, causal claims are contrastive in nature; they are true or false relative to a specific alternative. Thus, the golfer’s slice caused the hole-in-one, relative to the alternative in which she abstains from swinging, but not relative to the alternative in which she hits it squarely. In the latter case, we say that the ball landed in the cup despite the badly sliced shot.’

Hitchcock’s treatment of this case makes use of the idea of contrastive causation, which is, in this respect similar to my interpretation of the agency-theory, which contrasts different alternatives of action. However, his solution is wanting in that the contrastive cases which Hitchcock chooses are clearly not the ones suggested by Rosen, even if it is not spelled out explicitly. If we replace the action of refraining from swinging by the intended alternative, swinging in a straight way rather than slicing, then the problematic assessment of the example is restored.

Mellor (1995) treats the example in a slightly different way. According to his theory of causation, the causal relation between individual events, like the golf player’s slicing the ball and the holing derives from facts about these events, namely the fact *that* he sliced the ball, and *that* the

ball was holed. The contrast of the former of these two facts would be the fact that the gold player did not slice the ball at all, reducing chances of holing in to zero, according to Mellor. So far this treatment resembles Hitchcock's. Mellor then turns to the question whether it is true in the first place to claim that the golf player holes out because he pulls his drive, rather than saying despite that fact, or saying because he pulls the drive so that a tree is hit first. Denying the implications of the counterexamples at the level of causation as it is intuited enables Mellor to bypass the question of chance-raising. According to Mellor, conflicting intuitions reach a stand-off as far as the causal intuition is concerned, and that theory should command denying a causal connection between the aspect of slicing the ball and the holing.

Although my own theory does not operate with objective chances, I would still want to suggest a possible solution that does not imply a denial of the causal impact of slicing the ball. Might it not simply be incorrect, or at least ill-defined, to say that the slicing lowered the probability of the outcome? Why should the probability have been lowered in the first place? The slicing achieved an absolute success in holing the ball. It would not do so *in general circumstances*, and therefore will entail a smaller chance compared to cases of some reference class of broadly similar situations, but this is not the point at stake, since the example addresses a *particular* situation. I hold, similar to the case of the cat causing the accident in section 7.1.1, that the slicing increased the chance of the ball being holed to 1 (or close to 1 under genuine indeterminism), since any other value begs the question when and how such a value is determined.

Generally, slicing golf balls rather than hitting them properly does not increase their chances of being holed, and so indeed they aren't in any situation distinct from the episode that the example recounts. Similar considerations hold true for many other examples that introduce spurious, haphazard causal connections. If an agency-theory contents itself with governing only type level judgments, the theory would be vindicated thus.

Still, the token level problem needs to be addressed, since the easy solution of being content with a type level theory has been ruled out in the previous section. Then questions similar to the ones of the preceding section arise. How does the agency-theory deal with the problem? Would it be an adequate way to proceed for an agent to slice the gold ball in order to hole it? Obviously not, but in the particular case this is what caused the success. So we need to ask what the basis is concerning the judgment that the chance decreases by driving the ball *in the specific situation*. I would claim that there is none, because the reference-class for determining the probability consists solely of the specific situation, whose final result is a success. Notice that this objection against the common intuition does not deny that chances are real, but it denies that the agent is in any position to judge the *value* (even in a mere comparative sense) of that chance.

In order to judge the chance of a specific situation *objectively*, an agent probably needs a huge amount of information. For example, judging what the objective chances of the outcome of throwing a die amount to, an agent needs to assess in detail the geometrical and physical properties of the die, e.g. ascertain equal size of all the die's faces, symmetry, a central balance point (equal density), it needs to be ensured that the die will be rolled properly, probably also that a proper toss result is the authoritative one that stands, etc. There is not much doubt that the information delivered by the example is very underspecified to comply with these requirements. On the other hand, there is additional information available to the agent *after* having made the swing, and this additional information changes the evaluation drastically. Before the swing, the agent will assess the situation in accordance with a *ceteris paribus* law. Generally, slicing the golf ball will lower the probability of holing the ball, on the basis of similar experience. The underlying model features two

variables, connected causally according to a probabilistic model, with past frequencies in similar cases delivering the value of the probability. However, in that particular case, as we can see with hindsight, slicing the ball exactly in *that way* such that it bounces off the tree in *that angle*, it must have been the case that the chance of the ball's holing was significantly increased from a low default value, so that any other kind of slicing would have failed to bring about the effect. This way of evaluating the situation confirms that the slicing was a cause (contrary to Mellor), it confirms that the contrast consists in swinging in a straight direction (as the counterexample's intention is, and contrary to Mellor and Hitchcock), and it confirms that the chance of the singular event of holing was increased (contrary to Rosen, Hitchcock, and Mellor, but not contrary to intuition).

Woodward (1990) considers an example with the following structure: C_1 and C_2 are both causes and probability raisers of E , and work independently of each other. On a particular occasion, all three occur, but it was only C_2 that brought about E , since C_1 's operation has been a failure at that time. So C_1 raises the probability of E (and probably also its chance on this occasion) without causing E . Now that is an example of (2), the case when events which appear to raise the probabilities of other events do not cause those events. For the agency-theory, the structure of the example does not pose a challenge. Although the agency-formula (see chapter 5, first paragraph) asserts an equivalence of causation and agency properly understood, and agency relies on the notion of raising a probability, it does so only on type level, where C_1 is clearly judged correctly a cause of E . As explained in the preceding section, the derived account for token level causation does not allow for this kind of reasoning (see the example concerning button and bulb of the preceding section, where I explain that an agent has to engage in more complicated counterfactual reasoning specific to the situation in question, which sometimes requires additional information about how the events are specified; see also 7.1.4).

7.1.3 Over-determination

Evaluating cases of over-determination requires the distinction between token and type level theories, too. Again, token level theories tend to be more affected by them than type level theories. In line with what has been stated in section 7.1.1, I want to address both levels, while keeping in mind that the token level is derivative and depends on what the primary account, the type level account, allows.

