Organon F 24 (4) 2017

Guest Editors: Richard David-Rus & Lukáš Bielik

Contents

Organon F 24 (4) 2017: 436-437

Current Topics in the Philosophy of Science

The four papers assembled in this special issue on the philosophy of science *were originally presented at The Inaugural Conference of the East European Network for Philosophy of Science (EENPS) in June 24-26, 2016, in Sofia (Bulgaria). The driving idea behind initiating this network and making such a conference a reality was to bring the philosophers of science working (primarily, though not exclusively) within the former 'Eastern Block' together and enhance their co-operation with other colleagues not just from East European countries but also from other European countries and countries outside Europe. Even though the present papers are but a fragment of the conference's contributions, they clearly witness the quality and fruitfulness of the inaugural conference in particular and the EENPS's agenda in general.*

The papers figuring in this special issue of Organon F *address some of the topics that are currently discussed in the philosophy of science. Namely, Lilia Gurova's "A Reason to Avoid the Causal Construal of Dispositional Explanations" goes in line with non-reductionist (and non-causal) accounts of dispositional explanations. Gurova's paper provides an argument against a general treatment of dispositional explanations as causal ones and offers also a positive reason to account for the distinctiveness of dispositional explanations. Mario Günther's "Learning Conditional and Causal Information by Jeffrey Imaging on Stalnaker Conditionals" presents a unified account of learning uncertain conditionals and causal information modelled by Jeffrey imagining on Stalnaker conditionals. Moreover, his paper also comes with what seems to be a general solution to Douven's (2012) examples and the Judy Benjamin Problem. Duško Prelević's "Hempel's Dilemma and Research Programmes: Why Adding Stances Is Not a Boon" paper addresses two distinct approaches to Hempel's dilemma with respect to physicalism. Prelević argues for construing physicalism rather as a (Lakatosian) research programme than as a stance. Finally, in "The Role of Priors in a Probabilistic Account of 'Best Explanation'", Anton Donchev argues for a probabilistic interpretation of the 'best explanation' in terms of both, likelihoods and priors of alternative hypotheses with an emphasis on taking priors seriously. He also invites us to have a closer*

look at those conditions where Inference to the Best Explanation and Bayesian Confirmation Theory lead to the same (kind of) conclusions.

We believe that these papers make a vivid contribution to the corresponding areas of philosophy of science. We thereby thank the authors for being interested in submitting their papers to this special issue. It was a pleasure for both of us to be involved in the process of preparing and elaborating this issue.

We'd also like to thank Marian Zouhar, the editor-in-chief of Organon F*, for inviting us to edit this issue. We are extremely grateful to our reviewers, for their excellent suggestions and criticisms, and an inspiring co-operation during the whole process. From the beginning, we've been receiving a great support and encouragement from our colleagues in the EENPS' Steering Committee, especially from Daniel Kostić and Lilia Gurová and from the EPSA (European Philosophy of Science Association) Steering Comitee, especially from Stephan Hartmann and Roman Frigg who stimulated the efforts of initating the network and organizing the inaugural conference. Thank you!*

As a special bonus, we are happy to include in this volume also an interview with Professor Joseph Agassi, on occasion of his astonishing jubilee, by Zuzana Parusniková (Institute of Philosophy, Czech Academy of Sciences). This is a magnificent occasion for us and Organon F *to wish professor Agassi all the best, especially strong health and fresh intellect, for the up-coming years.*

Moreover, we are also delighted to include two book reviews – the former by Martin Zach, and the later by Jaroslav Peregrin. Both of them fit nicely into this issue.

So, without any further ado, we invite you to enjoy the reading!

Richard David-Rus
rusdavid@gmail.com

Lukáš Bielik
bielikluc@gmail.com

References

DOUVEN, I. (2012): Learning Conditional Information. *Mind & Language* 27, No. 3, 239-263.

Organon F 24 (4) 2017: 438-455

A Reason to Avoid the Causal Construal of Dispositional Explanations

LILIA GUROVA[1]

ABSTRACT: Those who argue that dispositional explanations are genuine explanations usually construe them as causal explanations. There are several well-known arguments against the causal efficacy of dispositions, but there are as well demonstrations that on some minimal conditions, dispositions could be viewed as causally relevant to the effects which they are taken to explain. Although the latter position is generally tenable, it may be shown that in some important cases it is not a good idea to commit to a causal construal of dispositional explanations. The argument goes as follows: (1) Dispositional explanations are valued for certain specific extra-inferences which they allow us to draw; (2) The causal construal of dispositional explanations can account for some of these extra-inferences only on the assumption that the disposition is a common cause of its manifestations; (3) However, under certain circumstances, the common cause assumption is refuted on theoretical or empirical grounds; Therefore, (4) under certain circumstances, the causal construal of dispositional explanations cannot account for what these explanations are valued for. The latter conclusion is a reason to argue that in some cases at least, the causal construal of dispositional explanations should be avoided.

KEYWORDS: Dispositions – dispositional explanations – extra-inferences – non-causal construal of dispositions – surplus meaning – trait explanations in psychology.

[1] Received: 21 January 2017 / Accepted: 27 July 2017

✉ Lilia Gurova
Department of Cognitive Science and Psychology
New Bulgarian University
21 Montevideo street, 1618 Sofia, Bulgaria
e-mail: lilia.gurova@gmail.com

1. Introduction

Those who argue that dispositional explanations are genuine explanations usually construe them as causal explanations either by assigning a direct causal role to the explanatory dispositions (or to their 'causal bases')[2] or by representing dispositions as parts of, or referring to, larger complexes which play the causal/explanatory role (cf. Hempel 1965; Vanderbeeken & Weber 2002). There are several well-known arguments against the causal construal of dispositions which allegedly demonstrate that dispositions could not play a causal role.[3] There are at the same time demonstrations that on some minimal conditions, dispositions could be viewed as causally relevant to the effects which they are taken to explain (cf. McKitrik 2005). Although the latter position is generally tenable, it can be shown that, in some cases at least, it is not a good idea to commit to the causal construal of dispositional explanations. The main argument goes as follows: (1) Dispositional explanations are valued for certain specific extra-inferences which they allow us to draw and which raise our understanding of explained phenomena; (2) The causal construal of dispositional explanations can account for some of the extra-inferences which the dispositional explanations allow for only on the assumption that the disposition is a common cause of its manifestations; (3) However, under certain circumstances, the common cause assumption is refuted on either theoretical or empirical grounds, or both; Therefore, (4) under certain circumstances, the causal construal of dispositional explanations cannot account for the extra-inferences which these explanations are valued for. The latter conclusion is a reason to argue that the causal construal of dispositional explanations

---

[2] The view that dispositions play a direct causal role has been defended by Nancy Cartwright. She, however, prefers the term 'capacities' instead of 'dispositions' (see Cartwright 1999). Armstrong is a famous defender of the view that dispositions could be assigned a causal role if, and only if, we identify them with their underlying causal bases (see Armstrong, Martin & Place 1996).

[3] The most popular are the so-called "Analyticity Argument" (cf. Armstrong 1968; Mackie 1973; Block 1990; Dardis 1993; Jackson 1995) and the "Exclusion Argument" (cf. Kim 1990; Block 1990); see McKitrk (2005) and Choi & Fara (2012) for a concise presentation of both arguments. Hüttemann (2009) has raised an additional objection against the causal view of dispositions: the latter could not be construed as causes of their manifestations because they do not precede their manifestations in time.

should be avoided if we value the inferential benefits provided by these explanations.

After introducing some preliminaries in section 2, the premises (1) – (3) of the argument against the causal construal of dispositional explanations are discussed in more detail in sections 3 and 4. Section 5 presents the view that dispositional explanations are better viewed as forming a distinct type of explanation and that the explanatory virtues of these explanations build on the extra-inferences which they allow for. The last section 6 summarizes the rationale for the proposal to give up the causal construal of dispositional explanations and analyze them instead in terms of their inferential virtues.

2. Some preliminaries

2.1. Two ways to present a dispositional explanation

A dispositional explanation could be presented in the following way:

(2.1) 'X did B in the situation S because X has the *dispositional property* D'.

Here are some examples of dispositional explanations:

E1: 'The vase broke when it fell on the floor because the vase is fragile.'

E2: 'John hit Mary when she provoked him because John is aggressive.'

X in the sentence (2.1) stands for the object (the agent) which possesses the dispositional property D. B is usually called a *manifestation* of the dispositional property D and the situation S contains, or coincides with, the *stimulus condition,* which activates the manifestation B (cf. Choi & Fara 2012).

In (2.1) the stimulus condition (the situation S) is presented as a part of the *explanandum,* i.e. it precedes the because-clause. It is possible, however, to reformulate (2.1) in the following way:

(2.2) 'X did B because X has the dispositional property D and X was in the situation S.'

In (2.2) the stimulus condition (the situation S) is presented as a part of the *explanans,* i.e. as a part of the because-clause. If we reformulate the examples E1 and E2 in accordance to (2.2), we'll receive the following explanations:

E12: 'The vase broke because it is fragile and it fell on the floor.'

E22: 'John hit Mary because John is aggressive and Mary provoked him.'

Intuitively, the two forms of the presented dispositional explanations (the forms E1 and E2 on the one hand and E12 and E22 on the other hand) have different meanings but without undertaking an additional analysis, we don't seem to have any good reason to prefer the one form instead of the other. The distinction between the forms (2.1) and (2.2) should be taken seriously in any analysis of dispositional explanations, especially by those who construe dispositional explanations as causal explanations. This is because the choice between (2.1) and (2.2) determines what kind of a causal role one is allowed to assign to dispositions. For example, if we construe a dispositional explanation following (2.1) form we must assign a direct causal role to the dispositional property D, but if we choose (2.2) form we may assume that dispositions are not independent causal factors as they play a causal role only in conjunction with the stimulus conditions which have evoked their manifestations.

2.2. Dispositional vs. categorical properties

In both forms of dispositional explanations the role of the stimulus condition S is crucial, although this role, as it was shown above, is different in (2.1) and (2.2). Some are tempted to assume that stimulus conditions are indispensable parts of dispositional explanations because it is a distinctive characteristics of all *dispositional properties* (e.g. the properties of being fragile, soluble, aggressive, vulnerable etc.) that they are always manifested under some stimulus conditions while *categorical properties* (e.g. the properties of being made of wood or glass, being round, having a particular mass etc.) are present under all conditions (cf. Choi & Fara 2012).

Philosophers, however, have never shared a common view about the dispositional/categorical divide. According to the so-called *categoricalists,* a famous representative of which is D. Armstrong (see Armstrong 1997), all real properties are categorical properties. On this view, the terms which seemingly refer to dispositional properties are mere shortcuts for categorical properties. In contrast to *categoricalism*, the view called *dispositionalism* states that (at least some of) the real properties in the world are essentially dispositional, i.e. irreducible to any categorical properties.[4] As we shall see in the next section 2.3, categoricalism and dispositionalism entail different views on the explanatory status of dispositions. The categoricalists usually claim that dispositions, being at best shortcuts for categorical properties, are causally inert and thus non-explanatory. Most of the dispositionalists recognize the causal efficacy of dispositions and their irreducible explanatory role. If, however, we embrace a view where the explanatoriness of dispositions is disentangled from their causal status, we are not anymore forced to take a side in the debate about the proper ontology of dispositions.

2.3. Different views on the explanatory status of dispositional explanations

One can recognize in philosophical literature three major views on the explanatory status of dispositions (cf. Mumford 1998):

(a) Dispositions do not play any explanatory role.

This is the position defended by most of the categoricalists (see above) who insist that dispositions, if they exist at all, are causally inert and, therefore, non-explanatory.

(b) Dispositions play only a heuristic role pointing to where to look for genuine causal explanations.

---

[4] The extreme version of dispositionalism stating that *all* properties are essentially dispositional can be found in Popper (1959) or Mumford (2004). Another kind of extreme dispositionalism is the view that all properties are at once dispositional and qualitative, i.e. categorical (see Heil 2005).

Part of the categoricalists tolerate a temporary use of dispositional explanations in situations where there is a lack of information about the alleged categorical bases of the dispositional properties.

(c) Dispositional explanations are genuine causal explanations.

Those who subscribe to (c), however, differ significantly in their views on how dispositional explanations should be construed as causal explanations. Very few of them, for instance, claim that dispositions possess causal powers.[5] Dispositions are usually taken to have a causal role either in a couple with their causal bases (cf. Armstrong, Martin & Place 1996), or in conjunction with the situations in which they are manifested (see Hempel 1965). On the other hand, those who criticize the causal role of dispositions provide arguments against the possibility for dispositions to play a direct causal role (cf. Armstrong 1968; Mackie 1973; Block 1990; Kim 1990; Dardis 1993; or Jackson 1995).

Besides the three major views (a) – (c), a recent view (d) states that:

(d) Dispositional explanations are genuine non-causal explanations.

According to this view, dispositions do not cause, neither on themselves nor along with other factors, the explained phenomena (cf. Hütteman 2009). Hütteman builds his argument for a non-causal construal of dispositional explanations on the claim that dispositions cannot be construed as causes as they do not precede their manifestations in time. However, as McKitrik (2005) has shown, it is possible to construe dispositional explanations as causal explanations if we embrace a sufficiently weak, "disposition-friendly" criterion for causal relevance which does not include the clause that causes must be independent from their effects and temporally precede them.

In this paper I'll argue against the causal construal of dispositional explanations on a different basis. It will be demonstrated that even in cases where the causal construal of dispositional explanations is possible, this construal leads to assumptions which are unacceptable for theoretical and empirical reasons.

---

[5] Nancy Cartwright is a famous defender of this view – see Cartwright (1999); see also Heil (2005).

3. What dispositional explanations are good for?

According to an influential view (see Quine 1969; Armstrong 1973), dispositional explanations are at best (temporary) substitutes for genuine causal explanations. I am not going to discuss here the arguments for this view.[6] It suffices to note, that if there are dispositional explanations which are irreducible to non-dispositional ones in disciplines as diverse as physics and psychology,[7] these explanations probably play a role which exceeds that of a substitute and which deserves a more careful analysis.

Let us consider again the simple example E1: 'The vase broke when it fell on the floor because the vase is fragile.' A categoricalist would argue that the explanation E1 could be reduced to the following one:

E1*: 'The vase broke when it fell on the floor because the vase is made of glass and the crystalline structure of glass makes it fragile.'

At first glance, E1* does not only serve as a good substitute for E1 but it even looks a "deeper" explanation as far as in addition to explaining why this particular vase broke, it explains as well why the vase is fragile (the vase is fragile because of its crystalline structure). From this perspective, E1* does look superior to E1.

But let's take a different perspective. Let's ask about what one can infer from each of these explanations. Given E1*, we are entitled to expect that not only this particular vase will break if it falls on the floor but any object of a similar mass, which is made of glass having the same crystalline structure will break too, if it falls on the floor from the same or a bigger height.

---

6 A simplified form of the standard argument goes as follows: all genuine explanations are causal explanations; dispositions are causally inert (although they can refer to, or be grounded in, causally efficient categorical properties); therefore, dispositions in themselves could not play an explanatory role.

7 Quantum mechanics, as it is understood today, seems to leave no room for non-dispositional interpretations of the properties of the fundamental particles. See, e.g., Bigaj (2012) for a nice explanation of why such properties as the spin of an electron are best understood as irreducible dispositional properties. In a similar vein, many personality psychologists and philosophers of psychology view personality traits as dispositions which are not reducible to neurophysiological, genetic or other biological or physical categories (see Wiggins 1973; Cervone 2004; Borsboom 2015; Gurova 2017).

Given E1, however, we are entitled to expect much more. For instance, we are justified to suppose that a fragile ceramic vase (or another fragile ceramic object) will break if we drop it on the floor. The same should be expected about a fragile match house, or a fragile egg, if they are dropped on the floor, although they do not have a crystalline structure like the glass vase from E1. In other words, the dispositional explanation E1 has a bigger *inferential content* (i.e. it allows for a larger number of inferences to be drawn) than its non-dispositional substitute E1*. One can ask here, why should we care about the explanations' inferential content? What follows may count as an answer of this question: If we agree with the widely supported claim that the primary goal of any explanation is to enhance our understanding of the explained phenomenon (cf. Friedman 1974; Lipton 2004), and if we agree that a distinctive mark of understanding is the ability to go "beyond the information given" (see Bruner 1957), then we may also agree that the inferential content of a given explanation (the extra-inferences which this explanation allows for) is a good measure of the explanation's capacity to lead us "beyond the information given".

In fact, in many areas where dispositional explanations are used, they are appreciated exactly for their capacity to suggest inferences which go "beyond the information given". In psychology, for instance, many insist that dispositional explanations carry "surplus meaning" where "surplus meaning" is just another term referring to the extra-inferences which a given explanation allows for. The following citation from two eminent personality psychologists is representative for the latter view:

> [an] explanation becomes useful only when it provides surplus meaning and allows inferences which go beyond the observed data. … Traits are defined as enduring dispositions, and are hypothesized to be related to outcome variables; thus trait explanation carries with it the implication that long-term predictions can be made. (McCrae & Costa 1995, 243)

Indeed, given E2, i.e. given the knowledge that 'John hit Mary when she provoked him because John is aggressive', we may reliably predict that John has probably attributed a hostile intention to Mary's provocation, as well as we may expect that John will not hesitate to harm somebody if John sees the harm as a means to achieving his goals.