Obviously, causal over-determination is *prima facie* a challenge to difference-making theories. My interpretation of the agency-account falls into the class of difference-making theories. The agent's action makes a difference to the intended effect. Therefore, the agency-account needs to be defended against objections stemming from that phenomenon. The objection says that if an effect is over-determined by two or more causes, none of them, taken separately, makes a difference to the effect, since each of the other causes already brings about the effect. One of the classical examples is Hall (2004): 'Two children, Billy and Suzy, are throwing rocks at a bottle. Suzy's rock hits the bottle first, just before Billy's. Suzy's rock causes the bottle to break, even though Billy's would have done so if she had missed.' This example clearly affects the token level *only*. On type level, throwing stones at bottles clearly is a difference-maker (and an effective means) to breaking bottles, since the situations with someone else throwing *and hitting* simultaneously a targeted bottle are rare and average out when regularities are considered. Examples of coincidental over-determination are easily dealt with on type level. But this does not hold true for the token level; also, there are more elaborate examples where problems do not go away that easily.

The type level equivalent of coincidental, token level over-determination is *systematic* over-determination. An example of it can be constructed from Mumford (2009), originally due to Hugh Mellor: 'A nuclear reactor [...] has the capacities to explode. When it is about to do so, a safety mechanism cuts in and shuts the reactor down.' The original point is that sometimes causal capacities have no chance of manifesting themselves, and an analogous relation exists between type level difference-makers and their chance of making a difference. The example has the right structure, if we count the normal operation of the reactor as a cause of its not exploding, which is over-determined by the safety mechanism. Otherwise one can adapt the scenario slightly. For example, one can easily think of several layers of security mechanisms in the nuclear reactors, where one mechanism takes over if the mechanism of the previous level fails. Systematic over-determination is an ill-conceived concept in an absolute sense, though, and therefore fails to be a counter-example against probability-raising on type level. First, the design of such a structure already takes into account that some primary factor *is* a cause, and therefore a difference-maker *prior* to the implementation of the over-determining super-structure, which, in the considered example, is implemented by the additional safety levels of a reactor. Secondly, there is always a hypothetical scenario, and therefore a reference class, where the additional causes are absent or fail to activate. Therefore, over-determining structures never touch the correct assessment that takes causes to be difference-makers on type level. Notice that, for the scenario of the failing additional causes, James Woodward's minimal definition of causes as difference-makers in *some*, rather than *all*, background conditions applies (see Woodward (2003b), page 40).

Turning to token level over-determination, things become more complicated. I will in the remainder of this section defend *TLC* of section 7.1.1, my formulation of a necessary and sufficient criterion for the token level judgment, and discuss examples unrestricted by domain considerations, except for causal models that follow the paradigm of neural diagrams, which will be discussed in the following section.

Christopher Hitchcock is one of the authors most involved with these issues, so I am going to discuss his treatment of the problem and compare it to my solution. The general approach he recommends is his principle *PSE*, the principle of precise specification of events, while I do not need a special purpose model and will stick with the general formula instead.

PSE: 'Suppose that on a particular occasion, events occur that instantiate types *C* and *E*. Even if, relative to the relevant background conditions, *C* raises the probability of *E*, if there are more precise specifications of the events in question, *C'* and *E'*, such that *C'* does not raise the probability of *E'*, then we should not say that *C* causes *E*.' (Hitchcock (2004))

There is a type-token slip in this formula, concerning the final line, which conflicts with the assertion that we are considering *a particular occasion*. It should be '... then we should not say that *c* causes *e* [where *c* is the particular event instantiating the type (!) *C*, and *e* the particular event instantiating the type *E*']'. Apart from this error, the formula is applicable to the following case: Barney smokes (*c*), and Barney likes sunbathing (not assigned a variable). Barney contracts cancer (*e*), and more specifically, he contracts skin cancer (*E'*). Although smoking (*C*) generally causes cancer (*E*), it is not true to say that Barney's smoking (*c*) causes his contracting cancer (*e*), so 'we should not say that *c* causes *e*'. That is because the more precise specification of what happened is Barney's contracting skin cancer (*E'*), and any more precise specifications of his smoking cigarettes (*C'*) has not raised the probability of *E'*. (Notice that, for consistency, we could have replaced *C'* and *E'* by the small letters equivalent to refer to corresponding particular events corresponding to more specific

specifications of the respective *types C* and *E*. I refrained from that, since I wanted to keep the alteration of the original formula minimal.)

The formula can be applied to classical examples of over-determination. Suppose there are two firings at a vase, each of them having an a priori chance of 0.5 of hitting the vase. Next, we are assured that one shot hits the vase, while the second would have missed it. The first shot is deemed the cause. The solution by applying *PSE* tells us that there is a more specific way of describing the shattering of the vase as it actually happens due to the first bullet, such that we have to retract from the a priori assessment of 0.5 assigned to the second shot's contribution to the effect's chance. The adapted value will be zero given a sufficiently precise description of the shattering.

There is another solution, which Hitchcock also discusses in Hitchcock (2004), where he says: 'Another possibility emerges if instead we focus on the trajectory of the first bullet. Suppose now that there is some time *t* at which the bullet fired from the first gun is determined to hit the vase. If we hold fixed the trajectory of this bullet at time *t*, then the firing of the second gun no longer makes a difference to the probability of shattering: The probability is 1 either way.'

I concur with both solutions, and *TLC* incorporates both. It ensures that the cause has to be a real difference-maker, which the second shot would not be given the first shot at some point completely determines the outcome in all possible scenarios. And it asserts that a cause must be a difference-maker for the effect *as it happens*. The constraint 'as it happens' must be read carefully, though. It is not sufficient merely to affect the outcome in its accidental features. For example, the second bullet might interfere with the trajectory of some of the vase's flying fragments, and by that have an influence on how the effect unfolds. The difference between causing and affecting applies here. It is therefore relevant that the way in which the cause affects the effect makes a critical difference to its individuation, which, in the considered case, is given intuitively by the time the vase started to shatter. This ensures that the first shot is indeed the cause of *the* shattering.

But Hitchcock is worried about additional empirical assumptions, such as taking for granted that in cases of over-determination more specific information can be given about how an effect-event unfolds, such that the cause can be determined. The second, and most substantial concern expressed (*ibid.*) is genuine indeterminism, which undermines the second of his two treatments of cases of over-determination. Thirdly and finally, there is a worry that the probabilistic approach to causation has to seek help from process theories, which consider causation as involving continuous processes. Only by means of such an assumption can we assume that there is a point in time when the cause completely determines the event. But I think that all three worries do not carry over to account for how an agent would ground its judgment of highlighting one of several causes in a structure of over-determination as *the* responsible cause of the effect. As far as the first and the third worry are concerned, my interpretation of the agency-account has no problems in making those assumptions. First, the event has already happened, so if one wants to supplement one's assessment concerning a particular past event by further assumptions of what would have been measurable, there is a priori nothing wrong with doing that. Secondly, nothing in my account of action, and therefore cause, speaks against the assumptions that events are measurable in ever more detailed a way. This would even be a very conservative, default assumption, just like the more conservative assumption that causes merely tend to bring about their effects rather than necessitating them. So it seems that it is the opposite of that assumption, i.e. expecting a discrete specification with limited available information, which is in need of further evidence. The next section on neural diagrams will show that the assumption is even likely to explain some unexpected intuitions in neural diagram structures. As far as the genuine determinism is concerned, I can refer to

section 7.1.2 again. If an agent is forced to make an assertion concerning alternative outcomes of an actual event in the absolute sense that 'caused' requires, where such a sense is precluded by an indeterministic context, then the judgment can only be irrational. This is a paradigmatic example of using a LoA beyond its specification.

Over-determination is also discussed in another context. It concerns the possible epiphenomenalism of a mental decision that leads to an observable bodily movement. The effect of the bodily movement would be over-determined by a preceding *physical* cause if we assume the causal closure of the physical world. It seems that every agency-account of causation should say something concerning this issue. In chapter 6 I developed a scenario which allows the concept of action to emerge while asserting the causal closure of the physical world. This seems to favour an account that equates mental phenomena, including decisions, as an epiphenomenon. An alternative account, which asserts that free decisions are not only construed as such with hindsight (or projected as such for the purpose of planning), but real *in situ*, would indeed undermine the whole argument of identifying causes with actions. That being said, the category of the subjective world, in the sense of a dual-aspect theory, is in harmony with my account, since physicalistic models like Markovian causal networks lack the resources to express crucial ideas such as the identity of actor, information processor, and beneficiary of an action. Also, although it is consistent with my account to assume that the subject is determined in its decision by the way it is configured to react to information, believing in the modifiability of one's own structure does make a difference to one's behaviour. Thus, reflecting on and then regretting past decisions can change the future response to the same stimulus. Therefore, in a certain sense there is a causal impact of the mental; however, not *in situ*, but in the long run. A concern of possible over-determination in that sense can be circumvented by an appropriate design of the level of abstraction of the question. Thus, the question asking what determines the agent at the time when the stimulus arrives at the agent must be answered by referring to the physical structure as the responsible cause. This is basically the same solution as the one that dissolved the 'contradiction' of seeing information as a cause of observable behaviour rather than the event that precipitated the flow of information about it.

7.1.4 Counterexamples from causal diagrams

Hitchcock (2004) admonishes that in some cases of causal over-determination, we are disposed to ask further questions about the case, in deciding which of two possible causes the actual cause of an effect was. I agreed with that assessment in section 7.1.3., and added that, if there is no prospect of getting further information, a model for causal evaluation might just be supplemented by hypothetical data if that helps to underpin a judgment. Examples of causation are normally drawn from realistic, real world scenarios, and must then be mapped to a model. Intuition supposedly has some way of modelling the conceived situation in a wider sense, and a theory of causation tries to capture the right intuition, at least for those cases when intuition yields a univocal result. Ned Hall has developed a model for representing causal structures in the form of graphical diagrams. Interestingly, it is not examples drawn from a real-world situation, but from the diagram itself that he wants us to judge according to intuition. This intuition is then compared to what theory would say about the case. But the diagrams are *already* the model. That means that both intuition and a theory-driven model are applied to what has been processed – and possibly distorted – by a *specific*

modelling technique. In that regard, the approach is interesting, but also very problematic.⁶⁰ In this section I will discuss some examples from Hall (2007). Since graph-based representations do seem to play a role as an important modelling paradigm (e.g. for visualizing causal Bayesian networks), and are therefore possibly relevant to the internal representation that human agents' intuition uses (cf. Sloman (2005)), I will use this space to show how the levels of abstraction appropriate to my theory deal with these cases.

For his analyses of causal structures, Hall uses a specific graph-based model he calls 'neuron diagrams'. The nodes of such a diagram are called 'neurons', the event of their firing or not firing is modelled by assigning the arbitrarily chosen values of 1 and 0 respectively. If the tail of an edge connecting two neurons has a blob, like the edge connecting C and B in the diagram given below, then the connection is inhibitory, otherwise it is excitatory. The model is deterministic, and the relations can also be modelled by an analogous system of structural equations.

By means of neural diagrams, Hall considers several cases of token level causation, for example the following structure:

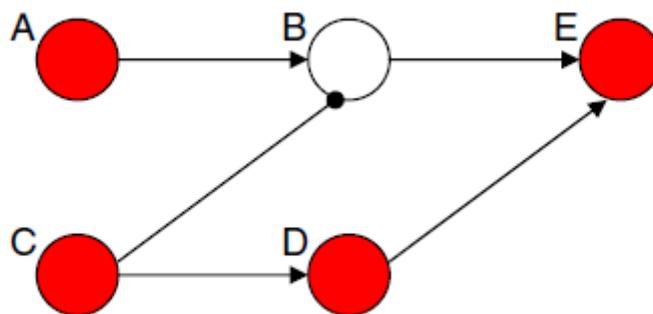


Diagram 7.1: Hall's example of causal pre-emption

A is firing, but C is likewise firing and, by virtue of being a preventative of B, it blocks A's influence along the A – B – E path. C is also a cause of D, which is a cause of E. According to Hall, and I would agree with this assessment, C is the actual cause in this instance of neuron-firings. Hence, we seek an account that identifies C as the actual cause of E, while A is causally irrelevant to E in this situation.

Hall deploys his 'H-account' (which resembles Woodward's concept of 'actual cause' in Woodward (2003b)) to predicate A negatively and C positively as cause of E. Next to A and C firing for being considered as candidate causes, and the effect E firing, we require that 'C is a cause of E just in case there is a path from C to E, such that for zero or more off-path variables $X_1; \dots; X_n$ with actual values $v_1; \dots; v_n$, the conditional

if $(C = 0 \ \& \ X_1 = v_1 \ \& \ \dots \ \& \ X_n = v_n)$, then $E = 0$

In the considered case, application of the rule to the path A – B – E requires the conditional: if $(A = 0 \ \& \ C = 1)$, then $E = 0$. But that is false, so A is not the actual cause. Applied to C – D – E, the conditional: if $(C = 0 \ \& \ B = 0)$, then $E = 0$ needs to be true, which it is. Thus C is the actual cause of E.