To summarize, dispositional explanations are valued for two types of inference which they allow for. Given the dispositional explanation (2.1), for example, we may derive that:

(3.1) X is expected to do B_1 in S_1, or B_2 in S_2, or … B_n in S_n, if B_1 – B_n are known possible manifestations of the dispositional property D, which X possesses.

Let's call the inferences like (3.1) 'inferences to different manifestations'. Given (2.1.), we are also entitled to assume that:

(3.2) Any object (agent) Y, which is different from X, will do B* in S* if he possesses the dispositional property D.

In (3.2) B* and S* stand for any manifestation and stimulus condition which are identical or similar to B and S. Let's call the (3.2) like inferences 'inferences to different objects (agents)'.

Let's see now what happens with these two types of inference when we construe dispositional explanations as causal explanations.

4. What happens when dispositional explanations are construed as causal explanations?

In the previous section, we saw that, at least in some areas, the higher inferential content of dispositional explanations has been recognized as their main explanatory virtue. Now we have to see what happens when we try to account for this virtue by assigning a causal role to the explanatory dispositions.

4.1. The inferences to different manifestations

Let's consider again the example E2: 'John hit Mary when she provoked him because John is aggressive'; and let's assign the following values to the variables B, S and D:

B = 'John hit Mary.'
S = 'Mary provoked John.'
D = 'John is aggressive.'

Then if we use one of the "disposition-friendly" criteria for causal relevance (cf. McKitrick 2005), e.g. the probabilistic criterion,[8] the following inequality must be satisfied in order to claim that the disposition D is causally related to the *explanandum* (B, S):

(4.1) $P(B, S|D) > P(B, S|\text{non-}D)$[9]

Let's assume now that (4.1) is satisfied, i.e. the disposition D (John's aggressiveness) is causally related to the *explanandum* (B, S) ('John hit Mary when she provoked him.'). As it was shown in section 3, explanations like E2 are valued because they allow us to predict other behavioral acts of the agent who possesses the explanatory disposition D. Let's now, for the sake of simplicity, take into account the following prediction about John's understanding of Mary's intentions in the same situation S:

C: 'John attributed hostile intentions to Mary.'

Then the explanation of C would be:

E2*: 'John attributed hostile intentions to Mary when she provoked him because John is aggressive.'

In order to view E2* as a valid causal explanation, the following inequality must hold:

(4.2) $P(C, S|D) > P(C, S|\text{non-}D)$

If both (4.1) and (4.2) are satisfied, taken together, they imply that D is a common cause of B and C. Being a common cause, D screens off the correlation between its two manifestations. However, the correlation between B and C is the only empirical fact we know for sure. There is a plenty of evidence e.g. for the existence of a direct connection between the various

[8] The inferences that follow hold even if we use a different criterion for causal relevance. The probabilistic criterion has been chosen only because it is considered "disposition-friendly" (McKitrick 2005), i.e. it is not expected to bring additional problems for the causal construal of dispositional explanations.

[9] The inequality (4.1) should be read as follows: the probability of the appearance of B in S given D is higher than the probability of the appearance of B in S given non-D.

violent reactions to particular provocations and the attribution of a negative intention to the provocateur (see Dodge 2006). However, when we construe the dispositional explanations as common cause explanations, this construal forces us to assume that the correlations between the manifestations of the dispositional property are spurious rather than standing for real connections. On the other hand, there is little to no evidence that specific biological structures exist that might play the role of the alleged common causes of the correlated behavioral acts (see Kehoe et al. 2012). In addition, theoretical considerations have been raised against the plausibility of the hypothesis that such biological common causes of traits' manifestations exist.[10] The situation in personality psychology thus reminds us about the situation in quantum mechanics where the assumption that dispositional properties like the spin of an electron are grounded in (still unknown) categorical physical properties led to theoretical conceptions which are not supported by the available experimental evidence as well as by theoretical results such as the Bell's theorem.[11]

Nothing significantly changes if we use the E22 form of the explanation: 'John hit Mary because John is aggressive and Mary provoked him.' In this case the following equations must hold in order to construe E22 and E22* as causal explanations, in accordance with the disposition-friendly probabilistic criterion for causal relevance:

(4.3) $P(B|S, D) > P(B|S, \text{non-}D)$

(4.4) $P(C|S, D) > P(C|S, \text{non-}D)$

[10] Lamiell (1987) was probably the first who tried to draw attention to the fact that the behaviorally defined traits have been elicited using statistical methods such as factor analysis in between subject studies which do not allow us to infer that the elicited structure exists within the particular subjects; see also Rorer (1990); Borsboom et al. (2003); Cervone (2004); and Borsboom (2015). A different argument against the interpretation of traits as hidden causes of their observable manifestations was raised by Wiggins (1973). His argument builds on the premise that the considerations involved in drawing the boundaries between the different trait categories reflect some socially important distinctions rather than biological ones.

[11] A series of proofs known under the label "the Bell's theorem" demonstrate that local hidden variables cannot (causally) account for the quantum measurement correlations, which the quantum mechanics predicts – see Bell (1964); see Myrvold (2016) for a recent discussion on the Bell's theorem's implications.

Again, the common cause (S, D) screens off the correlation between B and C, which in this particular example is unacceptable for both empirical and theoretical reasons, as it was shown above. There is empirical evidence for the connection between hostile attributions and aggressive reactions to provocations and there is a theoretical model built on this evidence which has been well confirmed (cf. Dodge 2006). At the same time there is no convincing evidence that the alleged common causes stand for real biological structures and there are good theoretical arguments against such hypotheses.

4.2. The inferences to different objects (agents)

Let's go back again to the example E2: 'John hit Mary when she provoked him because John is aggressive' and remind that this dispositional explanation allows us to predict that another person, say Billy, who has the same dispositional property (has an aggressive personality) will act in a similar way B* in a situation S* which is similar to S.

Let's assume that B* stands for 'Billy offended Sally', S* stands for 'Sally provoked Billy' and D* stands for 'Billy is aggressive'. Then if we apply again the probabilistic criterion for causal relevance to the following dispositional explanation

E3: Billy offended Sally when she provoked him because Billy is aggressive

we'll receive

(4.5) $P(B^*, S^*|D^*) > P(B^*, S^*|\text{non-}D^*)$

As far as D* is similar but not identical to D (i.e. we do not have good reasons to assume that John's aggressiveness is exactly the same as Billy's aggressiveness), we cannot say that the two events (B, S) and (B*, S*) have a common cause, we can only say that they have *similar* causes. Therefore, we are not forced here to screen off the correlations between (B, S) and (B*, S*), but even if we were, that would not create any problem because no one expects a direct causal link between the events 'John hit Mary when she provoked him' and 'Billy offended Sally when she provoked him'.

To sum up, the causal construal of dispositional explanations leads to a common cause assumption only when we try to account for the inferences to different manifestations. In some of these cases the implied common cause assumption goes against the available empirical data and theoretical considerations. However, the causal construal does not lead to any serious problems when we interpret causally the inferences to different objects (agents). Probably because the causal account of dispositional explanations does not face serious problems most of the time, many are tempted to assume that it is generally tenable but it is not as the analysis of the causal construal of the inferences to different manifestations has shown.

One can ask at this point: but what are we left with when we abandon the causal construal of dispositional explanations for the reasons stated above? Or asking the same question in slightly different words, what in the end is the proper construal of dispositional explanations? In the next section I'll try to defend the view that dispositional explanations are better viewed as a distinct type of explanation, which has to be analyzed in terms of the extra-inferences (inferences to different manifestations and inferences to different objects/agents) that these explanations allow us to draw.

5. Dispositional explanations as a distinct type of explanation

The main views of scientific explanation in the philosophy of science today[12] set different requirements for the *explanans* and (or) for the relation between the *explanans* and the *explanandum* (see Table 1 below).

Dispositional explanations could not be easily subsumed under either of the views presented in Table 1. They, for example, do not explicitly refer to any laws and some dispositionalists (e.g. Mumford 2004) have even argued that they do not need to. Thus, unless we make some problematic stipulations, dispositional explanations could not be construed as covering-law explanations. We have already shown why, in some cases at least, dispositional explanations should not be treated as causal explanations. But

[12] See Skow (2016) for a recent review.

View on explanation	Requirements about the *explanans* and the *explanans/explanandum* relation
The covering-law model	The *explanans* contains at least one deterministic or probabilistic law or a law-like sentence. The *explanans* implies, deductively or inductively, the *explanandum.*
The causal theories	The *explanans* stands for events (states, processes etc.) which are causally relevant to the events (states, processes etc.) represented by the *explanandum.*
The unificationist view	The *explanans* implies, deductively or inductively, different *explananda.*

Table 1. The specific requirements for *explanans* and the *explanans/explanandum* relation that have been set by the three major views on explanation.

what about the unificationist account? On the one hand, dispositional properties do play a unifying role with respect to their different manifestations and thus an explanation which refers to such a property unifies different *explananda.* On the other hand, as Skow (2016) has already noted, unification seems to be a *consequence* of having an explanation rather than a *condition* that must be satisfied in order to have an explanation. Indeed, in the case of dispositional explanations, we must have an explanation already stated in either of the forms (2.1) or (2.2) in order to be able to draw the inferences to multiple manifestations that bring unification of different *explananda*. In addition, unification does not account for the specifics of dispositional explanations, e.g. for the important role of the stimulus conditions, as well as for the two types of extra-inferences that are constitutive for the explanatory benefits of dispositional explanations.

For the reasons stated above, it is safe to conclude that dispositional explanations are better viewed as a distinct form of explanation that satisfies the following conditions:

(i) The explanation can be presented in one of the forms (2.1) or (2.2), which means that the *explanans* must refer to a dispositional property D, and either the *explanans* or the *explanandum* must contain information about the stimulus condition S;

(ii) The explanation should allow for extra-inferences to different manifestations (3.1) as well as for inferences to different objects/agents (3.2) and these extra-inferences must have meaningful (and possibly true) interpretations.

The main advantages of the view that dispositional explanations form a distinct type of explanation, which satisfies the conditions (i) and (ii) are that this view makes salient the specific explanatory virtues of dispositional explanations and allows for analyzing and comparing different concrete explanations in terms of these virtues.

6. Conclusions

Dispositional explanations are most valued for the extra-inferences, which they allow for. The explanation of a particular phenomenon, or a behavioral act, which relates the explained phenomenon (behavioral act) to a particular disposition, allows us to predict that other manifestations of the same disposition may be expected in the same or in a different stimulus condition. Such predictions are called here "inferences to different manifestations". Dispositional explanations allow us to predict as well that a different object/agent possessing the same dispositional property will exhibit similar manifestations, i.e. they allow for what was called here "inferences to different objects/agents". The causal construal of dispositional explanations successfully accounts for the inferences to different objects/agents but it fails to account properly for the inferences to different manifestations. This is because the causal construal of dispositional explanations entails that the explanatory dispositions are common causes of their manifestations. As far as the common causes screen-off the correlations

between their effects, the common cause assumption leads to conclusions which, in some cases at least, are either unacceptable for theoretical reasons or incompatible with the available empirical evidence, or both. Such unfortunate consequences of the common cause assumption are a serious reason to argue that the causal construal of dispositional explanations should be avoided, or applied with a great caution, and that dispositional explanations are better and safely analyzed in terms of their specific inferential virtues which present them as a distinct type of explanation.

Acknowledgments

I would like to thank the two anonymous reviewers for their valuable comments on the first draft of this paper. They helped a lot to strengthen and make clearer the main argument. An earlier version of this argument was presented at the Inaugural conference of the East European Network for Philosophy of Science (Sofia, June 24-26 2016), as part of the symposium "Explanation and understanding in science". I am indebted for the thoughtful discussion to all those who took part in it but my special thanks are for the co-organizers of the symposium (Lukáš Bielik, Richard David-Rus and Daniel Kostić) who made all these possible.

References

Armstrong, D. M. (1968): *A Materialist Theory of the Mind*. London: Routledge.

Armstrong, D. A. (1973): *Belief, Truth, and Knowledge*. Cambridge: Cambridge University Press.

Armstrong, D. A. (1997): *A World of States of Affairs*. Cambridge: Cambridge University Press.

Armstrong, D. M., Martin, C. B. & Place, U. T. (1996): *Dispositions: A Debate*. London: Routledge.

Bell, J. S. (1964): On the Einstein Podolsky Rosen Paradox. *Physics* 1, No. 3, 195-200.

Bigaj, T. (2012): Ungrounded Dispositions in Quantum Mechanics. *Foundations of Science* 17, 205-221.

Block, N. (1990): Can the Mind Change the World? In: Boolos, G. S. (ed.): *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press, 137-170.

Borsboom, D. (2015): What is Causal about Individual Differences? A Comment on Weinberger. *Theory and Psychology* 25, No. 3, 362-368.

BORSBOOM, D., MELLENBERGH, G. J., VAN HEERDEN, J. (2003): The Theoretical Status of Latent Variables. *Psychological Review* 110, 203-219.

BRUNER, J. (1957): Going beyond the information given. In: Bruner, J., Brunswik, E., Festinger, L., Heider, F., Muenzinger, K. F., Osgood, C. E. & Rapaport, D. (eds.): *Contemporary Approaches to Cognition.* Cambridge (Mass.): Harvard University Press, 41-69.

CARTWRIGHT, N. (1999): *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

CERVONE, D. (2004): Personality Assessment: Tapping the Social-Cognitive Architecture of Personality. *Behavioral Therapy* 35, 113-129.

CHOI, S. & FARA, M. (2012): Dispositions. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy*. (Spring 2016 Edition), available at: https://plato.stanford.edu/archives/spr2016/entries/dispositions/

DARDIS, A. (1993): Sunburn: Independence Conditions on Causal Relevance. *Philosophy and Phenomenological Research* 53, No. 3, 577-598.

DODGE, K. A. (2006): Translational Science in Action: Hostile Attributional Style and the Development of Aggressive Behavior Problems. *Development and Psychopathology* 18, No. 3, 791-814.

FRIEDMAN, M. (1974): Explanation and Scientific Understanding. *The Journal of Philosophy* 71, 5-19.

GUROVA, L. (2017): Are Causal Accounts of Explanation Always Useful? In the Case of Personality Trait Explanations They Are Probably Not. In: Massimi, M., Romejn, J.-W. & Schurz, G. (eds.): *EPSA15 Selected Papers*. Cham: Springer, 167-177.

HEIL, J. (2005): Dispositions. *Synthese* 144, No. 3, 343-356.

HEMPEL, C. (1965): *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: The Free Press.

HÜTTEMANN, A. (2009): Dispositions in Physics. In: Damschen, G., Schnepf, R. & Stüber, K. R. (eds.): *Debating Dispositions*. Berlin: Walter de Gruyter, 223-237.

JACKSON, F. (1995): Essentialism, Mental Properties and Causation. *Proceedings of the Aristotelian Society* 95, 253-268.

KEHOE, E. G., TOOMEY, J. M., BALSTERS, J. H., BOKDE, A. L. W. (2012): Personality Modulates the Effects of Emotional Arousal and Valence on Brain Activation. *Social Cognitive and Affective Neuroscience* 7, 858-870.

KIM, J. (1990): Explanatory Exclusion and the Problem of Mental Causation. In: Villanueva, E. (ed.): *Information, Semantics, and Epistemology*. Oxford: Blackwell, 36-56.

LAMIELL, J. T. (1987): *The Psychology of Personality: An Epistemological Inquiry*. New York: Columbia University Press.

LIPTON, P. (2004): *Inference to the Best Explanation*. 2nd ed. London: Routledge.

McCrae, R. & Costa, P. (1995): Trait Explanations in Personality Psychology. *European Journal of Personality* 9, 231-252.

Mackie, J. L. (1973): *Truth, Probability and Paradox*. Oxford: Oxford University Press.

McKitrik, J. (2005): Are Dispositions Causally Relevant? *Synthese* 144, 357-371.

Myrvold, W. C. (2016): Lessons of Bell's Theorem: Nonlocality, Yes; Action at a Distance, not Necessary. In: Bell, M. & Gao, S. (eds.): *50 Years of Bell's Theorem*. Cambridge: Cambridge University Press, 237-260.

Mumford, S. (1998): *Dispositions.* Oxford: Oxford University Press.

Mumford, S. (2004): *Laws in Nature*. London: Routledge.

Popper, K. (1959): *The Logic of Scientific Discovery*. London: Hutchinson & Co.

Rorer, L. G. (1990): Personality Assessment: A Conceptual Survey. In: Pervin, L. A. (ed.): *Handbook of Personality: Theory and Research.* New York: Guilford, 693-720.

Skow, B. (2016): Scientific Explanation. In: Humphreys, P. (ed.): *The Oxford Handbook of Philosophy of Science*. Oxford: Oxford University Press, 524-543.

Quine, V. W. O. (1969): Natural Kinds. In: Rescher, N. (ed.): *Essays in Honor of Carl G. Hempel.* Dordrecht: D. Reidel, 5-23.

Vanderbeeken, R. & Weber, E. (2002): Dispositional Explanations of Behavior. *Behavior & Philosophy* 30, 43-59.

Wiggins, J. S. (1973/1997): In Defense of Traits. In: Hogan, R., Johnson, J. & Briggs, S. (eds.): *Handbook of Personality Psychology.* San Diego: Academic Press, 95-113.

Organon F 24 (4) 2017: 456-486

Learning Conditional and Causal Information by Jeffrey Imaging on Stalnaker Conditionals

MARIO GÜNTHER[1]

ABSTRACT: We show that the learning of (uncertain) conditional and/or causal information may be modelled by (Jeffrey) imaging on Stalnaker conditionals. We adapt the method of learning uncertain conditional information proposed in Günther (2017) to a method of learning uncertain causal information. The idea behind the adaptation parallels Lewis (1973c)'s analysis of causal dependence. The combination of the methods provides a unified account of learning conditional and causal information that manages to clearly distinguish between conditional, causal and conjunctive information. Moreover, our framework seems to be the first general solution that generates the correct predictions for Douven (2012)'s benchmark examples and the Judy Benjamin Problem.