⁶⁰ The problematic aspect is encapsulated by Hall's saying '[T]his is one of the advantages of working with such diagrams: they provide clear tests for any analysis of causation.' (Italics added), *without* saying in which way we have to qualify the result of such a test.

The H-account has in this case managed to capture intuition. But it is questionable whether intuition tells us anything substantial if applied, not to a real-world example, but to a diagram, and therefore it is likewise questionable whether any account that tries to track intuition as much as possible in that context tells us something interesting about causation. The problem, I think, consists in how intuition deals with the neural diagram scenario and its constraints. These consist in a completely discretised picture of the events – there are only on- and off-neurons, time does not seem to play a role, nor do properties of the connection between two nodes (their length is irrelevant, etc). Also, a neuron's firing *necessitates* the firing of its descendant. Apart from that, the model is causally sufficient, such that a neuron's only causal antecedents are those in the model, and it is complete, such that when a neuron has some descendant nodes, these comprise all the effects the firing of this neuron has. All of these constraints would fail to hold true in an absolute sense in any real implementation of the diagram, and even more in the real-world case from which the scheme is, according to Hall (p. 111, *ibid.*), derived. So the question is whether we can expect intuition to tell us something authoritative about such cases. I assume that intuition does not judge the neural diagram case, but a potential *physical implementation* of such a diagram, where the above constraints cannot be expected to hold true.

The formula *TLC* tracks this intuition by trying to take into account as best as it can the agent-perspective in its variant of hindsight-judgment. Assuming that perspective, *A* is not a real difference-maker, since it is in fact blocked by *B*. *C* is a real difference-maker for how *E* happened compared to some potential outcome in the counterfactual scenario of not having acted on *C* (= *C* switched off). First, a reasonable default assumption about a physical realisation of the causal structure of diagram 7.1 is that the time of *E'* (in case it would have been brought about via *A* – *B* – *E*) would be different from that of *E*. And secondly, because it is possible that the mechanisms connecting *A* and *B*, or *B* and *E*, might have failed. If, however, the internal model of evaluation abides strictly by the rules of the neural diagram, *C* would fail to qualify as a cause, because it would fail to meet the criterion of being a difference-maker to *E* in any significant sense. But the cause of this failure to track intuition might be that intuition does *not* abide by these – unnatural – rules of neural diagrams. This is why I have doubts that Hall's diagrams will ever bring about a clear-cut account of what it means to be a token level cause.

Interestingly, in the second half of the paper, Hall acknowledges a distinction that closely resembles the distinction between spontaneous and non-spontaneous processes, which is absolutely central to my account of where the concept of agency comes from, since the agent is itself subject to that distinction; that is, not only as observer of an external system. In his terms, the corresponding primordial dichotomy is between default and deviant states, which are the states of an object when nothing acts upon it and when something in some way acts upon it respectively. Hall admits that, by virtue of the concept of 'acting upon an object', only a causal criterion can be given for that distinction. But the distinction is crucial to the evaluation of the counterfactual that we need for an account of a token level cause, and again I agree with that assessment. I also refer to sections 6.2.7 and 6.3.1 on the question of conceptual circularity. Hall observes that the evaluation of the counterfactual 'Had *C*, the putative cause, not happened, what would the rest of the objects of the system, and the effect in particular, have done?' The removal of the cause-event requires us to set the cause-object to a default state, which is distinguished from all others in that we would assume this state to be occupied by the object if we did not know anything else about the causal history of the object.

Allegedly, the notions of default and deviant behaviour help to shed some more light on further considered examples. But the question remains whether some sort of final account can be achieved if one continually tries to track the results of a possibly misguided intuition (see Hitchcock (2009) for further problems of the neural diagram approach). As a final comment on the approach, consider what the motivation of these retrospective causal evaluations could consist in, seen from an agency-stance: either an agent wants to learn about a possible causally relevant property new to its internal model, or it seeks to update the empirical data concerning the strength of the influence of a probabilistic cause to a more objective value. In the first of these cases, something about causal structure can be learned; in the second case, the probability measure that governs the internal model can be updated. Both cases are ruled out in the neural diagram considerations, since we know both the causal structure and the (degenerated) probabilities. The second motivation is assigning responsibility to a single neuron. As the cases of early pre-emption above, but also examples of over-determination show, this can be done, but it is often arbitrary to some degree, and it seems more informative to show in all openness the known causal structure. Then responsibility can be shared in a fair way. But this is obviously not what the account seeks to achieve; instead the goal is a perfect projection of a causal structure to a single object that carries the causal responsibility in an absolute sense.

7.2 Different constraint-levels inform different causal concepts

Based on the findings of the previous chapters, one can now analyse the import the concept of causation given by objectified agency has for standards of model validity. First, the hierarchy of causal concepts established by different constraint levels will be outlined, and proofs of their hierarchical relation will be delivered. Subsequently, the concept of causation derived from agency will be situated within that hierarchy. The result of 4.2 has been the negative finding that we should not expect the causal Markov condition to follow immediately from the notion of an intervention, while in this section the constraint level of causal models that does comply with causation as derived from agency will be situated on the hierarchical map of causal concepts.

A model that is not level invariant (and therefore neither modular, nor coefficient independent), but rather orders correlations according to customary causal inference, is useful for causal predictions by conditioning on evidence variables. Gillies (2002) has called such a network a 'propensity network'. A concept of causality from the classical literature that corresponds to such a moderately constrained graph is advocated in Schlick (1932). A further constraint of causal models is given by level invariance (*LI*). *LI* critically hinges on *interventions*, and therefore marks the transition from interpreting the antecedents of inferences as propensity-raisers (or, if other accounts of probability are used, from frequency-, chance-, or degree-of-belief-raisers) to genuine causes. If a system is modular according to the *MD* condition, *CMC* is satisfied under both interpretations of probability described in 7.1.1, and the correct skeleton and the v-structures can often⁶¹ be determined on the basis of observational data. Modularity implies level invariance, but not vice versa

Proof:

Modularity \rightarrow *level invariance*: Let S be a system of equations, and V_i one of the endogenous variables. Due to modularity, the equation determining V_i can be replaced by an operation that sets V_i 's value directly, and this operation has no repercussions on any of the remaining equations of the system. Accordingly, it is either the case that some considered equation from the set of remaining equations of the system contains occurrences of V_i as independent variables. Then none of the parameters of these equations have been changed by the intervention, so the equation is level invariant (since invariant *simpliciter*) with respect to V_i . Or it is the case that such equation does not contain V_i as independent variable, in which case it is vacuously true that the equation is level invariant with respect to V_i .