KEYWORDS: Causal dependence – Douven's examples – Imaging – Judy Benjamin Problem – Learning – Stalnaker conditional.

[1] Received: 21 January 2017 / Accepted: 28 June 2017

✉ Mario Günther
Ludwig-Maximilian-University Munich
Munich Center for Mathematical Philosophy
& Graduate School for Systemic Neurosciences
Chair of Logic and Philosophy of Language
Geschwister-Scholl-Platz 1, D-80539 Munich, Germany
e-mail: Mario.Guenther@campus.lmu.de

1. Introduction

"How do we learn conditional information?" Igor Douven et al. present this question for consideration in a series of papers (cf. Douven & Dietz 2011; Douven & Romeijn 2011; Douven 2012; Pfeifer & Douven 2014, especially section 6). Douven (2012) contains a survey of the available accounts that model the learning of conditional information. The survey comes to the conclusion that a general account of probabilistic belief updating by learning (uncertain) conditional and causal information is still to be formulated. Pfeifer & Douven (2014) analyses the state of the art even more pessimistically by writing that "no one seems to have an idea of what an even moderately general rule of updating on conditionals might look like," even if we restrict the scope of the account to indicative conditionals (Pfeifer & Douven 2014, 213). We aim to provide such a general account of updating that unifies the learning of (uncertain) conditional and causal information.

In Günther (2017), we proposed a method of learning conditional information. We have shown that the predictions of the proposed method align with the intuitions in Douven (2012)'s benchmark examples and can generate predictions for the Judy Benjamin Problem. In this paper, we adapt the method of learning conditional information to a method of learning causal information. The adapted method allows us to causally conceive of the information conveyed by the conditionals uttered in Douven's examples and the Judy Benjamin Problem.

It may come as a surprise that we propose an account of learning that involves (Jeffrey) imaging. After all, the standard view on learning that α is Bayesian updating on α, while David Lewis's imaging on α is widely conceived of as modeling the supposition of α. But learning a conditional may – according to the suppositional view on conditionals – be interpreted as learning what is true under a supposition (about which we may be uncertain). In particular, learning the conditional "If α, then γ" is thus equivalent to learning the conditional information that γ is the case under the supposition that α is the case.

Douven aims to provide an account of learning conditional information that is an empirically adequate account of human reasoning. Douven & Verbrugge (2010) submitted the thesis whether the acceptability of an indicative conditional 'goes by' the conditional probability of its consequent

given the antecedent to empirical testing, and claim that the experiments speak against the thesis.[2] Their results indicate that conditional probabilities do not correspond to probabilities of conditionals, which was proved by Lewis (1976), if conditionals are understood as Stalnaker conditionals. Those formal and empirical results obviously provide a severe challenge for Bayesian analyses of learning conditionals, where conditional probabilities usually take center stage.

Moreover, Zhao et al. (2012) obtained empirical results that indicate a fundamental difference between supposing and learning. In particular, supposing a conditional's antecedent α seems to have less impact on the credibility of the consequent γ than learning that α is true. We will provide a framework that allows us both, to distinguish between the learning of 'factual' and conditional information and to generate empirically testable predictions.

In Section 2, we introduce Douven's desideratum for accounts of learning (uncertain) conditional information. His own proposal is based on the explanatory status of the antecedent. In Section 2.1, we sketch his argumentation against the method of imaging on the Stalnaker conditional as an account of learning conditional information. The reason for Douven's dismissal of the method is that the rationality constraints of Stalnaker models are not sufficient to single out a model, which may count as a representation of a belief state.

In Section 3, we review the method of learning (uncertain) conditional information proposed in Günther (2017), where we showed that Douven's dismissal is unjustified. We met Douven's challenge for possible worlds models by imposing two additional constraints: interpreting the meaning of a Stalnaker conditional in a minimally informative way and supplementing the analysis by a default assumption. Moreover, we generalised Lewis's imaging method in order to account for uncertain information as well.

In Section 4, we adapt the method of learning conditional information to a method of learning causal information. The adaptation is inspired by Lewis's notion of causal dependence and replaces the default assumption by the assumption that the antecedent makes a difference. In Section 4.1, we apply our adapted method of learning causal information to Douven's

[2] The 'goes by' is Lewis's formulation that may be found in Lewis (1976, 297).

examples and the Judy Benjamin Problem. In Section 5, we formally implement Douven's idea concerning the explanatory status of the antecedent within our framework.

2. Douven's account of learning conditional information via the explanatory status of the antecedent

Igor Douven propounds a broadly Bayesian model of learning conditional information. As the standard Bayesian view of learning, Douven's account assumes that learning the unnested indicative conditional "If α, then γ" implies that the posterior degree of belief for γ given α is set to approximately 1, i. e. $P^*(\gamma \mid \alpha) \approx 1$. In contrast to standard Bayesian epistemology, explanatory considerations play a major role in his model of updating on conditionals.

Douven proposes a desideratum for any account of learning conditional information, viz. a criterion that determines whether an agent raises, lowers, or leaves unchanged her degree of belief $P(\alpha)$ for the antecedent upon learning a conditional.

He even writes that we "should be [...] dissatisfied with an account of updating on conditionals that failed to explain [...] basic and compelling intuitions about such updating, such as, in our examples" (Douven 2012, 3). Douven's methodology consists in searching for an updating model that accounts for our intuitions with respect to three examples, the Sundowners Example, the Ski Trip Example and the Driving Test Example. The three examples represent the classes of scenarios, in which $P(\alpha)$ should intuitively remain unchanged, be increased and decreased, respectively. He dismisses any method of learning conditional information that cannot account for all of the three examples. He emphasises that no single account of learning uncertain conditional and/or causal information is capable of solving all of his examples. Taking the examples as benchmark, he also dismisses the Stalnaker conditional as a tool to model the learning of conditional information.

The core hypothesis of Douven's account is that the change in explanatory quality or 'explanatory status' of the antecedent α during learning the information results in a change of the degree of belief for α. If the explanatory status of α goes up, that is α explains γ well, then the degree of belief

after learning the conditional increases, i. e. $P^*(\alpha) > P(\alpha)$; if the explanatory status of α goes down, $P^*(\alpha) < P(\alpha)$; if the explanatory status remains the same, a variant of Jeffrey conditioning is applied that has the property that $P^*(\alpha) = P(\alpha)$. Following Richard Bradley, Douven calls this Jeffrey conditioning over a restricted partition 'Adams conditioning on $P^*(\gamma \mid \alpha) \approx 1$'.[3]

Douven and Romeijn proposed a solution to the Judy Benjamin Problem. The problem indicates that the revision method that minimises the Kullback-Leibler divergence leads to counterintuitive results for learning uncertain conditional information. Their solution uses the variant of Jeffrey conditioning mentioned above. However, their proposed method fails to account for examples where the probability of the antecedent is supposed to change, since it has the invariance property that $P^*(\alpha) = P(\alpha)$, for all α, and thus disqualifies as a general account of learning conditional information (cf. Douven & Romeijn 2011, 648-655; Douven 2012, 9-11).

2.1. Douven's dismissal of the Stalnaker conditional

Douven claims that Stalnaker conditionals are not suited to model the learning of conditional information. He argues for this claim by pointing out that a learning method based on the Stalnaker conditional "makes no predictions at all about any of our examples" (Douven 2012, 7). The cited reason is that we would not be able to exclude certain Stalnaker models as rational representation of a belief state.

Douven provides three possible worlds models for his point. Each model consists of four worlds such that all logical possibilities of two binary variables are covered. He observes that imaging on "If α, then β" interpreted as a Stalnaker conditional has different effects: in model I, the probability of the antecedent α, i. e. $P(\alpha)$ decreases; in model II, $P(\alpha)$ remains unchanged; and in model III $P(\alpha)$ increases. According to Douven this flexibility of the class of possible world models is a problem rather than an advantage, since there would be no rationality constraints to rule out certain models as rational representations of a belief state.

---

[3] The partition is restricted according to the odds for the consequent of the learned conditional. For details, see Bradley (2005, 351-352); and Douven & Romeijn (2011, 650-653).

Consider a scenario of the class, where the antecedent remains unchanged (e.g. the Sundowners Example). The problem is, so Douven argues, that there are no criteria to exclude models I and III as rational representations of a belief state, in which $P(\alpha)$ should not change. In Douven's words:

> In fact, to the best of my knowledge, nothing said by Stalnaker (or Lewis, or anyone else working on possible worlds semantics) implies that, supposing imaging is the update rule to go with Stalnaker's account, models I and III [...] could not represent the belief state of a rational person; [...] In short, interpreting "If A, B" as the Stalnaker conditional and updating on it [...] by means of imaging offers no guarantee that our intuitions are respected about what should happen – or rather not happen – after the update [...]. Naturally, it cannot be excluded that some of these models – and perhaps indeed all on which [...] [the degree of belief in the antecedent] changes as an effect of learning [the conditional] – are to be ruled out on the basis of rationality constraints that I am presently overlooking, perhaps ones still to be uncovered, or at least still to be related to possible worlds semantics as a tool for modelling epistemic states. It is left as a challenge to those attracted to the view considered here to point out such additional constraints. (Douven 2012, 8-9)

In Günther (2017), we met the challenge Douven mentions in the quote. We discovered two constraints that singled out Stalnaker models that plausibly represent the belief states in Douven's benchmark examples. Imposing the two additional constraints amounts to interpreting the meaning of a Stalnaker conditional in a minimally informative way and supplementing the analysis by a default assumption.

3. Review of the Method of Learning Conditional Information by Jeffrey Imaging on Stalnaker Conditionals

Günther (2017) puts forward a method of learning conditional information by Jeffrey imaging on Stalnaker conditionals. The learning method may be summarised as follows. (i) We model an agent's belief state as a Stalnaker model. (ii) The agent learns conditional information by (ii).(a)

interpreting the received conditional information as a Stalnaker conditional; (ii).(b) constraining the similarity order by the meaning of the Stalnaker conditional in a minimally informative way and respecting the default assumption; and (ii).(c) updating her degrees of belief by Jeffrey imaging on this Stalnaker conditional (together with further contextual information, if available).

We outline the method of learning conditional information by presenting its constituents, i. e. the semantics of the Stalnaker conditional, Jeffrey imaging and the meaning of 'minimally informative'. Afterwards, we put the constituents together.

3.1. The Stalnaker conditional

The idea behind a Stalnaker conditional may be expressed as follows: a Stalnaker conditional $\alpha > \gamma$ is true at a world w iff γ is true in the most similar possible world w' to w, in which α is true (cf. Stalnaker 1975).[4] We denote the set of possible worlds that satisfies a formula α by $[\alpha]$. Thereby, we identify the set $[\alpha]$ with the proposition expressed by α. In symbols, $[\alpha] = \{w \in W \mid w(\alpha) = 1\}$, where each w of the set of worlds under consideration W may be thought of as a Boolean evaluation.

A Stalnaker conditional is evaluated with respect to a Stalnaker model, i. e. a model of possible worlds where each world w is equipped with a total order such that w is the unique center of the respective order and, for non-contradictions α, it is guaranteed that there exists a unique most similar world $min_{\leq w}\,[\alpha]$ from w that satisfies α. The accessibility relation of a Stalnaker model is reflexive and connective.

Let us state more precisely the meaning of a Stalnaker conditional using the notations just introduced. "If α, then γ" denotes according to Stalnaker's proposal the set of worlds (or equivalently the proposition) containing each world whose most similar α-world is a world that satisfies γ. In

[4] Note that Stalnaker's theory of conditionals aims to account for both indicative and counterfactual conditionals. We set the complicated issue of this distinction aside in this paper. However, we want to emphasise that Douven's examples and the Judy Benjamin Problem only involve indicative conditionals.

symbols, $[\alpha > \gamma] = \{w \mid w \models \alpha > \gamma\} = \{w \mid \min_{\leq w}[\alpha] = \emptyset \text{ or } \min_{\leq w}[\alpha] \models \gamma\}$.[5]

Finally, note that any Stalnaker model validates the principle called 'Conditional Excluded Middle' according to which $(\alpha > \gamma) \vee (\alpha > \neg\gamma)$. The reason for the validity of Conditional Excluded Middle is that, for any $w \in W$, the single most similar α-world $\min_{\leq w}[\alpha]$ is either a γ-world, or else a $\neg\gamma$-world. This principle will come in handy when modeling the learning of uncertain information.

In the next section, we introduce Lewis's imaging method, which we will generalise in the subsequent section.

3.2. Lewis's imagining

David Lewis developed a probabilistic updating method called 'imaging' (cf. Lewis 1976). We introduce a notational shortcut: for each world w and each (possible) antecedens α, $w_\alpha = \min_{\leq w}[\alpha]$ be the most similar world of w such that $w_\alpha(\alpha) = 1$. Invoking the shortcut, we can then specify the truth conditions for Stalnaker's conditional operator $>$ as follows:

(1) $w(\alpha > \gamma) = w_\alpha(\gamma)$, if α is possible.[6]

Definition 1. Probability Space over Possible Worlds

We call $\langle W, \wp(W), P\rangle$ a probability space over a finite set of possible worlds W iff

(i) $\wp(W)$ is the set of all subsets of W,

(ii) and $P : \wp(W) \mapsto [0, 1]$ is a probability measure, i.e.

(a) $P(W) = 1$, $P(\emptyset) = 0$, and

(b) for all $X, Y \subseteq W$ such that $X \cap Y = \emptyset$, $P(X \cup Y) = P(X) + P(Y)$.

As before, we conceive of the elements of $\wp(W)$ as propositions. We define, for each α, $P(\alpha) = P([\alpha])$. We see that W corresponds to an arbitrary tautology denoted by $\top$ and $\emptyset$ to an arbitrary contradiction denoted by $\bot$.

---

5 See Günther (2017) for a more thorough presentation of Stalnaker models. See Stalnaker & Thomason (1970) for Stalnaker and Thomason's original presentation of the Stalnaker semantics.

6 We assume here that there are only finitely many worlds. Note also that if α is -

Definition 1 allows us to understand a probability measure P as a probability distribution over worlds such that each w is assigned a probability $P(w) > 0$, and $\sum_w P(w) = 1$. We may determine the probability of a formula α by summing up the probabilities of the worlds at which the formula is true.[7]

$$(2) \qquad P(\alpha) = \sum_w P(w) \cdot w(\alpha)$$

Now, we are in a position to define Lewis's updating method of imaging.

Definition 2. Imaging (Lewis 1976, 310)

For each probability function P, and each possible formula α, there is a probability function P^α such that, for each world w', we have:

$$(3) \qquad P^\alpha(w') = \sum_w P(w) \cdot \begin{Bmatrix} 1 \text{ if } w_\alpha = w' \\ 0 \text{ otherwise} \end{Bmatrix}$$

We say that we obtain P^α by imaging P on α, and call P^α the image of P on α.

Intuitively, imaging transfers the probability of each world w to the most similar α-world w_α. Importantly, the probabilities are transferred, but in total no probability mass is additionally produced and no probability mass is lost. In formal terms, we have always $\sum_{w'} P^\alpha(w') = 1$. Any α-world w' keeps at least its original probability mass (since then $w_\alpha = w'$), and is possibly transferred additional probability shares of $\neg\alpha$-worlds w iff $\min_{\leq w} [\alpha] = w'$. In other words, each α-world w' receives as its updated probability mass its previous probability mass plus the previous probability shares that were assigned to $\neg\alpha$-worlds w such that $\min_{\leq w} [\alpha] = w'$. In this way, the method of imaging distributes the whole probability onto the α-worlds such that $P^\alpha(\alpha) = \sum_{w(\alpha)=1} P(w(\alpha)) = 1$, and each share remains 'as close as possible' at the world at which it has previously been located. For an illustration, see Figure 1.

---

[7] We assume here that each world is distinguishable from any other world, i. e. for two arbitrary worlds, there is always a formula such that the formula is true in one of the worlds, but false in the other. In other words, we consider no copies of worlds.

Lewis proved the following theorem, which relates the semantics of the Stalnaker conditional and the method of imaging on its antecedent.

Theorem 1. (Lewis 1976, 311)
The probability of a Stalnaker conditional equals the probability of the consequent after imaging on the antecedent, i. e. $P(\alpha > \gamma) = P^{\alpha}(\gamma)$, if α is possible.

Note that α in Theorem 1 may itself be of conditional form $\beta > \delta$ for any formulas β, δ.

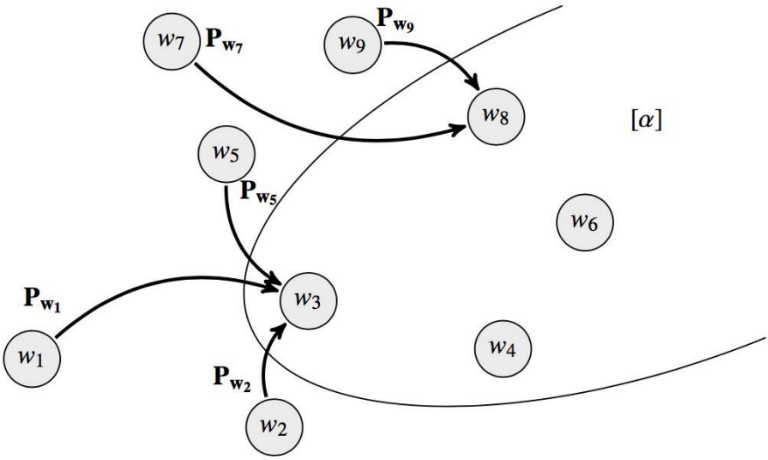

Figure 1: A set of possible worlds. The area delineated by the elliptical line represents the proposition or set of worlds $[\alpha] = \{w_3, w_4, w_6, w_8\}$. The thick arrows represent the transfer of probability shares from the respective $[\neg\alpha]$-worlds to their most similar $[\alpha]$-world. Similarity is graphically represented by topological distance between the worlds such that w_3, for instance, is the most similar or 'closest' $[\alpha]$-world to w_2.

3.3. Jeffrey imagining

The case of learning uncertain conditional information, i. e. $P(\alpha > \gamma) = k$ for $k \in [0, 1]$ but unequal to 0 or 1, requires to generalise Lewis's imaging method of Definition 2. In analogy to Jeffrey conditionalisation, we call the generalised method 'Jeffrey' imaging. Jeffrey imaging is based on Lewis's imaging and the fact that in a Stalnaker model the principle of Conditional Excluded Middle prescribes that $\neg(\alpha > \gamma)$ is equivalent to $\alpha > \neg\gamma$. We know, for all $w \in W$, presupposed $\alpha > \gamma$ is possible, both (I) that $\sum_w P^{\alpha > \gamma}(w)$ sums up to 1 and (II) that $\sum_w P^{\alpha > \neg\gamma}(w)$ sums up to 1. The idea is that if we form a weighted sum over the terms of (I) and (II) with some parameter $k \in [0, 1]$, then we obtain again a sum of terms $P_k^{\alpha > \gamma}(w)$ such that $\sum_w P_k^{\alpha > \gamma}(w) = 1$. Note, however, that we present the more general case $P_k^{\alpha}(w)$ in the definition below.

Definition 3. Jeffrey Imaging

For each probability function P, each possible formula α (possibly of conditional form $\beta > \delta$), and some parameter $k \in [0, 1]$, there is a probability function P_k^{α} such that for each world w' and the two similarity orderings centred on w_α and $w_{\neg\alpha}$, we have:

$$(4) \qquad P_k^{\alpha}(w') = \sum_w \left(P(w) \cdot \begin{cases} k & \text{if } w_\alpha = w' \\ 0 & \text{otherwise} \end{cases} + P(w) \cdot \begin{cases} 1-k & \text{if } w_{\neg\alpha} = w' \\ 0 & \text{otherwise} \end{cases} \right)$$

We say that we obtain P_k^{α} by Jeffrey imaging P on α, and call P_k^{α} the Jeffrey image of P on α. Note that in the case where $k = 1$, Jeffrey imaging reduces to Lewis's imaging.

Theorem 2. Properties of Jeffrey Imaging

(i) $\sum_{w'} P_k^{\alpha}(w') = 1$

(ii) $P_k^{\alpha}(\alpha) = k$

(iii) $P_k^{\alpha}(\neg\alpha) = (1 - k)$

(iv) $P_k^{\alpha}(\gamma) = k \cdot P(\alpha > \gamma)$ [8]

[8] The proofs of the properties can be found in Günther (2017).

We see that in total the revision method of Jeffrey imaging does neither produce additional probability shares, nor destroy any probability shares. In contrast to Lewis's imaging, Jeffrey imaging does not distribute the whole probabilistic mass onto the α-worlds, but only a part thereof that is determined by the parameter k.

In particular, as compared to Lewis's imaging, Jeffrey imaging may be understood as implementing a more moderate or balanced movement of probabilistic mass between α- and $\neg\alpha$-worlds. For an illustration, see Figure 2.

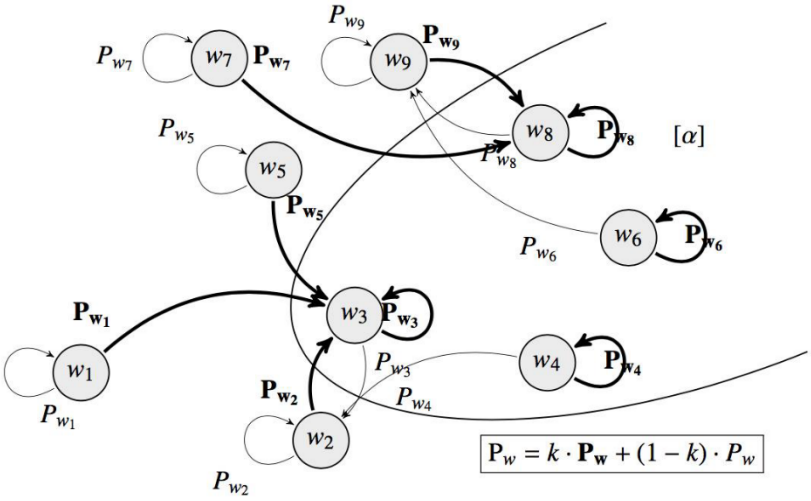

Figure 2: An illustration of the probability kinematics of Jeffrey imaging. The Jeffrey image P_k^α is characterised by a 'k-inertia' of the probabilistic mass from the respective α-worlds, and a '$(1-k)$-inertia' of the probabilistic mass from the respective $\neg\alpha$-worlds. Each thick arrow represents the transfer of the probability share $k \cdot P(w)$ to the closest α-world from w. Each thin arrow represents the transfer of the probability share $(1-k) \cdot P(w)$ to the closest $\neg\alpha$-world from w.

It is easy to show that P_k^{α} is a probability function. In a possible worlds framework, such a proof basically amounts to showing that the probability shares of all the worlds sum up to 1 after Jeffrey imaging. Therefore, property (i) of Theorem 2 provides minimal justification for applying Jeffrey imaging to probabilistic belief updating.

3.4. Putting the constituents together

Now we outline the method of learning conditional information put forward in Günther (2017). The method comprises three main steps:

(i) We model an agent's belief state as a Stalnaker model such that all and only those logical possibilities are represented as single worlds, which are relevant to the scenario under consideration. For instance, if only a single conditional "If α, then γ" is relevant and nothing else, then W contains exactly four elements as depicted in Figure 3.[9]

(ii) An agent learns conditional information "If α, then γ" iff (a) the agent interprets the received conditional information as a Stalnaker conditional $\alpha > \gamma$; (b) changes the similarity order $\leq$ by the meaning of $\alpha > \gamma$ in a minimally informative way and respecting the default assumption; and (c) updates her degrees of belief by Jeffrey imaging on the minimally informative meaning of $\alpha > \gamma$.

(iii) Finally, we check whether or not the result of Jeffrey imaging obtained in step (ii).(c) corresponds to the intuition associated with the respective example.

Step (ii) constitutes the core of the learning method and requires further clarification:

(a) In the agent's belief state, i.e. a Stalnaker model, the received information is interpreted. In the case of conditional information, the received information is interpreted as Stalnaker conditional. Hence, if the agent receives the information "If α, then γ", she interprets the information as meaning that the most similar

[9] In other words, we consider "small" models of possible worlds and do not allow for copies of worlds, i. e. worlds that satisfy the same formulas.

α-world (from the respective actual world) is a world that satisfies γ (presupposed α is possible). Technically, the interpretation (i.e. the meaning) of $\alpha > \gamma$ (relative to the Stalnaker model) is the proposition $[\alpha > \gamma] = \{w \in W \mid \min_{\leq w} [\alpha] \in [\gamma]\}$, where w is the respective actual world.

(b) The similarity order(s) is/are changed upon receiving conditional information. The proposition $\{w \in W \mid \min_{\leq w} [\alpha] \in [\gamma]\}$ depends on the similarity order $\leq$. The learning method prescribes that $\leq$ is specified, or adjusted, such that from each world the most similar α-world is a γ-world whenever possible. In other words, the method demands a maximally conservative, or equivalently minimally informative, interpretation of the received information. This amounts to specifying or adjusting the orders $\leq_w$ such that as many worlds as possible satisfy the received information. On the one hand, we can describe this interpretation as maximally conservative in the sense that no worlds are gratuitously excluded. On the other hand, we may think of possible worlds as information states. Then the exclusion of possible worlds corresponds to a gain of information. If an agent interprets the received information in a maximally conservative way, then as few as possible worlds or information states are excluded. In this sense, her gain of information is minimal.

The learning method assumes that the agent changes her similarity order respecting a default assumption. This default assumption states that the most similar $\alpha > \gamma$-world from any excluded $\alpha > \neg\gamma$-world is an $\alpha \wedge \gamma$-world, if there is more than one candidate. Formally, the default assumption expresses that $\min_{\leq w(\alpha > \neg\gamma)=1} [\alpha > \gamma] \models \alpha \wedge \gamma$, if $\min_{\leq w(\alpha > \neg\gamma)=1} [\alpha > \gamma]$ is underdetermined.[10] A justification for the default assumption is provided in Günther (2017).

[10] Relying on the default assumption solves a well-known problem of underdetermination: it might well be that, for instance, in the Stalnaker model depicted in Figure 3 w_3 or w_4 is the more similar $\alpha > \gamma$-world to w_2 than w_1 is. However, we will see in the examples below that additional (contextual) information may sometimes fully determine the epistemic states under consideration such that we do not always need to rely on the default assumption.

(c) Jeffrey imaging is applied on the minimally informative meaning of the Stalnaker conditional $\alpha > \gamma$. The application of Jeffrey imaging determines a probability distribution after learning the (uncertain) conditional information.

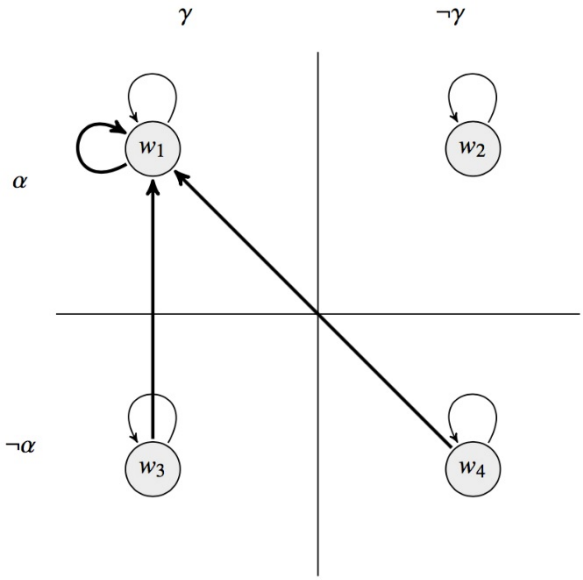

Figure 3: A four-worlds Stalnaker model for a case, in which the only received and relevant information is "If α, then γ". The reflexive thin arrows illustrate that each world w is the most similar to itself under the respective similarity order $\leq_w$. The thick arrows illustrate the change of the similarity order such that the received and interpreted information $[\alpha > \gamma]$ is minimally informative. Here, the minimally informative meaning of $\alpha > \gamma$ is $[\alpha > \gamma] = \{w \in W \mid w \models \alpha > \gamma\} = \{w_1, w_3, w_4\}$. Note that world w_2 is its own most similar α-world, but does not satisfy γ, i.e. $\min_{\leq w_2}[\alpha] \not\models \gamma$ and thus $\min_{\leq w_2}[\alpha > \gamma] \neq w_2$. Relying on the default assumption of step (ii).(b), $\min_{\leq w_2}[\alpha > \gamma] = w(\alpha \wedge \gamma) = w_1$. In words, the method prescribes that w_1 is the most similar $\alpha > \gamma$-world to w_2. This illustrates that the minimally informative meaning of $[\alpha > \gamma]$ implies that $\neg\gamma$ is excluded under the supposition of α. Hence, imaging on the minimally informative meaning of $\alpha > \gamma$ 'probabilistically excludes' w_2 and the probability share of w_2 will be fully transferred to w_1.

The proposed learning method has the following property that allows us to distinguish conditional and conjunctive information. If there is no further contextual information available to the agent receiving information, then learning the conditional information $\alpha > \gamma$ is less informative than learning the information $\alpha \wedge \gamma$. For, the proposition $[\alpha \wedge \gamma]$ is in the proposed framework always a strict subset of the minimally informative proposition $[\alpha > \gamma]$.

4. An adaptation of the method to the learning of causal information

In Section 2, we have seen that Douven invokes explanatory considerations in order to model the learning of conditional information. His account presupposes an explanatory reading of the learned conditional information, which may be of the form "If α, then γ". While we are skeptical about the presupposition that any conditional can or should be read as (a part of) an explanation or causal dependence, we admit that conditionals often figure in explanations. Hence, the method of learning conditional information proposed in Günther (2017) should be able to account for the learning of causal information conveyed by conditionals; otherwise, the proposed method suffers a major drawback.

In this section, we sketch how the proposed method may be adapted to a method of learning causal information. The adaptation is inspired by Lewis's analysis of causal dependence in terms of counterfactuals. Douven claims that, in any account of explanation that relies on a Stalnaker model, "to explain" means to "provide causal information", where "causal" refers to a Lewis-style analysis.[11]

[11] Cf. Douven (2012, 8-9, especially footnote 7); and Lewis (1973c). Furthermore, Douven claims that Lewis's and Stalnaker's semantics for conditionals are "exactly the same" (Douven 2012, 8). However, there is a difference between Stalnaker's and Lewis's semantics. In a Stalnaker model, there is always a single most similar world (or no world) to the actual world, whereas Lewis's semantics allows for a set of worlds (or no world) whose elements are equally similar to the actual world. A consequence of the difference is that Lewis's 'official' semantics for conditionals, i.e. the system VC, does not validate the principle of Conditional Excluded Middle, whereas Stalnaker's logic C2 for conditionals does. In Lewis's nomenclature, system C2 is labelled by VCS. Cf.

We write $\alpha \Rightarrow \gamma$ for the causal reading of "If α, then γ". According to Lewis's idea of causal dependence, $\alpha \Rightarrow \gamma$ is satisfied iff $\alpha > \gamma$ and $\neg\alpha > \neg\gamma$. We may apply the proposed method of learning conditional information by taking the minimally informative meaning of $\alpha \Rightarrow \gamma$ into account (instead of the one of $\alpha > \gamma$), if we substitute the default assumption. We call the adaptation the 'method of learning causal information'.

The substitution of the default assumption to what we call 'causal difference assumption' runs as follows. Assume we have no further contextual knowledge. Then, the most similar $\alpha \Rightarrow \gamma$-world from any excluded $\alpha \Rightarrow \neg\gamma$-world is a $(\alpha \wedge \gamma)$-world, if the excluded $\alpha \Rightarrow \neg\gamma$-world satisfies α. Furthermore, the most similar $\alpha \Rightarrow \gamma$-world from any excluded $\alpha \Rightarrow \neg\gamma$-world is a $(\neg\alpha \wedge \neg\gamma)$-world, if the excluded $\alpha \Rightarrow \neg\gamma$-world satisfies $\neg\alpha$. In symbols,

$$(5) \qquad min_{\leq w_{\alpha \Rightarrow \neg\gamma}}[\alpha \Rightarrow \gamma] = \begin{Bmatrix} w_{\alpha \wedge \gamma} & if\ w_{\alpha \Rightarrow \neg\gamma}(\alpha) = 1 \\ w_{\neg\alpha \wedge \neg\gamma} & if\ w_{\alpha \Rightarrow \neg\gamma}(\alpha) = 0 \end{Bmatrix}$$

The causal difference assumption is justified, if we understand causal dependence as difference making à la Lewis (cf. Lewis 1973c). The antecedent α makes the difference as to whether γ or $\neg\gamma$. Hence, $\alpha \Rightarrow \gamma$ means that worlds in which α obtains are worlds in which γ obtains, and accordingly that worlds in which $\neg\alpha$ obtains are worlds in which $\neg\gamma$ obtains. It is built in the analysis of causal dependence, so to speak, that the difference making factors (α and $\neg\alpha$) are more dissimilar than the ensuing effects.

Note that causal dependence is more informative than conditional dependence. For, the minimally informative meaning of $[\alpha \Rightarrow \gamma]$ is always a strict subset of the minimally informative meaning of $[\alpha > \gamma]$. The reason is that causal dependence, by definition, conveys in addition to the indicative conditional information also the information $[\neg\alpha > \neg\gamma]$. In brief, if an agent learns $\alpha \Rightarrow \gamma$, our adapted method prescribes that the $\alpha \wedge \neg\gamma$-worlds

---

Lewis (1973b; 1973a); and, for details, Unterhuber (2013, especially chap. 3.2, 3.3.3 and 3.3.4). The non-identity of Lewis's and Stalnaker's semantics implies that the notion of causal dependence employed in our method of learning causal information is not equivalent to Lewis's notion of causal dependence. While the method relies on Lewis's idea, we stick to Stalnaker's semantics in this paper.

transfer their probability shares to the most similar $\alpha \wedge \gamma$-world, and the $\neg\alpha \wedge \gamma$-worlds transfer their probability shares to the most similar $\neg\alpha \wedge \neg\gamma$-world. In other words, if the antecedent α is a difference maker, then the probability mass of those worlds w that do not satisfy $\alpha \Rightarrow \gamma$ is shifted to the most similar $\alpha \Rightarrow \gamma$-world w' that agrees with the Boolean evaluation for α, i. e. $w(\alpha) = w'(\alpha)$.

4.1. Douven's examples and the Judy Benjamin Problem

We apply now our adapted method of learning causal information to Douven's examples and the Judy Benjamin Problem.