Level invariance (not \rightarrow) *modularity*: Cartwright's factory example serves as a counterexample. The structure does not comply with the causal Markov condition, so *a fortiori* the system is not modular, but its equations are level invariant: it is possible to manipulate both E_1 and E_2 by manipulating C without invalidating the equations for E_1 and E_2 . One can also manipulate any of them, e.g. E_1 , by manipulating the parameter a_{E_1} connecting C and E_1 , or by manipulating the error variable u_{E_1} , without invalidating the equation that determines E_1 . But because of the correlation between the parameters (or the error terms, depending on the mathematical modelling), the intervention would have repercussions on E_2 , therefore modularity is not satisfied.

The strongest causal model in terms of the probabilistic constraints is the one that features coefficient independence (*CI*). In such a case, not only are the effects of a common cause independent of each other, but also all causes of a given effect and their corresponding mechanisms

⁶¹ See the constraints described section 4.2.3.

are independent of each other. For a deterministic *SEM*, this relates to independence of all coefficients, including those of a single equation. In a corresponding probability specification, *CI* relates to manipulability of an effect via a single cause while holding the values of its other causes fixed at any chosen level. In common causal terms, this corresponds to the exclusive presence of disjunctive causes, rather than cooperative causes.⁶²

Proof:

Coefficient independence -> Modularity: If coefficients of all equations from a system of equations are independent, then there is a modular intervention for any given variable V_i from the system, by appropriately setting the coefficients in the equations for that variable. These settings have no repercussions on other equations of the system.

Modularity (not ->) coefficient independence: Examples of cooperative causes, e.g. oxygen level, combustibles and spark plugs cooperatively causing fire. The relation of extensity of fire and oxygen level is proportional, but the parameter that quantifies the proportionality depends on the parameters that regulate the influences of the spark plugs and the combustibles. Such a system can be modular, but there is evidently no coefficient independence.

Those theories that have an underlying mechanistic ontology, and also assume that causal relations are ultimately deterministic (e.g. Judea Pearl's theory), equate causal with modular systems, which comply with *CMC* under the assumption of acyclicity and independent error variables (causal sufficiency). Woodward, Hausman, and Steel consider genuine indeterministic causation, but they also rely on a mechanistic ontology. Since each arrow in a causal graph, or, alternatively, each factor in a *SEM*, is taken to represent a distinct mechanism, the systems they consider as causal are likewise modular systems. Cartwright's six assumptions about causal systems, given in Cartwright (2007), are consistent with *LI*. Accordingly, she explicitly endorses this constraint level in Cartwright (2002). Modular systems, in her framework, are particular causal systems that are more strongly constrained than causal systems in general. Jon Williamson's epistemic theory does not endorse *CMC* generally, since its endorsement is dependent on prior causal beliefs that allow for both modular and non-modular systems.

The concept of causation derived from my interpretation of the agency-account is situated at the constraint-level of *LI*, as far as *LoA3* is concerned. This can be illustrated with the simple example of a level invariant system given in Hausman and Woodward (1999), who consider the regression equation:

$$Y = aX + U$$

Level invariance states that this equation represents truly the relation between X and Y , no matter whether we read this equation in an observational or in an interventional context. *LoA3* interprets the equation thus:

$$y:Y \Leftarrow a x:X + u:U,$$

⁶² An example of such a structure is a charged particle whose trajectory can be manipulated by a gravitational and an electric field. Manipulating any of the two fields contributes to the overall acceleration of the particle, and the contribution of one field is independent of the value of the other field, so the causes are disjunctive.

where ' \leq ' is to be interpreted as 'causes'. The equation's variables are thus interpreted as tokens that instantiate corresponding types. If repetitively instantiated, the equation gives rise to an observable regularity. The equation is not symmetrical and therefore does not allow for isolating x , or allowing any prediction of x when y is intervened on. The locality of the concept is reflected by the presence of additional factors (U), which render the relation between x and y a *ceteris paribus* rule. By contrast, the constraint of modularity of a system of equations is not a necessary condition. Since the model of Cartwright's factory is level invariant, it is a causal model according to the concept of causation derived from objectifying agency. Therefore, it is a counterexample against the claim that causal systems are situated at the level of modularity in our considered constraint-hierarchy. Modular systems are therefore *specific* causal systems.

8 Summary, final conclusion, and outlook

8.1 Summary

In the quest for the right approach to explain the meaning of a causal proposition I was guided by two convictions. First, there must be some significance to the fact that an agent feels that a causal connection obtains between two observables if she has herself performed the intervention on the observable representing the cause, and, secondly, that there is significance to our concept of causation stemming from the fact that living organisms have to be agents in order to maintain their structure, by exploiting the asymmetry between spontaneous and non-spontaneous processes. This asymmetry depends on drawing boundaries and is otherwise a statistical phenomenon. The agency-account of causation by Huw Price and Peter Menzies immediately expresses the first conviction. I have attempted to defend the account against some frequently articulated objections. This, however, is just a necessary step to vindicate the approach. To further corroborate it, I have looked for a connection between the first and the second conviction. The bridge between the two turned out to be information theory and a phenomenon which I have called ‘causation by information’. Information is conveyed via physical signals, but if the event this signal stands for causes some behaviour, rather than any physical property of the signal being responsible for that, then the observed behaviour is an action. To explain this phenomenon further, the thermodynamics of non-spontaneous processes were examined. Although I have left open the possibility that the arrow of time might be a construction of the agents’ perspective as deliberators, I took the asymmetry of time as a premise of my argument, and also assumed that entropy rises within time. I treated both these assumptions as (non-causal) facts. Causation arises from this metaphysical backdrop as a phenomenon that allows localised, improbable structures to exist by consuming free energy. The causal arrow does not follow the thermodynamic arrow in the sense that A is the cause of B because A is the state of lower and B the state of higher entropy, but instead in the sense that A drives the local process B by virtue of its higher entropic magnitude. Following this lead, I found the scenario of Maxwell’s demon as a significant object of the combined study of thermodynamics, information theory, causation, and action. It allowed the derivation of not only the finding that contingent information, the resource needed to guide purposeful action, needs to be physically represented, in order to prevent breaches of the second law of thermodynamics. But it also enabled an account of how efficient causation and causation by information are connected, since there are two different ways to describe the chain of events during the demon’s sorting of particles. The demon obeys a rule, such that causation by information gives rise to an observable regularity, although actions are involved. The rule was interpreted as a function, a mapping from the signals containing contingent information to the optimal reaction. But if there is a function, then there is a computational model that computes this function. My interpretation of this computational model, which belongs to the agent and constitutes its identity as information addressee, acting agent, and beneficiary of the action, also delivered an account of free action. With that, I found an interpretation of the required spontaneity of the A in the judgment of ‘ A causes B ’. The thermodynamic model grounded the asymmetry, locality, and regularity of agency. An identification of the concepts of agency and causation let causation inherit these three properties that inform actions. I subsequently devised three levels of abstraction that are respectively applicable to the different perspectives previously assumed. The concept of levels of abstraction was shown to be sufficiently flexible to allow specifications of causal concepts as parts of a wider family of concepts. I have concentrated in this dissertation on providing my own positive account of how causal judgments can be analysed, yet I