4.1.1. A possible worlds model for the Sundowners Example

Example 1. The Sundowners Example (Douven & Romeijn 2011, 645-646)

Sarah and her sister Marian have arranged to go for sundowners at the Westcliff hotel tomorrow. Sarah feels there is some chance that it will rain, but thinks they can always enjoy the view from inside. To make sure, Marian consults the staff at the Westcliff hotel and finds out that in the event of rain, the inside area will be occupied by a wedding party. So she tells Sarah:

(6) If it rains tomorrow, we cannot have sundowners at the Westcliff.

Upon learning this conditional, Sarah sets her probability for sundowners and rain to 0, but she does not adapt her probability for rain.

We model Sarah's belief state as the Stalnaker model depicted in Figure 4. W contains four elements covering the possible events of R, $\neg R$, S, $\neg S$, where R stands for "it rains tomorrow" and S for "Sarah and Marian can have sundowners at the Westcliff tomorrow".

Let us assume that Sarah interprets the conditional uttered by her sister Marian as conveying the causal information $R \Rightarrow \neg S$. As Douven himself points out, the intuition in the Sundowners Example derives from the verdict that whether or not it rains makes the difference as to whether or not they have sundowners, but not the other way around: having sundowners simply has no effect whatsoever on whether or not it rains (cf. Douven

2012, 8). Hence, the change of belief between R and $\neg R$ is more far-fetched than between S and $\neg S$. In other words, the worlds along the horizontal axis are more similar than the worlds along the vertical axis. Since $R \Rightarrow \neg S \equiv (R > \neg S) \wedge (\neg R > S)$, $R \Rightarrow \neg S$ expresses both that S is excluded under the supposition of R and $\neg S$ is excluded under the supposition of $\neg R$. By the causal difference assumption, we obtain $\min_{\leq w1} [R > \neg S] = w_2$ and $\min_{\leq w4} [\neg R > S] = w_3$. Lewis's imaging method results in a shift of probability shares along the horizontal axis of Figure 4.

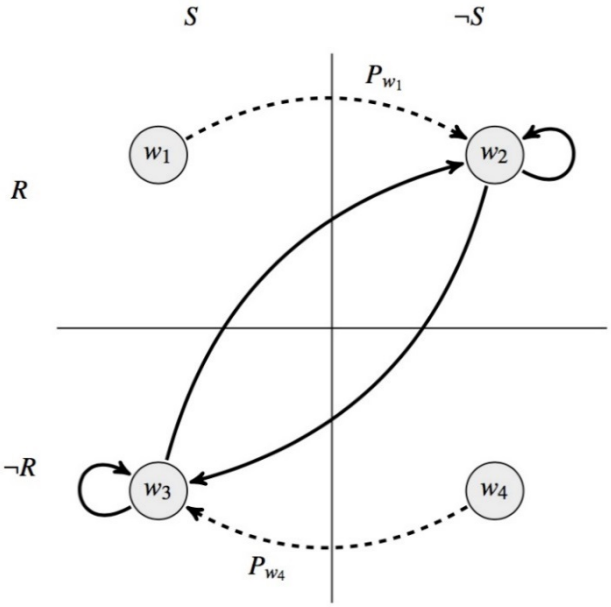

Figure 4: A Stalnaker model for Sarah's belief state in the Sundowners Example. The thick arrows illustrate the change of the similarity order such that the received information, causally understood as $R \Rightarrow \neg S$, is minimally informative. Here, the minimally informative meaning of $R \Rightarrow \neg S$ is $[R \Rightarrow \neg S] = [R > \neg S] \cap [\neg R > S] = \{w_2, w_3\}$. The dashed arrows represent the respective transfers of probability.

Imaging on the minimally informative proposition $[R \Rightarrow \neg S] = \{w_2, w_3\}$ results in $P^{R\Rightarrow\neg S}(w') = \sum_w P(w) \cdot \begin{cases} 1 & if\ w_{R\Rightarrow\neg S} = w' \\ 0 & \text{otherwise} \end{cases}$:

(7) $$P^{R\Rightarrow\neg S}(w_1) = 0$$
$$P^{R\Rightarrow\neg S}(w_2) = P(w_1) + P(w_2)$$
$$P^{R\Rightarrow\neg S}(w_3) = P(w_3) + P(w_4)$$
$$P^{R\Rightarrow\neg S}(w_4) = 0$$

We see immediately that both intuitions associated with the Sundowners Example are satisfied, viz. $P^{R\Rightarrow\neg S}(R) = P(R) = P(w_1) + P(w_2)$ and $P^{R\Rightarrow\neg S}(R \wedge S) = P(w_1) = 0$. We conclude that the method of learning causal information yields the intuitively correct results.[12]

4.1.2. A possible worlds model for the Ski Trip Example

Example 2. The Ski Trip Example (Douven & Dietz 2011, 33)

Harry sees his friend Sue buying a skiing outfit. This surprises him a bit, because he did not know of any plans of hers to go on a skiing trip. He knows that she recently had an important exam and thinks it unlikely that she passed. Then he meets Tom, his best friend and also a friend of Sue, who is just on his way to Sue to hear whether she passed the exam, and who tells him:

(8) If Sue passed the exam, her father will take her on a skiing vacation.

Recalling his earlier observation, Harry now comes to find it more likely that Sue passed the exam.

We model Harry's belief state as the Stalnaker model depicted in Figure 5. W contains eight elements covering the possible events of E, $\neg E$, S, $\neg S$,

[12] Note that the Sundowners Example seems to be somewhat artificial. It seems plausible that upon hearing her sister's conditional, Sarah would promptly ask 'why?' in order to obtain some more contextual information, before setting her probability for sundowners and rain to 0. After all, she 'thinks that they can always enjoy the view from inside'.

B, $\neg B$, where E stands for "Sue passed the exam", S for "Sue's father takes her on a skiing vacation", and B for "Sue buys a skiing outfit".

We assume that Harry interprets the conditional uttered by his friend Tom as conveying the causal information $E \Rightarrow S$. Furthermore, The Ski Trip Example assumes that Harry is equipped with the following contextual knowledge: Sue buying a skiing outfit may causally depend on the invitation of Sue's father to a skiing vacation, in symbols $S \Rightarrow B$. Finally, Harry observed Sue buying a skiing outfit, and thus has the factual information that B.

In total, Harry learns the minimally informative proposition $[(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B] = \{w_1\}$. Since w_1 is the only world that is not probabilistically excluded, we do not need to appeal to the causal difference assumption in this example.

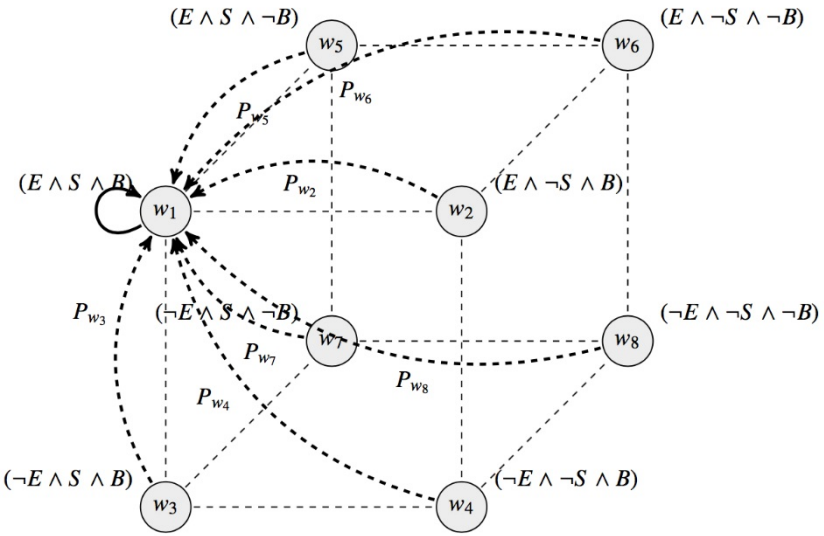

Figure 5: An eight-worlds Stalnaker model for Harry's belief state in the Ski Trip Example. Harry learns the minimally informative proposition $[(E \Rightarrow S) \wedge (S \Rightarrow B)] = \{w \in W \mid (\min_{w\leq}[E] \in [S]) \wedge (\min_{w\leq}[\neg E] \in [\neg S\,]) \wedge (\min_{w\leq}[S] \in [B]) \wedge (\min_{w\leq}[\neg S\,] \in [\neg B])\} = \{w_1, w_8\}$. Since Harry also obtains the factual information B, we can also exclude the $\neg B$-world w_8. (The arrows follow the convention of Figure 4.)

Imaging on the minimally informative proposition $[(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B] = \{w_1\}$ results in the following probability distribution, where we do not display the vanishing probabilities:

$$P^{(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B}(w') = P^*(w') = \sum_w P(w) \cdot \begin{cases} 1 & if\ w_{(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B} = w' \\ 0 & \text{otherwise} \end{cases}.$$

(9) $P^*(w_1) = 1$

The result meets the intuition associated with the Ski Trip Example: $P^*(E) > P(E)$, since $P^*(E) = P^*(w_1)$ and $P(E) = P(w_1) + P(w_2) + P(w_5) + P(w_6)$. Later on, we will see that the probabilities of the worlds w_2, w_3, w_4 would not have vanished entirely, if either $E \Rightarrow S$ or $S \Rightarrow B$ (or both) had conveyed only uncertain information.

In Günther (2017), we needed the default assumption to model the Ski Trip Example. If we appeal to the causal interpretation in the Ski Trip Example, we do neither need the default nor the causal difference assumption any more.

4.1.3. A possible worlds model for the Driving Test Example

Example 3. The Driving Test Example (Douven 2012, 3)

Betty knows that Kevin, the son of her neighbours, was to take his driving test yesterday. She has no idea whether or not Kevin is a good driver; she deems it about as likely as not that Kevin passed the test. Betty notices that her neighbours have started to spade their garden. Then her mother, who is friends with Kevin's parents, calls her and tells her the following:

(10) If Kevin passed the driving test, his parents will throw a garden party.

Betty figures that, given the spading that has just begun, it is doubtful (even if not wholly excluded) that a party can be held in the garden of Kevin's parents in the near future. As a result, Betty lowers her degree of belief for Kevin's having passed the driving test.

We model Betty's belief state as the Stalnaker model depicted in Figure 6. W contains eight elements covering the possible events of D, $\neg D$, G, $\neg G$,

S , $\neg S$, where D stands for “Kevin passed the driving test”, G for “Kevin’s parents will throw a garden party”, and S for “Kevin’s parents have started to spade their garden”.

Assume Betty interprets the conditional uttered by her mother as the causal information $D \Rightarrow G$. Furthermore, Betty infers from her contextual knowledge that because Kevin’s parents are spading their garden, they will not throw a garden party, in symbols $S \Rightarrow \neg G$. Finally, Betty knows that Kevin’s parents have started to spade their garden, and thus has the factual information that S.

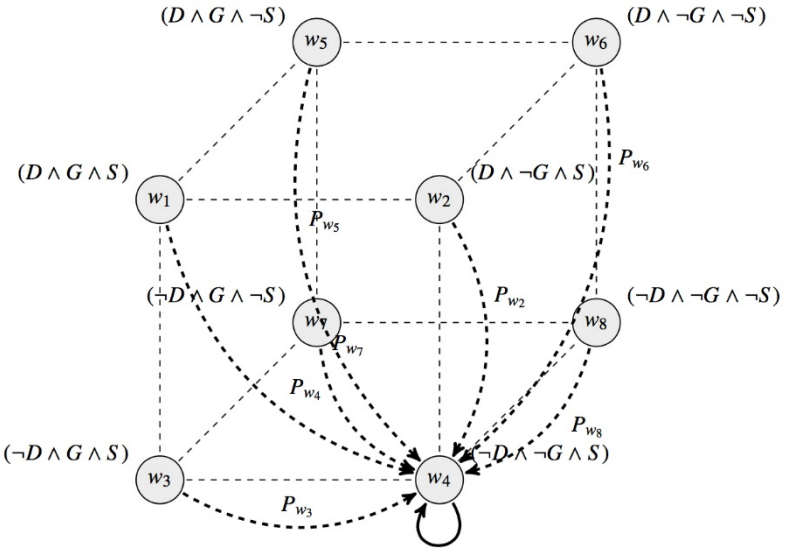

Figure 6: An eight-worlds Stalnaker model for Betty’s belief state in the Driving Test Example.

In total, Betty learns the minimally informative proposition $[(D \Rightarrow G) \wedge (S \Rightarrow \neg G) \wedge S] = \{w_4\}$. In Figure 6, we see that the Driving Test Example is structurally similar to the Ski Trip Example.

Imaging on the minimally informative proposition $[(D \Rightarrow G) \wedge (S \Rightarrow \neg G) \wedge S] = \{w_4\}$ results in the following probability distribution, where we do not display the vanishing probabilities:

$$P^{(D\Rightarrow G)\wedge(S\Rightarrow\neg G)\wedge S}(w') = P^*(w') = \sum_w P(w) \cdot \begin{cases} 1 & if\ w_{(D\Rightarrow G)\wedge(S\Rightarrow\neg G)\wedge S} = w' \\ 0 & \text{otherwise} \end{cases}:$$

(11) $P^*(w_4) = 1$

Our method yields again the correct result regarding the intuition associated with the Driving Test Example: $P^*(D) < P(D)$, since $P^*(D) = 0$ and $P(D) = P(w_1) + P(w_2) + P(w_5) + P(w_6) > 0$.

The following Judy Benjamin Problem will illustrate that if Betty thinks that the conditionals $D \Rightarrow G$ or $S \Rightarrow \neg G$ (or both) convey uncertain information, then the probability shares for some other worlds will not reduce to zero. This fact fits nicely with the Driving Test Examples's remark that "given the spading that has just begun, it is doubtful [or uncertain] (even if not wholly excluded) that a party can be held in the garden of Kevin's parents". We will treat the application of our method to the learning of uncertain causal information in the next section.

4.1.4. A possible worlds model for the Judy Benjamin Problem

We apply now our method of learning causal information to a case, in which the received causal information is uncertain. We show thereby that the method may be generalised to those cases in which the learned causal information is uncertain, provided we use Jeffrey imaging. Following the presentation in Hartmann & Rad (2017), we consider Bas van Fraassen's Judy Benjamin Problem (cf. van Fraassen 1981, 376-379).

Example 4. The Judy Benjamin Problem (Hartmann & Rad 2017, 7))

A soldier, Judy Benjamin, is dropped with her platoon in a territory that is divided in two halves, Red territory and Blue territory, respectively, with each territory in turn being divided in equal parts, Second Company area and Headquarters Company area, thus forming four quadrants of roughly equal size. Because the platoon was dropped more or less at the center of the whole territory, Judy Benjamin deems it equally likely

that they are in one quadrant as that they are in any of the others. They then receive the following radio message:

(12) I can't be sure where you are. If you are in Red Territory, then the odds are 3 : 1 that you are in Second Company area.

After this, the radio contact breaks down. Supposing that Judy accepts this message, how should she adjust her degrees of belief?

Douven claims that the probability of being in red territory should, intuitively, remain unchanged after learning the uncertain information. Furthermore, the probability distribution after hearing the radio message, i. e. P*, should take the following values:

(13) $P^*(R \wedge S) = \frac{3}{8}$ $\quad P^*(R \wedge \neg S) = \frac{1}{8}$

$P^*(\neg R \wedge S) = \frac{1}{4}$ $\quad P^*(\neg R \wedge \neg S) = \frac{1}{4}$

We model Judy Benjamin's belief state as the Stalnaker model depicted in Figure 7. W contains four elements covering the possible events of R, $\neg R$, S, $\neg S$, where R stands for "Judy Benjamin's platoon is in Red territory", and S for "Judy Benjamin's platoon is in Second Company area". The story prescribes that the probability distribution before learning the uncertain information is given by:

(14) $P(R \wedge S) = P(R \wedge \neg S) = P(\neg R \wedge S) = P(\neg R \wedge \neg S) = \frac{1}{4}$

In the previous examples, our agents implicitly learned Stalnaker conditionals of the form $\alpha > \gamma$ with certainty. According to Theorem 1, this amounts to the constraint that $P(\alpha > \gamma) = P^{\alpha}(\gamma) = 1$ (provided α is not a contradiction). Given this constraint and since P^{α} is a probability distribution, we have $P^{\alpha}(\neg\gamma) = 1 - P^{\alpha}(\gamma) = 0$. This means that we were able to probabilistically exclude any $\neg\gamma$-world under the supposition of α.

Now, our agent Judy Benjamin learns uncertain causal information, i.e. she implicitly learns Stalnaker conditionals with uncertainty. According to Theorem 1 and since $R \Rightarrow S$ is equivalent to $(R > S) \wedge (\neg R > \neg S)$, this amounts in the Judy Benjamin Problem to the constraint that $P(R \Rightarrow S) = \frac{3}{4}$.

By our method, we obtain $P(R \Rightarrow \neg S) = \frac{1}{4}$. In contrast to learning causal information with certainty, we cannot subtract the whole probabilistic mass from the $\neg S$ -worlds under the supposition of R, and accordingly from the S-worlds under the supposition of $\neg R$. However, Judy Benjamin is informed from an external source about the proportion to which she should gradually 'exclude' or downweigh the probability share of $R \Rightarrow \neg S$-worlds. Equivalently, we may say that the most similar $R \Rightarrow S$-world (from any $R \Rightarrow \neg S$)-world) obtains a gradual upweight of probability such that it receives $\frac{3}{4}$ of the probability shares of the $R \Rightarrow \neg S$-worlds; in turn, however, this $R \Rightarrow \neg S$-world then receives a probability share from the $R \Rightarrow S$-world weighed by $\frac{1}{4}$. Note that in Stalnaker models $R \Rightarrow \neg S$ is equivalent to $R > \neg S \wedge \neg R > S$.

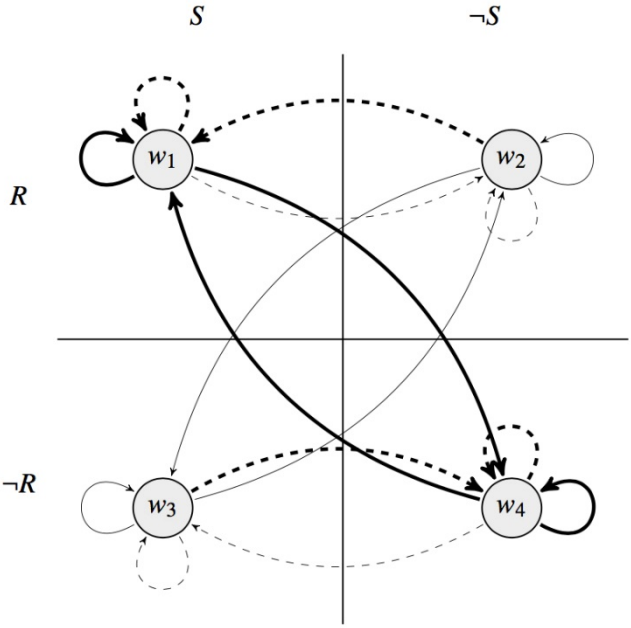

Figure 7: A Stalnaker model for Private Benjamin's belief state in the Judy Benjamin Problem. The thick arrows illustrate the specification of a similarity order $\leq'$ such that the received information $[R \Rightarrow S]$ is minimally informative. Note that each world having two outgoing thick arrows (one

for $R > S$ and one for $\neg R > \neg S$) satisfies $R \Rightarrow S$. The thin arrows illustrate the specification of another similarity order $\leq \neq \leq'$ such that the received information $[R \Rightarrow \neg S]$ is minimally informative. Each world having two outgoing thin arrows (one for $R > \neg S$ and one for $\neg R > S$) satisfies $R \Rightarrow \neg S$). In sum, the similarity orders are specified such that one makes $[R \Rightarrow S] = \{w_1, w_4\}$ a minimally informative proposition and the other makes the complement proposition $[R \Rightarrow \neg S] = \{w_2, w_3\}$ a minimally informative proposition. By the causal difference assumption, we obtain $min_{\leq'_{w2}}[R \Rightarrow S] = w_1$ and $min_{\leq'_{w3}}[R \Rightarrow S] = w_4$. Furthermore, we obtain $min_{\leq_{w1}}[R \Rightarrow \neg S] = w_2$ and $min_{\leq_{w4}}[R \Rightarrow \neg S] = w_3$. The thick dashed arrows represent the transfer of $k \cdot P(w)$, while the thin dashed arrows represent the transfer of $(1 - k) \cdot P(w)$. The application of Jeffrey imaging on $[R \Rightarrow S]$ with $k = \frac{3}{4}$ leads to the following calculation for the probability distribution: $P_{3/4}^{R \Rightarrow S}(w_1) = 3/4 \cdot P(w_1) + 3/4 \cdot P(w_2)$, and $P_{3/4}^{R \Rightarrow S}(w_2) = 1/4 \cdot P(w_1) + 1/4 \cdot P(w_2)$, and $P_{3/4}^{R \Rightarrow S}(w_3) = 1/4 \cdot P(w_3) + 1/4 \cdot P(w_4)$, and $P_{3/4}^{R \Rightarrow S}(w_4) = 3/4 \cdot P(w_3) + 3/4 \cdot P(w_4)$.

We apply now Jeffrey imaging to the Judy Benjamin Problem, where a source external to Judy provides her with the information that $k = \frac{3}{4}$.

$$(15) \quad P_k^{R \Rightarrow S}(w') = \sum_w \left(P(w) \cdot \begin{Bmatrix} k \text{ if } w_{R \Rightarrow S} = w' \\ 0 \quad \text{otherwise} \end{Bmatrix} + P(w) \cdot \begin{Bmatrix} 1 - k \text{ if } w_{R \Rightarrow \neg S} = w' \\ 0 \quad \text{otherwise} \end{Bmatrix} \right)$$

Given the probability distribution before the learning process in Equation (14), Judy obtains the following probability distribution after being informed that $P(R \Rightarrow S) = \frac{3}{4}$:

$$(16) \quad P_{3/4}^{R \Rightarrow S}(w_1) = P_{3/4}^{R \Rightarrow S}(R \wedge S) = \frac{3}{8}$$

$$P_{3/4}^{R \Rightarrow S}(w_2) = P_{3/4}^{R \Rightarrow S}(R \wedge \neg S) = \frac{1}{8}$$

$$P_{3/4}^{R \Rightarrow S}(w_3) = P_{3/4}^{R \Rightarrow S}(\neg R \wedge S) = \frac{1}{8}$$

$$P_{3/4}^{R \Rightarrow S}(w_4) = P_{3/4}^{R \Rightarrow S}(\neg R \wedge \neg S) = \frac{3}{8}$$

The probability distribution of (16) does not conform to Douven's intuitively correct distribution of (13), while the desideratum $P_{3/4}^{R>S}(R) = P(R) = \frac{1}{2}$ is met. Note that the learning of causal information results in $P_{3/4}^{R\Rightarrow S}(\neg R \wedge \neg S) = \frac{3}{8}$, which may be plausible for cases of causal dependence. However, we do not think that the conditional of the Judy Benjamin Problem is meant to express a causal dependence relation. In Günther (2017), we treated the received uncertain conditional as merely carrying uncertain conditional information. Applying the method of learning uncertain conditional information allowed us to offer a solution to the Judy Benjamin Problem that agrees with Douven's desired distribution of (13).

The Judy Benjamin Problem illustrates quite vividly the main difference between learning conditional and causal information. A merely conditional understanding of the conditional in the Judy Benjamin Problem does not affect the (row of) $\neg\alpha$-worlds, whereas the difference-making or causal dependence interpretation of the conditional affects the (row of) $\neg\alpha$-worlds.

5. Stalnaker inferences to the explanatory status of the antecedent

The method of learning causal information provides a formally precise implementation for when and how Douven's explanatory status of the antecedent should change. Recall his idea from Section 2 that the explanatory power of the antecedent with respect to the consequent determines the probability of the antecedent after learning the conditional. The idea is related to abduction, nowadays more commonly referred to as 'inference to the best explanation', or at least to a good explanation. The schema of such an inference runs as follows: α explains γ (well), and γ obtains. Therefore, α is true, or at least more likely.

We may interpret a Stalnaker agent's learning of $\alpha \Rightarrow \gamma$ as inference to a good explanation. Suppose an agent believes the fact γ and receives the information $\alpha \Rightarrow \gamma$. Then the agent infers that α explains γ (well). For, $\alpha \Rightarrow \gamma$ implies that $\neg\gamma$ would be the case, if α were not the case. But γ is the case and thus indicates that α is the case as well. The Ski Trip Example is an

instance of this type of reasoning. Harry learns $E \Rightarrow S$, $S \Rightarrow B$ and the fact B. He infers by our method of learning causal information that S explains B and, in turn, that E explains S. Consequently, $P^{(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B}(E) \geq P(E)$. In general, $P_k^{(\alpha \Rightarrow \gamma) \wedge \gamma}(\alpha) \geq P(\alpha)$, if $k > \frac{1}{2}$. In such a case, we call α the antecedent in a 'Stalnaker inference to a good explanatory status of the antecedent', or simply the antecedent in a 'Stalnaker inference to a good explanans'.

In the Driving Test Example, Kevin's passing the driving test (D) is at odds with the parent's spading their garden (S). D does not explain S (well). There is rather a tension between the occurrence of D and S. We can again formally implement the reasoning. Suppose S and $S \Rightarrow \neg G$, where G stands for "Kevin's parents will throw a garden party". Betty receives the information that $D \Rightarrow G$. S and $S \Rightarrow \neg G$ implies that G is not the case. By $D \Rightarrow G$, we may therefrom infer that D is not the case either. For, if D were the case, G would be the case. Consequently, $P^{(D \Rightarrow G) \wedge (S \Rightarrow \neg G) \wedge S}(D) \leq P(D)$. In general, $P_k^{(\alpha \Rightarrow \gamma) \wedge \neg \gamma}(\alpha) \leq P(\alpha)$, if $k > \frac{1}{2}$. In such a case, we call α the antecedent in a 'Stalnaker inference to a bad explanans'. Notice that our framework allows for a probabilification of the Stalnaker inferences, if uncertain causal information is learned.

6. Conclusion

We have seen that Douven's dismissal of the Stalnaker conditional as a tool to model the learning of conditional and causal information is unjustified. Rather, this type of learning may be modelled by Jeffrey imaging on the meaning of Stalnaker conditionals under the following condition: the similarity order of the Stalnaker model is changed in a way such that the meaning of the conditional is minimally informative. Both methods of learning information align with the intuitively correct results in Douven's benchmark examples. However, Douven's intuitions about the Judy Benjamin Problem are only met, if we understand the conditional Judy receives as conveying merely conditional information.

We have shown that the method of learning (uncertain) conditional information proposed in Günther (2017) may be adapted to a learning method of (uncertain) causal information. The adaptation is based on the Stalnaker

conditional, for which Lewis's idea of causal dependence is implemented. The two methods come with two different assumptions, i.e. the default assumption and the causal difference assumption, respectively. The combination of the two methods provides a unified framework that manages to clearly discern between a merely conditional and a causal reading of the conditional "If α, then γ". Hence, the general method cannot be attacked for not being applicable to conditionals that (are supposed to) express causal dependences. In detail, if no further contextual information is available, conjunctive information is strictly more informative than causal information, which is in turn strictly more informative than conditional information. For, the minimally informative conjunctive, causal and conditional propositions stand in the following strict subset relation: $[\alpha \wedge \gamma] \subset [\alpha \Rightarrow \gamma] \subset [\alpha > \gamma]$.

The causal dependence reading can be used to formalise Douven's explanatory status of the antecedent. We thereby convey the explanatory status a precise formal meaning that may be used to operationalize Douven's idea that explanatory considerations play a core role in learning conditionals. Furthermore, the results suggest that we should distinguish between a merely conditional or suppositional interpretation and a causal dependence interpretation of a conditional. A supposition should not affect those cases, in which the antecedent is not satisfied, whereas a difference-making conditional should. Based on this distinction, we hope that the proposed framework can help psychologists of reasoning to provide an empirically adequate account of actual reasoning behaviour with respect to the learning of conditional and causal information.

The advantages of our unified framework of learning uncertain information, as compared to alternative accounts, will be assessed in a follow-up paper. We plan to compare our account in detail to Douven's account of learning conditional information and Bayesian accounts of learning conditionals. In particular, we will show that the Bayesian account of Hartmann & Rad (2017) – that minimizes the Kullback-Leibler divergence on a fixed Bayesian network – has severe problems to capture the merely conditional interpretation of conditionals. As a consequence the Judy Benjamin Problem remains troublesome for their account.

References

BRADLEY, R. (2005): Radical Probabilism and Bayesian Conditioning*. *Philosophy of Science* 72, No. 2, 342-364.

DOUVEN, I. (2012): Learning Conditional Information. *Mind & Language* 27, No. 3, 239-263.

DOUVEN, I. & DIETZ, R. (2011): A Puzzle about Stalnaker's Hypothesis. *Topoi* 30, No. 1, 31-37.

DOUVEN, I. & ROMEIJN, J.-W. (2011): A New Resolution of the Judy Benjamin Problem. *Mind* 120, No. 479, 637-670.

DOUVEN, I. & VERBRUGGE, S. (2010): The Adams Family. *Cognition* 117, No. 3, 302-318.

GÜNTHER, M. (2017): Learning Conditional Information by Jeffrey Imagining on Stalnaker Conditionals. *Journal of Philosophical Logic*, online first. DOI 10.1007/s10992-017-9452-z.

HARTMANN, S. & RAD, S. R. (2017): Learning Indicative Conditionals. *Unpublished manuscript*, available at: http://www.rafieerad.org/conditional.pdf.

LEWIS, D. (1973a): *Counterfactuals*. Blackwell Publishers.

LEWIS, D. (1973b): Counterfactuals and Comparative Possibility. *Journal of Philosophical Logic* 2, No. 4, 418-446.

LEWIS, D. (1973c): Causation. *Journal of Philosophy* 70, No. 17, 556-567.

LEWIS, D. (1976): Probabilities of Conditionals and Conditional Probabilities. *The Philosophical Review* 85, No. 3, 297-315.

PFEIFER, N. & DOUVEN, I. (2014): Formal Epistemology and the New Paradigm Psychology of Reasoning. *Review of Philosophy and Psychology* 5, No. 2, 199-221.

STALNAKER, R. C. (1975): A Theory of Conditionals. In: Sosa, E. (ed.): *Causation and Conditionals*. Oxford University Press, 165-179.

STALNAKER, R. C. & THOMASON, R. H. (1970): A Semantic Analysis of Conditional Logic. *Theoria* 36, No. 1, 23-42.

UNTERHUBER, M. (2013): *Possible Worlds Semantics for Indicative and Counterfactual Conditionals? A Formal Philosophical Inquiry into Chellas-Segerberg Semantics*. Ontos.

VAN FRAASSEN, B. C. (1981): A Problem for Relative Information Minimizers in Probability Kinematics. *The British Journal for the Philosophy of Science* 32, No. 4, 375-379.

ZHAO, J., CRUPI, V., TENTORI, K., FITELSON, B., & OSHERSON, D. (2012): Updating: Learning versus Supposing. *Cognition* 124, No. 3, 373-378.

Organon F 24 (4) 2017: 487-510

Hempel's Dilemma and Research Programmes: Why Adding Stances Is Not a Boon

DUŠKO PRELEVIĆ[1]

ABSTRACT: Hempel's Dilemma is intended to force physicalists to make an unfavourable choice between the current physics and a future physical theory. The problem with the first horn of the dilemma is related to the fact that current physics is, strictly speaking, inconsistent, while the problem with its second horn is that we do not know how a future, completed physical theory will look like. In this paper, the two strategies of avoiding the dilemma are compared and assessed: the attitudinal approach, according to which physicalism is a stance or an attitude, and Lakatosian approach, according to which physicalism is best understood as a research programme. It is argued that the latter approach ought to be preferred over the former approach because, among other things, it better explains how some physicalists and their opponents sometimes switch the sides, as well as why different physicalists undertake different activities within a given time interval.

KEYWORDS: Attitudinal approach – Hempel's Dilemma – paradigm – physicalism – research programme – stance.

[1] Received: 31 January 2017 / Accepted: 5 August 2017

✉ Duško Prelević
Department of Philosophy
Faculty of Philosophy, University of Belgrade
Čika Ljubina 18-20, 11000 Belgrade, Serbia
e-mail: dusko.prelevic@f.bg.ac.rs; dprelevic@yahoo.com

1. Physicalism and Hempel's Dilemma

As one popular survey reports (see Chalmers & Bourget 2014, 476), most philosophers today are physicalists, that is, they think that everything is physical. They also believe that physics can explain the nature of the universe or at least that the fundamental level of reality is the subject-matter of physics. Physicalism can be spelled out in many ways, and, as pointed out by Robert Kirk, a less committing way is to say that the language of physics is (at least in principle) capable of describing all the facts about the universe, while the language of any other science is at best a re-description of the same reality.[2]

This means that even if one allows for non-physical facts and non-physical properties, these facts and properties, according to physicalists, do not belong to the fundamental level of reality: At best, non-physical properties might supervene on physical properties, which means that once all physical facts (plus the laws of physics) are fixed, everything else will be settled as well. If so, then even if some other scientific discipline (e.g., biology, psychology, economy, etc.) uses a vocabulary different from the one used in physics, all these disciplines would tell us something about one and the same reality: the physical reality.

It is likely that physicalism is a background for many/most scientists today. In physics, the debates over the correct interpretation of quantum mechanics and the validity of the causal closure of the physical, according to which all physical effects are fully determined by prior physical occurrences (and the laws of physics), might serve as an illustration. Although physics and physicalism are not the same,[3] many contemporary physicists are physicalists. For example, physicists typically endorse the principle of causal closure. However, there is more than one interpretation of quantum mechanics, one of which is the so-called Wigner's hypothesis, according to which consciousness might cause the wave function collapse. Such a hypothesis contravenes the principle of causal closure and goes in favor of

---

2 This is what Kirk calls "minimal physicalism"; see Kirk (2006) for more details.

3 That is because physicalism goes beyond physics by telling us that the fundamental level of reality can be fully described and explained by physics.

a dualistic ontology. Now, as Chalmers has pointed out,[4] physicists typically endorse the principle of causal closure (and, therefore, they reject Wigner's hypothesis) because they are physicalists, while, on the other hand, physicalists (for example, David Papineau when defending the causal argument for physicalism)[5] typically say that the principle of causal closure ought to be accepted because physicists typically endorse it.

In that respect, physicalism discourages work on certain theories (e.g. the work on the dualistic interpretation of quantum mechanics), while it encourages work on some other theories (e.g. the work on interpretations of quantum mechanics which aim to support physicalism).