hope it is not impossible to see how other approaches could be integrated into the account if the focus of causal analysis is a more specific one than mine. For example, it was shown that further constraints can be added to the generic three properties, like modularity or different variants of indeterminism. In Chapter 7 it was expounded that some problems concerning the judgments we make about causal scenarios can be addressed without explicit commitments to further metaphysical constraints. What defines the aforementioned family of concepts is the compliance with the above 'agency-formula', which occupies the root of a conceptual hierarchy that determines causal concepts. The levels of abstraction I have used throughout the thesis depend on the pre-set purpose of justifying generic causal judgments of the form 'A causes B'. Less conceptually orientated approaches, focusing more on epistemological or ontological issues, might construe the object of studies of causation very differently, e.g. the study of causation according to a causal process theory, or a mechanistic approach. In these cases, integrating a corresponding interpretation of causation into the conceptual family might be more difficult or not feasible at all, but that assessment requires another kind of analysis.

8.2 Final Conclusion and outlook

The proposition 'A causes B' means 'the occurrence of A would be an effective means by which a free agent could bring about the occurrence of B'. This dissertation is an attempt to defend this fundamental claim, and I am convinced that a conceptual reading of the equation is indeed plausible. The decisive argument that prioritises agency over causation is the non-circularity of this conceptual explanation, and the fact that agency makes sense at least from the first-person perspective.

My argument has conceded some points made by critics against the agency-approach. Indeed, the concept of causation given by the agency-approach is to some extent anthropomorphic, but not to a degree that would invalidate causal claims not actually involving human agents. It has been granted, too, that the above equation yields little information about causation, if the manipulability of effects via their causes is the only constraint that can be imposed on causation. Seen as a project of pure conceptual analysis, it is hard to make a case for the identity of actions and causes. Also, it is difficult to see how we can make sense of more complex causal structures, or of causation at token level.

My project has not been a reductive analysis of causation in terms of agency. However, agency inevitably belongs to causation. Agency, seen as a phenomenon, served as the starting point of the conceptual reconstruction – but it stayed with the concept of causation throughout. I contend that any causal judgment '*a* causes *b*', if we press for an explanation of the meaning of that judgment, will have to resort to the first-person notion of agency. Interventions require an interpretation as action – when we decide to push a button, we do not 'wipe out equations', 'break arrows', or do anything alike. Instead we directly act, exploiting thermodynamic asymmetry. The sense of the detour via thermodynamics of action was given by the fact that, without it, the agency-account would have stayed potentially true, but uninformative, as Hugh Mellor, Wolfgang Spohn, and others had noticed. The detour delivered further information about agency, and therefore on causation. But the analysis was not meant to fix the secondary intension of our object of investigation, vaguely identified via the notion of manipulability. Instead, the thermodynamics of actions deliver further information about how we have to evaluate causal structure in our world; it therefore delivers information that has to be considered as contingent. One can think of agency that is not thermodynamically constrained, and whose concept does not have thermodynamic origins.

Arguably, all causal judgments can be thought of as derived from causation in those alternative worlds as well. In that sense, the thermodynamic analysis I have given after defending the agency-approach against the objection of circularity is not different from analyses of causation in terms of mechanisms, of modular systems, etc. But the approach that I have taken allows a much more logical, step-by-step synthesis of concepts of causation. Since the further constraints on causation are contingent, the primary intension of causation as a relation of manipulability seems to remain the semantic hard core of causation – it is not a mere symptom of causation.

The causal concept is therefore based on agency and subjectivity, even if an analysis of its grounding seems to show that it is itself embedded in a thermodynamic metaphysical backdrop. To ask the binary question of what it *means* when we say *a* causes *b*, if *a* and *b* are taken from the context of a wider and more encompassing causal model, is a crass turning towards a subjective level of abstraction – otherwise one can well work with causal models while avoiding the question of spontaneity.

The specific arguments I have devised in order to vindicate of the ‘agency-formula’, by which term I have dubbed Menzies and Price’s identification of causation and agency, constitute the main points of this thesis. I could only indicate how different concepts of causation could be accommodated into a family of concepts, and where the connection between a concept of causation and a level of abstraction as a prior constraint of that concept lies. These questions concern the relation between different levels of abstraction for the evaluation of token level cases. Since difference-making on type level is the primary constraint that informs our concept of causation, which is tied to the strategy the agent needs to act prospectively, there seems to be a conflict between type and token level claims that I could not completely solve. It seems to me that more empirically informed levels of abstractions are called for, which should take over from more generic levels of abstraction.

I have outlined some parts of what I think of as the web of causal concepts, which is in parts a hierarchy, as the last section of this thesis has shown. The position of a causal concept within this hierarchy is determined by the number of constraints that the concept satisfies. Some approaches to causation start with fundamentally different categories (e.g. mechanisms, counterfactuals, causal processes). It might be part of future research projects to work out how the causal concepts these different approaches give rise to can be integrated into a more complete conceptual map of causation.

9 Bibliography

- Ahmed, A. (2007). Agency and Causation. Causation, Physics, and the Constitution of Reality. H. Price and R. Corry, Oxford University Press: 120-155.
- Armstrong, D. (1983). What is a Law of Nature?, Cambridge University Press.
- Arntzenius, F. (1993). "The common Cause Principle." PSA **2**: 227-237.
- Atkins, P. (2007). Four laws that drive the universe. Oxford, Oxford University Press.
- Bayne, T. (2011). The Sense of Agency. The Senses: Classic and Contemporary Philosophical Perspectives F. Macpherson.
- Bennett, C. H. (1982). "The thermodynamics of computation—a review." International Journal of Theoretical Physics **21**(12): 905-940.
- Bertalanffy, L. v. (1969). General System Theory. New York, George Braziller.
- Bogdan, R. J. (1988). "Information and semantic cognition: An ontological account." Mind and Language **3**(2): 81-122.
- Bogen, J. (2004). "Analysing Causality: The Opposite of Counterfactual is Factual." International Studies in the Philosophy of Science **18**(1): 3.
- Cartwright, N. (1989). Nature's Capacities and their Measurement, Oxford University Press.
- Cartwright, N. (1993). "Marks and Probabilities: Two Ways to Find Causal Structure." Scientific Philosophy: Origins and Development.
- Cartwright, N. (1999a). "Causal Diversity and the Markov Condition." Synthese **121**(1): 3-27.
- Cartwright, N. (1999b). The Dappled World : A Study of the Boundaries of Science, Cambridge University Press.
- Cartwright, N. (2001). "What is Wrong with Bayes Nets?" The monist **84** (2): 242-264.
- Cartwright, N. (2002). "Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward." Br J Philos Sci **53**(3): 411-453.
- Cartwright, N. (2004). "Causation: One Word, Many Things." Philosophy of Science **71**(5): 805-820.
- Cartwright, N. (2007). Hunting Causes and Using Them: Approaches in Philosophy and Economics, {Cambridge University Press}.
- Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford, Oxford University Press.
- Dennett, D. C. (1987). The intentional stance. Cambridge, Mass ; London, MIT Press.
- Dennett, D. C. (1991). Consciousness Explained, Little, Brown & Company.
- 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. New York, Cambridge University Press.
- Dretske, F. (1977). "Laws of Nature." Philosophy of Science **44**: 248-268.
- Dretske, F. (1981). Knowledge and the Flow of Information. Cambridge MA, MIT Press.
- Dretske, F. (1988). Explaining Behavior: Meaning in World of Causes, MIT Press.
- Dretske, F. (2000). Perception, Knowledge and Belief: Selected Essays (Cambridge Studies in Philosophy), Cambridge University Press.
- Drouet, I. (2008). "Is Determinism More Favorable than Indeterminism for the Causal Markov Condition?" Philosophy of Science Assoc. 21st Biennial Mtg.
- Druzdel, M. and H. Simon (1993). Causality in Bayesian belief networks. Proceedings of Ninth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann.
- Ducasse, C. J. (1968). Truth, Knowledge and Causation, London: RKP.
- Eagle, A. (2007). Pragmatic Causation. Causation, Physics, and the Constitution of Reality. H. Price and R. Corry, Oxford University Press: 156-190.
- Ehring, D. (2009). Causal Relata. Oxford Handbook of Causation: 387-413.
- Einstein, A., B. Podolsky, et al. (1935). "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?" Physical Review **47**(10): 777-780.
- Floridi, L. (2008). "The Method of Levels of Abstraction." Mind and Machine **18**(3): 303-329.
- Floridi, L. (2010). The Philosophy of Information. Oxford, Oxford University Press.

- Gillies, D. (2002). "Causality, Propensity, and Bayesian Networks." *Synthese* **132**(1): 63-88.
- Glennan, S. (2009). Singular and General Causal Relations: A Mechanist Perspective.
- Glymour, C. (1999). "Rabbit Hunting." *Synthese* **121**: 55-78.
- Hall, N. (2004). Two Concepts of Causation. *Causation and Counterfactuals*. J. Collins, N. Hall and L. Paul, MIT Press: 225-276.
- Hall, N. (2007). "Structural equations and causation." *Philosophical Studies* **132**: 109-136.
- Hausman, D. (1997). "Causation, Agency, and Independence." *Philosophy of Science* **64**: S15-S25.
- Hausman, D. and J. Woodward (2004). "Modularity and the Causal Markov Condition: A Restatement." *Br J Philos Sci* **55**(1): 147-161.
- Hausman, D. M. and J. Woodward (1999). "Independence, Invariance and the Causal Markov Condition." *British Journal for the Philosophy of Science* **50**(4): 521-583.
- Hitchcock, C. (2002). Probabilistic Causation. *Stanford Encyclopedia of Philosophy*.
- Hitchcock, C. (2004). Do All and Only Causes Raise the Probabilities of Effects? *Causation and Counterfactuals*. J. Collins, N. Hall and L. Paul, MIT Press.
- Hitchcock, C. (2007). What Russell Got Right. *Causation, Physics, and the Constitution of Reality*. H. Price and R. Corry, Oxford University Press.
- Hitchcock, C. (2009). "Structural equations and causation: Six counterexamples." *Philosophical Studies* **144**: 391-401.
- Hitchcock, C. and J. Woodward (2003). "Explanatory Generalizations, Part I: A Counterfactual Account." *Nous* **37**(1): 1-24.
- Hitchcock, C. R. (1995). "Salmon on explanatory relevance." *Philosophy of Science* **62**(2): 304-320.
- Holzer, R. and I. Shimoyama (1998). Bio-robotic Systems Based on Insect Fixed Behavior by Artificial Stimulation. *Robotics Research*. Y. Shirai and S. Hirose, Springer London: 401-407.
- Hoover, K. (2001). *Causality in Macroeconomics*, Cambridge University Press.
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*.
- Humphreys, P. (1989). *The Chances of Explanation: Causal Explanations in the Social, Medical, and Physical Sciences*. Princeton, N.J., Princeton University Press.
- Jeffrey, R. (1967). *Formal Logic - Its Scope and Limits*, Hackett.
- Kant, I. (1787). *Critique of pure reason*.
- Leibniz, G. W. (1695). "Specimen dynamicum."
- Lewis, D. (1973). "Causation." *Journal of Philosophy* **70**.
- Lewis, D. (1979). "Counterfactual Dependence and Time's Arrow." *Noûs* **13**(4): 455-476.
- Machamer, P., L. Darden, et al. (2000). "Singular and General Causal Relations: A Mechanist Perspective." *Philosophy of Science* **18**(1): 1-25.
- Mackie, J. L. (1974). *The Cement of the Universe: A Study of Causation*, Clarendon Press.
- Margolis, E. a. L., Stephen (2014). Concepts. *The Stanford Encyclopaedia of Philosophy*. N. Z. Edward.
- Maroney, O. (2009). Information Processing and Thermodynamic Entropy. *The Stanford Encyclopedia of Philosophy*. E. N. Zalta.
- Maudlin, T. (2007). *The Metaphysics within Physics*, Oxford University Press.
- Mellor, D. H. (1995). *The Facts of Causation*. London, Routledge.
- Menzies, P. (1996). "Probabilistic Causation and the Pre-emption Problem." *Mind* **105**: 85-117.
- Menzies, P. and H. Price (1993). "Causation as a Secondary Quality." *British Journal for the Philosophy of Science* **44**: 187-203.
- Mill, J. (1843). *A System of Logic: Ratiocinative and Inductive*, J. W. Parker.
- Mumford, S. (2009). Causal Powers and Capacities. *The Oxford Handbook of Causation*. H. Beebe, P. Menzies and C. Hitchcock, Oxford University Press.
- Norton, J. (2003). "Causation as Folk Science." *Philosopher's Imprint* **3**.
- Nozick, R. (1969). Newcomb's Problem and Two Principles of Choice. *Essays in Honor of Carl G Hempel*. N. Reschner, Springer.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*, Morgan Kaufmann.