Physicalism has also inspired establishing new disciplines. For example, Patricia Churchland's book *Neurophilosophy* (see Churchland 1986) had a great impact on establishing neuroethics (and perhaps later on neuroeconomics, neuroaesthetics, and so on), bringing together philosophers and scientists from various fields in order to discuss new problems. Also, many neuroscientists direct their research toward discovering neural mechanisms of yet unexplained mental processes just because they believe that all mental processes are physical. Otherwise, they would probably stop their research or they would redirect it, for example, toward dualistic or panpsychist sorts of explanation. So it is likely that physicalism is a background for many/most neuroscientists today too.

However, physicalism has been defended and characterized in more than one way. As is well known, physicalists respond to the main anti-physicalist arguments (such as the zombie argument, the knowledge argument, and so on) in different ways, and sometimes they even dispute among themselves over which responses are the most satisfactory ones. Physicalists also propose various positive arguments and accounts in order to justify their view.[6]

---

[4] See his talk (based on his collaboration with Kelvin McQueen) "Consciousness and the collapse of the wave function" (2014) which is available at: https://www.youtube.com/watch?v=DIBT6E2GtjA

[5] The causal argument runs as follows (cf. Papineau 2001, 9): All physical effects are fully determined by laws and prior physical occurrences; all mental occurrences have physical effects; the physical effects of mental causes are not fully overdetermined; therefore, mental occurrences must be identical with physical occurrences.

[6] Some of them will be sketched briefly in section 2.

Last but not least, the key notion physicalists use, the notion of "physical", underwent so many changes in the history of science. For example, the concept of matter has changed in light of new scientific discoveries (see, for example, Ney 2008a, 1034), and the same happened to some other basic notions in physics, such as space, time, mass, and the like. Further, science surprised us many times by positing new properties at the fundamental level, given that different physical theories considered different properties as fundamental. Bearing this in mind, one might be skeptical about the idea that there are necessary conditions for something to be a physical object. This opens the question on how far physicalists should go in accepting the changes of the key notions they use without ceasing to be physicalists.

This creates a tension between ontological and methodological commitments that arguably any physicalist should take. Namely, the ontological commitment binds physicalists to rule out a view that non-physical entities belong to the fundamental level of reality, while the methodological commitment binds them to accept everything physics says is true. Accepting the latter commitment threats to undermine the former, and *vice versa*. This, among other things, gives rise to a well-known dilemma for physicalists, posed by Carl Hempel (see Hempel 1980, 195), which is now called "Hempel's Dilemma".

The dilemma runs as follows: Physicalists, who claim that physics alone can explain the nature of the universe, should be more accurate and say exactly which physical theory they have in mind. At first glance, it seems that they have to choose between the current physics and a future physical theory,[7] which is rather an unpleasant choice: On the one hand, current physics is incomplete, and, strictly speaking, inconsistent, since the standard model of quantum mechanics, which is powerful in describing microphysical phenomena, is indeterministic, while general relativity, which accurately describes the universe on large scales, is deterministic (see, for example, Greene 2004, 333-335, for more details). Thus, taking the first horn of the dilemma (the so-called "currentism") is not attractive because it is irrational to believe in inconsistencies and take them as capable of providing a complete explanation of the universe. On the other hand, we

[7] Appealing to an already abandoned physical theory obviously would not be an option.

do not know how a future physical theory will look. This means that taking the second horn of the dilemma faces the "inappropriate extension worry" (see Wilson 2006, 68), which is based on the conceivability of a scenario in which a future physical theory posits irreducible non-physical entities (like phenomenal consciousness) at the fundamental level.[8] Such a scenario is likely the one in which physicalism is not true. Now, if physicalists deny in advance that such a scenario will happen, it would depart from ordinary scientific practice, to which physicalists appeal, since physics is, after all, an empirical science, and therefore it is possible, at least in principle, that it can surprise us (as it did many times in the history of science) by positing new properties at the fundamental level. On the other hand, if physicalists bite the bullet and claim that they will be ready to accept even the ideal physical theory that posits phenomenal consciousness at the fundamental level, then physicalism, according to the objection, turns out trivial and empty, because in that case anything goes (see, for example, Wilson 2006; Ney 2008a, 1037). The upshot of Hempel's Dilemma is that physicalism is either false or a trivial and contentless doctrine.

Hempel's Dilemma is a usual way to approach the problem of characterizing physicalism, and it serves as a fruitful guide that can help physicalists to spell out their view in a more precise way. In that respect, the dilemma primarily deals with the meaning of the physicalist claim, that is, it primarily concerns a *meta*-question about what physicalism is and, related to this, about what all physicalists have in common (see Prelević 2017, 5 for more details). Hempel's Dilemma deals with the question of truth of the physicalist claim too, since solutions that render physicalism false would not be considered plausible. Relatedly, competing solutions can be compared with regard to what extent they are realistic accounts of worth considering phenomena that will be addressed in due course.

Three strategies of dealing with Hempel's Dilemma have been proposed by now: defending currentism, defending futurism or trying to avoid the dilemma by claiming that physicalism is not a thesis that might be trivial or empty, but something else (e.g., a stance or a research programme). The

[8] Here, dualistic interpretation of quantum mechanics (Wigner's hypothesis), mentioned in this section, might serve as an illustration.

first two strategies have been widely defended and criticized.[9] In what follows, the focus will be on the third strategy.

2. Avoiding the dilemma

In the previous section, we have seen that Hempel's Dilemma, which is aimed to force physicalists to take an unfavourable choice between current physics and a future physical theory, presupposes that physicalism is a *thesis* that might be true, false, trivial or empty. Avoiding the dilemma consists in challenging such an assumption. Here, two ways of avoiding the dilemma will be addressed: the attitudinal approach and understanding physicalism as a research programme. In what follows, these two approaches above will be compared. It will be argued that the latter approach (presented in section 2.2) fares better than the former approach (presented in section 2.1) as to how some physicalists (and their opponents) sometimes switch the sides, as well as why different physicalists undertake different activities within a given time interval. These considerations, if correct, would license a view that the latter approach (properly understood) ought to be preferred over the former approach because it is a more realistic account of worth considering phenomena that are relevant for characterizing physicalism and resolving Hempel's Dilemma thereof.

2.1. The attitudinal approach

Let us start with the attitudinal approach, according to which physicalism is best understood as a stance (or an attitude). Alyssa Ney expresses such an attitude in the following slogan: "I hereby swear to go in my ontology everywhere and only where physics leads me" (Ney 2008, 11).

In philosophy of science, the notion of stance has been famously introduced by Bas van Fraassen (2002). He has done so because, among other things, he wanted to resolve the problem of justifying empiricism. Namely, if empiricism is the claim that experience is the one and only source of factual information, then there is a problem of how to justify the empiricist

[9] For arguments against currentism, see, e.g., Wilson (2006, 64-66); and Prelević (2017); for the disputes among futurists, see Wilson (2006); and Dowell (2006); for critiques of futurism, see, for example, Ney (2008a); and Prelević (2017).

claim itself, since such a claim cannot be supported by experience. Hence, adopting empiricism as a thesis would be self-defeating. For that reason, van Fraassen understands empiricism as a stance that commits its adherents to act in a certain way and, at the same time, being aware that adopting such a stance is not justified by providing an algorithm or something of that sort. By the same token, van Fraassen thinks that problems like Hempel's Dilemma can be avoided once we understand physicalism as a stance, and not as a thesis (see van Fraassen 2002, 49 for more details)

It has already been pointed out that van Fraassen's account does not match well with the standard classifications in the history of philosophy, since it allows us to count philosophers like Descartes, Leibniz and Chalmers – who arguably tried (or could have tried) to reconcile their ontologies with their preferred physical theories[10] – as physicalists, which is rather implausible (see Prelević 2017). Perhaps one way of dealing with this problem would be to include some metaphysical commitments in characterizing physicalism along the lines of James Ladyman's defence of what he calls the "scientistic stance" (see Ladyman 2011). Although this would depart from what van Fraassen originally had in mind – after all, van Fraassen's empirical stance was purported to be anti-metaphysical – it would still be in line with the view that physicalism is a stance rather than something else.

By having or taking a stance, van Fraassen means "having or adopting a cluster of attitudes, including a number of propositional attitudes which will generally include some beliefs" (van Fraassen 2004, 175; see also van Fraassen 2002, 47-48). Here, the main point is that stances are not theses (although they typically contain them) as well as that stances permit someone to endorse a belief *without* pretensions to claim that such a belief is

[10] For example, Descartes considered conservation laws (the "quantity of motion") a nondirectional (scalar) quantity (mass times speed; see, for example, Descartes 1985, 83–84; see also Woolhouse 1985; and Papineau 2001, 14-15), which made it possible for him to claim that mind can alter the direction of body's motion leaving the conservation laws intact. Leibniz famously criticised him on these matters (see, for example, Leibniz 1997), but given that he, like Descartes, endorsed the causal closure of physics, he proposed the doctrine of pre-established harmony instead of interactionist dualism. On the other hand, a dualistic interpretation of quantum mechanics, to which Chalmers sometimes appeal, contravenes the causal closure of the physical world (see section 1).

rationally mandated (cf. Teller 2004, 161). As Paul Teller suggests, the notion of stance can be clarified by using the analogy with adopting a policy:[11] truth values are not assigned to policies, policies commit us to act in a certain way or to make certain decisions, they may be overridden by some other criteria or policies, they may be interpreted or applied in more than one way, and so on.

Given that, as stressed above, stances permit someone to endorse a belief *without* pretensions to claim that such a belief is *rationally mandated*, van Fraassen's approach is confronted with the problem of "stance voluntarism", which refers to "the thesis that one can intentionally acquire or sustain a stance in the absence of any epistemic reasons for that stance" (Baumann 2011, 29). Such a thesis implies that contrary stances are rationally permissible (see Chakravartty 2011).[12]

In that respect, it is not surprising that van Fraassen's conception of stance is often compared with Kuhnian view of paradigms, since Kuhn (1962) famously argued that, during scientific revolutions, "paradigm shifts" occur in a way in which replaced and newly established paradigms are incommensurable. Paradigms are, simply put, frameworks within which scientific communities work. In his "Second Thoughts on Paradigms" (see Kuhn 1974), Kuhn understood paradigms as disciplinary matrices that consist in "a constellation of group commitments" which, among other things, include examplars (shared examples) that suggest new puzzles, approaches to resolving them, and serve as standards that enable those who do the research within the paradigm to measure the quality of the proposed solutions (cf. Rowbottom 2011, 115)

As Darrell Rowbottom has pointed out, stances are very similar yet not identical to paradigms. According to him, stances should not be understood as paradigms writ large, since paradigms, unlike stances, include examplars. Rowbottom thinks that introducing stances should not be understood merely as spelling out a known idea in a new fashion, but as appraising it as a boon. He thinks that the distinction between stances and paradigms

---

[11] Van Fraassen agrees with him on that by telling that it clarifies the epistemological aspects of the notion (see van Fraassen 2004, 179).

[12] I will stay neutral in due course on whether van Fraassen's view of stance voluntarism leads to latent irrationality or not (this objection can be found, for example, in Baumann 2011).

enables us to explain why different scientists undertake different activities, that is, "how and why there is a measure of dissent within the boundaries of the disciplinary matrix" (Rowbottom 2011, 115). Rowbottom's solution to this problem runs as follows: "My basic idea is that a disciplinary matrix implies a *set of permissible stances*, and that the difference in stances of individual scientists explains how and why a broad range of activities occur" (Rowbottom 2011, 117). At the end of his paper, Rowbottom conjectures that van Fraassen's notion of stance may be also used to explain Kuhnian conversions in science, yet he finishes his paper without developing such an idea.

2.1.1. Physicalism and conversions

In the previous section, it was stressed that both Kuhnian view of paradigm shifts and van Fraassenian view of stance voluntarism are aimed to support the thesis that conversions in science are not rationally mandated. However, in the context of the debate over the possibility of characterizing physicalism, these accounts are hardly acceptable.[13] After all, the fact that so many arguments have been proposed for or against physicalism (and alternative views as well) suggests that a rational choice between physicalism and the alternative views can be made within a given time interval, contrary to what Kuhn's incommensurability thesis and van Fraassen's stance voluntarism presuppose.

Here, it is worth mentioning that even if some physicalists appeal to the Kuhnian view of scientific revolutions, it would still not follow that they themselves experience paradigm shifts whenever they introduce their theories. For example, eliminativists like Daniel Dennett[14] and Paul and Patricia Churchland typically claim that phenomenal consciousness will be

---

[13] As is well known, Kuhn's incommensurability thesis as such has been criticized many times (see, for example, Newton-Smith 1981 for more details). However, the main point here is that even if such a thesis can help us get a better grasp of some interesting episodes in the history of science, it would still not be of any use for our understanding of the nature of physicalism. The same holds, *mutatis mutandis*, for van Fraassen's account.

[14] As for Dennett, many times he has challenged anti-physicalist arguments, such as the zombie argument and the knowledge argument, by arguing that they are bad thought

explained away within a future physical theory in almost the same way as it happened with some other theoretical terms in science, such as phlogiston, luminiferous aether, and the like (see, e.g., Churchland 1996). Given that Kuhn interpreted episodes like these as the cases of paradigm shifts, a natural guess is that at least some eliminativists think (or could have thought) that a corresponding paradigm shift will dissolve phenomenal consciousness too. Yet this would at best show that philosophers who appeal to Kuhnian insights on how revolutions in science occur do that in order to provide a *rational* support for their view rather than because of experiencing a paradigm shift. Here the structure of their arguments would be almost the same as of those used by some identity theorists or analytic functionalists who appeal to theoretical identifications established in natural sciences (such as that water is H_2O, that genes are DNA, and the like) in order to justify the claim that consciousness is a brain process, and the like. Such optimism is far from not being rationally mandated[15] at least from the perspective of philosophers who share it and in the absence of counter-arguments. So it is likely that physicalistic views like eliminativism are not incommensurable with anti-physicalistic views.

In addition, let us recall a few representative cases of conversion in philosophy of mind. One such example is Frank Jackson's conversion, whose version of the knowledge argument is widely discussed in contemporary philosophy of mind.[16] Here is what Jackson says on this issue in one interview:[17]

---

experiments (he calls them "intuition pumps"; see Dennett 1991, 282 for more details). This also reveals that his defence of physicalism is rationally mandated.

[15] Here, as well as in cases below, I just present briefly some well-known arguments of various physicalists and their opponents in order to shed a better light on the nature of their debates and enterprises. I do not commit myself to holding their arguments valid.

[16] Jackson's knowledge argument is intended to show that knowledge of completed physics (chemistry and neurophysiology) does not enable us to know everything about the world, since one who knows everything about a completed science of colour vision could still be, for instance, ignorant of what is it like to see red.

[17] See the interview: "Frank Jackson, Later Day Physicalist" (2011), which is available at: http://www.philosophersmag.com/index.php/tpm-mag-articles/14-interviews/22-frank-jackson-latter-day-physicalist.

> In 'Epiphenomenal Qualia' I explain why it's not such a disaster being an epiphenomenalist, but I came to think of this as a triumph of philosophical ingenuity over common sense. This is what someone who's done a good philosophy degree can somehow make seem all right, but if you look at it in a more commonsensical way it's actually pretty implausible. So the epiphenomenal stuff was just very hard to believe.

However, Jackson himself changed his mind definitely *after* realizing that a representationalist theory of consciousness (a version of intentionalism that goes in favor of physicalism) is a viable doctrine. Actually, he detected the key intuition behind the knowledge argument and tried to show how such an intuition conflicts with an attractive view of the nature of phenomenal concepts that can be defended on independent grounds (see Jackson 2007 for more details). He has also provided some reasons why, for example, he believes that alternative responses to the knowledge argument, such as the "missing-concept reply", are not convincing.[18] So it is likely that Jackson's conversion to physicalism was rationally mandated, contrary to what van Fraassen and Rowbottom would say in similar cases.

It is also interesting to notice that some main figures in the debate over the validity of the zombie argument have completely changed their views on these matters.[19] On the one hand, Robert Kirk, who introduced the zombie argument in 1970s (see Kirk 1974), has changed sides and started to argue that zombies are not just impossible, but inconceivable as well (see, for example, Kirk 2007), while on the other hand, David Chalmers, whose version of the zombie argument against physicalism has been in focus for

---

[18] This reply consists in claiming that inside her black-and-white room (in Jackson's thought experiment) Mary is unable to acquire phenomenal concepts, which does not entail by itself that phenomenal truths are not a priori deducible from corresponding totality of micro-physical truths (plus the laws of physics).

[19] The zombie argument, roughly, starts with the premise that zombies – our physical duplicates who, unlike us, do not have phenomenal consciousness – are conceivable, continues with the principle that conceivability entails metaphysical possibility, ending up with the conclusion that metaphysical possibility of zombies undermines physicalism, in one way or another.

last twenty years or so (see, for example, Chalmers 2010), originally had thought that zombies are impossible, albeit conceivable.

However, these conversions can hardly be regarded as the cases of Kuhnian paradigm shifts. Robert Kirk has tried to show that the zombie scenario implies a sort of epiphenomenalism that involves a contradiction (cf. Kirk 2007). As for Chalmers, here is what he says in a recently held interview about his conversion:[20]

> I wanted to write a big-picture treatment of consciousness in philosophy and science and at the same time put forward a positive theory of consciousness. In my first couple of years at Indiana I wrote two long articles (still unpublished except on the web) pursuing the connection between consciousness and the way we talk about consciousness, but I also gradually got drawn into issues about materialism and dualism. I had come to graduate school thinking of myself as a materialist (albeit one who was very impressed by the problem of consciousness), but I gradually realized that commitments I already had meant that materialism couldn't work, and I should be some sort of dualist or perhaps panpsychist.