- Pearl, J. (2000). Causality, Cambridge University Press.
- Pearl, J. (2001). "Bayesianism and Causality, or, Why I am only a Half-Baysian." Applied Logic Series, Vol. 24: Foundations of Bayesianism: 19-36.
- Pearl, J. (2003). "Reply to Woodward." Economics and Philosophy **19**: 341-344.
- Price, H. (1992). "The Direction of Causation: Ramsey's Ultimate Contingency." PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association.
- Price, H. (2007). Causal Perspectivalism. Causation, Physics, and the Constitution of Reality. H. Price and R. Corry. Oxford, Oxford University Press.
- Price, H. and P. Menzies (1993). "Causation as a Secondary Quality." British Journal for the Philosophy of Science **44**: 187-203.
- Price, H. and B. Weslake (2009). The Time-Asymmetry of Causation. The Oxford Handbook of Causation. H. Beebe, C. Hitchcock and P. Menzies, Oxford University Press.
- Psillos, S. (2002). Causation and Explanation, Acumen.
- Putnam, H. (1973). "Meaning and Reference." The Journal of Philosophy **70**(19).
- Quine, W. V. (1951). "Two Dogmas of Empiricism." The Philosophical Review **60**: 20-43.
- Quine, W. V. (1960). Word and object. Cambridge, Mass., MIT Press.
- Quine, W. V. (1969). Propositional Objects. Ontological Relativity and Other Essays. New York, Columbia University Press: 139-160.
- Reichenbach, H. (1956). The Direction of Time, University of California Press.
- Russell, B. (1913). "On the Notion of Cause." Proceedings of the Aristotelian Society **13**.
- Russo, F. and J. Williamson (2007). "Interpreting Causality in the Health Sciences." International Studies in the Philosophy of Science **21**(2): 157-170.
- Salmon, W. (1984). Scientific Explanation and the Causal Structure of the World, Princeton University Press.
- Salmon, W. (1994). "Causality Without Counterfactuals." Philosophy of Science **61**: 297-312.
- Schlick, M. (1932). "Causality in Contemporary Physics." Philosophical Papers II: 176-209.
- Searle, J. (1983). Intentionality - An Essay in the Philosophy of Mind, Cambridge University Press.
- Shannon, C. (1949). The mathematical theory of communication, University of Illinois Press.
- Simpson, E. H. (1951). "The Interpretation of Interaction in Contingency Tables." Journal of the Royal Statistical Society, Series B **13**: 238-241.
- Sloman, S. (2005). Causal Models, Oxford University Press.
- Sober, E. (2001). "Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause." Br J Philos Sci **52**(2): 331-346.
- Spirtes, P., C. Glymour, et al. (1993). Causation, Prediction, and Search, Springer-Verlag, New York.
- Spohn, W. (2001). Bayesian Nets Are All There Is to Causal Dependence. Stochastic Dependence and Causality, CSLI.
- Spohn, W. (2012). "Reversing 30 years of discussion: why causal decision theorists should one-box." Synthese **187**: 95-122.
- Steel, D. (2005). "Indeterminism and the Causal Markov Condition." British Journal for the Philosophy of Science **56**(1): 3-26.
- Steel, D. (2006). "Comment On Hausman & Woodward On The Causal Markov Condition." The British Journal for the Philosophy of Science **57**(1): 219-231.
- Suppes, P. (1970). A Probabilistic Theory of Causality, Amsterdam: North-Holland Publishing Company.
- von Weizsäcker, C. (2006). The Structure of Physics (Fundamental Theories of Physics), Springer.
- Weber, B. (2014). Life. The Stanford Encyclopedia of Philosophy. E. N. Zalta.
- Werner, E. (1991). "A united view of information, intention and ability." Proceedings of the Second European Workshop on Modelling Autonomous Agents and Multi-Agent Worlds.
- Wiener, N. (1954). The human use of human beings : cybernetics and society, rev. ed. Boston, Houghton Mifflin.

- Williamson, J. (2005). Bayesian Nets and Causality: Philosophical and Computational Foundations, Oxford University Press.
- Williamson, J. (2009). Probabilistic Theories of Causality. Oxford Handbook of Causation.
- Woodward, J. (1990). "Supervenience and Singular Causal Statements." Royal Institute of Philosophy Supplement 27: 211-246.
- Woodward, J. (2003a). "Critical Notice: Causality by Judea Pearl." Economics and Philosophy 19: 321-340.
- Woodward, J. (2003b). Making Things Happen - A Theory of Causal Explanation, Oxford University Press.
- Woodward, J. (2007). Causation with a Human Face. Causation Workshop 2008. Pittsburgh.
- Woodward, J. (2008). Invariance, Modularity, and All That: Cartwright on Causation. Nancy Cartwright's Philosophy of Science. L. Bovens, C. Hofer and S. Hartmann, Routledge Studies in the Philosophy of Science.
- Woodward, J. (2009). Agency and Interventionist Theories. The Oxford Handbook of Causation. C. H. Helen Beebe, Peter Menzies, Oxford University Press.