The passage above suggests that Chalmers has changed his view after a more careful reflection on the commitments he already had accepted, and realizing that those commitments are incompatible with physicalism (materialism). A natural guess is that he realized that his views on the relation between modality and apriority, semantics of phenomenal and micro-physical concepts, quantum mechanics, and the like, do not match well with physicalism.

These representative cases of conversion suggest that it is more likely that they are rationally mandated. They neither justify Kuhnian view of paradigm shifts, nor van Fraassenian stance voluntarism, which is considered to be a hallmark of the attitudinal approach.

---

[20] See the interview: "What Is It Like to Be a Philosopher?" (2016), which is available at: http://www.whatisitliketobeaphilosopher.com/#/david-chalmers/.

2.1.2. Varieties of physicalism

Now, let us check whether van Fraassen's attitudinal approach can explain why different physicalists undertake different activities. In order to show that this is not the case, let us start with noticing that the history of physicalism is to a great extent parallel with the history of analytic philosophy, primarily with respect to the question about how philosophers see the relationship between philosophy, science and metaphysics. Namely, when Otto Neurath coined the term "physicalism" in 1930s (see Neurath 1983), he thought, like other members of the Vienna Circle who were influenced by the work of the early Wittgenstein, that there are no meaningful propositions in philosophy (in traditional metaphysics, in particular), and also that philosophy is a quite different activity from science. Generally, in the age of the "linguistic turn" (Gustav Bergmann's phrase), philosophers who endorsed physicalism in one way or another typically tried to provide a reductive analysis of the mental (for example, by means of a dispositional analyses of mental states; see, for example, Carnap 1959; Ryle 1949) or to show that there is no room for the subjective aspects of conscious experience (*qualia*) in corresponding language-games (this was the upshot of Wittgenstein's the-beetle-in-a-box thought experiment; see Wittgenstein 1958, § 295), and the like.

Quine's critiques of the main ideas defended by philosophers of the Vienna Circle[21] inspired many philosophers of that time and led them to think that philosophy and science should not be separated, and that metaphysics (modal discourse and essentialism, in particular) ought to be rejected. In view of the last fact, it is not surprising at all that the proponents of the identity theory, such as Place (1956), famously claimed that their theory "is a reasonable scientific hypothesis". They also believed that statements like "Consciousness is a process in the brain" are contingently true, and that the past successes in providing physical explanations of biological and chemical phenomena give rise to a belief that corresponding theoretical identifications in psychology are available.

---

[21] See, e.g., Quine (1951) for his famous criticism of the analytic/synthetic distinction.

However, Barcan's and Kripke's insights on identity, modality and essence became influential, and increased philosophers' interest to take metaphysics seriously. Although Kripke has famously argued against physicalism (see Kripke 1980 for more details), very soon physicalists tried to reconcile their own views with Kripke's compelling examples of necessary a posteriori statements and his explanation of modal illusions. For example, some physicalists claimed that views like "token physicalism" are even strengthened by the Kripkean insights on the necessary a posteriori statements (see, for example, McGinn 1977), while some others tried to show that terms like "pain" are not rigid designators that pick out their objects of reference through the use of essential modes of presentations (see Lewis 1983; and, more recently, Grahek 2007).[22]

In 1990s Chalmers famously amended conceivability arguments against physicalism, such as the zombie argument, in order to show that his view is compatible with the standard Kripkean cases of the necessary a posteriori. He has elaborated the key notions used in the argument, applied the epistemic version of the two-dimensional semantics, setting up his argument to the effect that the burden of proof has been shifted to physicalists.

Physicalists react to Chalmers's zombie argument in various ways. Some think that phenomenal consciousness can be explained a priori in terms of the physical, while others think that, although there is an explanatory gap between the physical and the mental, this gap still does not entail that there is an ontological gap between the physical and the mental. In other words, the latter argue that the conceivability of zombies does not entail that they are metaphysically possible. There are also physicalists who are ready to redefine physicalism in order to save the day (see, for example, Leuenberger 2008).

This very brief and incomplete outline of some representative physicalists' strategies of dealing with the zombie argument illustrates that it is, contrary to the attitudinal approach, highly unlikely that physicalists voluntarily undertake different activities due to the stances they adopt.

---

[22] In "Mad Pain and Martian Pain", Lewis constructed thought experiments purported to show that "pain" is not rigid designator. On the other hand, Nikola Grahek argued that some interesting cases in neuroscience, such as pain asymbolia, suggest that feeling pain (painfulness) and being in pain can be departed from each other.

Further, it is not unusual that physicalists dispute among themselves over which responses to the anti-physicalistic arguments, such as the zombie argument, are the best. Here, the disputes over the validity of the phenomenal concept strategy can serve as good illustrations.[23] Let us recall that this strategy consists in providing an account that would support the claim that, due to a specific nature of phenomenal concepts, physicalism can be true despite the explanatory gap. Various accounts of that sort have been proposed by now: indexical account, recognitional account, quotational account, and so on (see, for example, Alter & Walter 2006 for more details). On the other hand, setting aside the criticisms coming from anti-physicalists, the phenomenal concept strategy has been criticized by some physicalists more than once. For example, Daniel Stoljar[24] argues that the proponents of the phenomenal concept strategy at best can show that psychophysical conditionals, in which it is claimed that a complete description of the world in physical terms necessitates a complete description of the world in phenomenal terms, are not a priori synthesizable, yet they are not capable of explaining why those conditionals are not a priori.[25] The proponents of phenomenal concept strategy typically try to handle such an objection by providing examples and arguing that the psychophysical conditionals are analogous with some other conditionals that are likely not a priori (see, e.g., Diaz-Leon 2008). It is evident that such a dispute is rationally mandated, contrary to what the attitudinal approach would predict.

---

[23] Intentionalism in philosophy of mind, which is sometimes taken to support physicalism (see, for example, Cutter & Tye 2011; Grahek 2007; Klein 2007), can also serve as a good illustration here, since its proponents often dispute among themselves over which version of intentionalism better explains interesting phenomena. At the same time, there are physicalists, such as Ned Block (see, e.g., Block 1997), who reject intentionalism, typically by claiming that such a theory cannot explain some interesting phenomena (such as blindsight, and the like). This suggests that the debates over the validity of intentionalism are rationally mandated too.

[24] Another critique of the phenomenal concept strategy, posed by a physicalist, comes from Tye (2009).

[25] According to Stoljar (2005, 478), a sentence is a priori synthezisable when "a sufficiently logically acute person who possessed only the concepts required to understand *its antecedent*, is in a position to know that it is true," while a sentence is is a priori when "a sufficiently logically acute person who possessed only the concepts required to understand it, is in a position to know that it is true."

Now, one might think that these considerations turn into a sociological analysis of the physicalist debate, and that it is not clear if such an analysis helps us to address Hempel's Dilemma.[26] As a response to this worry, it should be stressed that these considerations are just partly devoted to a sociological (or a historical) analysis of the physicalist debate: They are primarily aimed to shed a better light on the *rationality* lying behind the willingness of various physicalists to undertake different activities within the same research programme.[27] I hold it is a common practice in philosophy of science to compare competing accounts (for example, Popperian, Kuhnian, Lakatosian accounts, and the like) of the nature of science and scientific rationality by taking into account representative episodes in the history of science, and evaluating to what extent those accounts are realistic in explaining them. This method has been applied outside philosophy of science as well. For example, in his *The Philosophy of Philosophy*, Timothy Williamson writes: "The primary task of the philosophy of science is to understand science, not to give scientists advice. Likewise, the primary task of the philosophy of philosophy is to understand philosophy – although I have not rigorously abstained from the latter" (Williamson 2007, ix). So, I think it is legitimate to apply the same method in assessing competing solutions to Hempel's Dilemma. This means that the considerations above are relevant for assessing competing solutions to Hempel's Dilemma, and that, as it stands, they do not go in favor of the attitudinal approach.

2.2. The Lakatosian approach

Now, let us turn to another strategy of avoiding Hempel's Dilemma, namely that of understanding physicalism as a research programme. In philosophy of science the term "research programme" was famously introduced by Imre Lakatos (see Lakatos 1978), who thought that the units of evaluation in science are not theories but research programmes, within which particular theories and models are produced. According to Lakatos, research programmes guide one's research, and they consist in the "hard-

---

[26] I would like to thank an anonymous referee for drawing my attention to this issue.

[27] The same holds, *mutatis mutandis*, for the cases of conversion, presented in section 2.1.1.

core", positive heuristic, and negative heuristic (cf. Lakatos 1978, 47). The hard-core of a research programme contains basic claims (for example, the three principles of motion in Newton's mechanics), and it is always protected by negative heuristic that redirects potential counterevidence to inessential parts of the programme (to auxiliary hypotheses, etc.). Positive heuristic suggests paths worth of being pursued, the order of investigation, ways to construct models and theories, and so on (see Lakatos 1978, 50). While negative heuristic discourages work on certain theories and models, positive heuristic encourages work on some other theories and models. Also, many competing theories and models might be produced within a single research programme (Newton's mechanics and Darwinism might serve as good examples).

Although it might seem that many philosophers do not explicitly admit of being engaged in a research programme, the fact that they typically try to amend tenaciously their arguments from critiques is a good evidence that they actually are. Of course, philosophers sometimes switch to another research programme, quite the opposite to the one they endorsed earlier (some representative cases of conversion were presented in section 2.1.1).

Recently, a view that physicalism is a research programme has been proposed independently by Guy Dove and Duško Prelević.[28] According to Dove (2016, 5), physicalism is an "ongoing interdisciplinary research programme", the core theses of which are, respectively, that current physics inspired physicalists to count certain entities as physical, and that past suc-

[28] A view that physicalism is a specific *theory* through which materialist (metaphysical) research programme expresses itself at various times is defended by Seth Crook and Carl Gillet (see Crook & Gillett 2001, § 3). However, although physicalism is usually regarded as a descendant of the materialist worldview, it is still rich enough to be understood as a separate research programme: After all, commitment to physics as fundamental science is not a necessary part of materialist metaphysics, whereas, on the other hand, physicalism is arguably incompatible with some materialist views about material substance, like those that were famously criticized by George Berkeley, and the like.

Of course, a natural guess is that Lakatos himself, had he been asked, would have said that physicalism is a research programme: after all, he understood science and ideologies in the same way.

cess in providing physical explanations of biological and chemical phenomena may serve as positive exemplars of how mental phenomena should be explained.

However, putting current physics and past exemplars into the hard-core of physicalist research programme is unsatisfactory for two reasons. First, such an account does not say too much about the nature of current physics. Here, let us recall that today there is more than one interpretation of quantum mechanics, one of which is a dualistic interpretation (see section 1) that is by no means acceptable to our day physicalists. If so, then Dove's account is too permissive. On the other hand, it seems that by positive exemplars in biology and chemistry Dove means well known cases of reductive explanations (functional reductions), to which physicalists frequently appeal in philosophy of mind, such as the explanation of why water is H_2O, why heat is the motion of molecules, why genes are DNA, and the like. Yet, in our times, explanatory pluralism in philosophy of science is a more viable doctrine, which means that physicalists need not be constrained by just one sort of explanation. For example, many phenomena in biology are explained in a non-reductive (and even in a non-causal) way, by appealing to the same level phenomena or even to the higher-level phenomena, as is the case, for instance, with the statistical explanations in theoretical population biology (see Walsh 2015 for further details), and the topological explanations that are used to explain, for example, metabolic economy, synchronicity, stability, robustness, resilience, and the like (see Kostić 2016). Furthermore, there are physicalistic views in philosophy of mind, such as the higher-order theory of consciousness (see, for example, Rosenthal 2011), which aim to provide a (second-order) representational account of consciousness, in which physics or any lower-level theory plays no role. Thus, it seems that Dove's solution is too restrictive concerning the sorts of explanations available to our day physicalists in dealing with mental phenomena.

In contrast to the solution above, Prelević (2017) understands physicalism as a research programme by putting some positive aspects of the term "physical" into the hard-core, such as the claim that a necessary condition for something to be a physical object is to be located in space and time, that is, that what physics generally deals with is, as Chalmers puts it, "structure and dynamics of the world throughout space and time" (Chalmers 1996,

36).[29] According to Prelević, this is possible because the hard-core of the physicalist research programme need not be *fully* specified: further specifications of the core claims of the physicalist research programme belong to its *positive heuristic*, in which many physical models and theories have been proposed.[30] This way, it would be possible to handle the standard problems related to *via negativa*,[31] which concern the (im)possibility of delineating physicalism from views such as the Russellian monism, which, historically speaking, were not counted as physicalistic.[32] Another virtue of the proposal just sketched is that it avoids the problems typical of Dove's solution above, since, on the one hand, it does not rely upon accepting current physics as such, while, on the other hand, it is quite compatible with explanatory pluralism.

Now, let us assess the explanatory power of the Lakatosian solution (as proposed in Prelević 2017) to Hempel's Dilemma. First, it is easy to notice that understanding physicalism as a research programme matches well with the standard classifications in the history of philosophy. Within the Lakatosian account, philosophers like Descartes, Leibniz and Chalmers might easily be classified as philosophers who develop research programmes involving core theses different from the core these of physicalist research programme.

As for the cases of conversion, mentioned in section 2.1.1., they are quite in accordance with the Lakatosian approach too. Lakatos's view of

---

[29] Physicalist research programme also includes, according to Prelević, a view that listing the furnishings of the universe is the subject-matter of physics.

[30] In that respect, one's views about the real nature of space and time (for example, whether space is three-dimensional or configurational, whether space and time are independent of each other or it is better to speak about space-time, and the like), about how many properties belong to the fundamental level, about the nature of the laws of physics, and so on, depend on a physical theory one adopts (see Chalmers 1996, 119 for more details), which is a part of the positive heuristic of physicalist research programme.

[31] Here, *via negativa* is a view that "physical" is best defined negatively, like the "non-fundamentally mental". This view is originally introduced as a version of futurism (see Montero 1999), but it can be also incorporated into the hard-core of physicalist research programme.

[32] That is because, according to Russellian monism, neither physical properties nor mental properties are counted as fundamental; see Judisch (2006) for more details.

research programmes is in many respects akin to Kuhn's view of paradigms (see, for example, Kuhn 1970, 238), but one crucial difference among them was that Lakatos rejected Kuhn's incommensurability thesis, arguing that research programmes can be compared within a given time interval (for example, by comparing to what extent they are fruitful or degenerate). In that respect, the cases of conversion above can be understood as those in which philosophers simply switched to another research programme, and got started developing arguments and accounts in light of new challenges.[33] Finally, the Lakatosian approach alone can explain why different physicalists undertake different activities. As stressed above, this account can easily explain the differences between various physicalists by positing their novelties and disputes in *positive heuristic* (see Prelević 2017, footnote 11). Bearing this in mind, there is no need for introducing stances (in van Fraassen's sense) in order to explain why different physicalists undertake different activities within one and the same research programme. Representative examples, addressed in section 2.1.2, suggest that physicalists do that not because they voluntarily adopt different stances, but due to some arguments they find convincing.

3. Conclusion

Previous considerations suggest that research programmes can be pursued without invoking van Fraassenian stances, and that there is no need for adding stances in order to explain the nature of physicalism and resolve Hempel's Dilemma thereof. Choices that physicalists make are far from not being rationally mandated, contrary to what Rowbottom's view of the role of stances would predict. On the other hand, these practices are compatible with a view that physicalism is a research programme within which different solutions are proposed and compared. In view of the last fact, it is highly unlikely that anything should prompt us to introduce van Fraassenian notion of stance in order to understand what physicalism is, what all physicalists have in common, and how to explain the differences among

[33] This is evident, for example, in the case of Chalmers, who updates his amendments of the zombie argument from time to time, taking into consideration new criticisms (see, for example, Chalmers 2010).

them. In other words, adding stances, which is a hallmark of the attitudinal approach, is not a boon.

Acknowledgments

An earlier draft of this paper was presented at the international conference EENPS 2016, which took place in Sofia (24–26 June, 2016). I would like to thank the organizers and the audience for their input. I am also very much indebted to Daniel Kostić for stimulating discussion of these matters, and the anonymous reviewer for *Organon F* for valuable and very helpful comments and suggestions provided.

This research was supported by Ministry of Education, Science and Technological Development of the Republic of Serbia (project: *Logico-epistemological foundations of science and metaphysics*, No. 179067).

References

ALTER, T. & WALTER, S. (eds.) (2006): *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford: Oxford University Press.

BAUMANN, P. (2011): Empiricism, Stances, and the Problem of Voluntarism. *Synthese* 178, No. 1, 27-36.

BLOCK, N. (1997): On a Confusion about a Function of Consciousness. In: Block, N., Flanagan, O. & Güzeldere, G. (eds.): *The Nature of Consciousness: Philosophical Debates*. Cambridge (Mass.): MIT Press.

BOURGET, D. & CHALMERS, D. (2014): What Do Philosophers Believe? *Philosophical Studies* 170, No. 3, 465-500.

CARNAP, R. (1959): Psychology in Physical Language. In: Ayer, A. (ed.): *Logical Positivism*. New York: The Free Press, 165-198.

CHAKRAVARTTY, A. (2011): A Puzzle about Voluntarism about Rational Epistemic Stances. *Synthese* 178, No. 1, 37-48.

CHALMERS, D. (1996): *The Conscious Mind*. Oxford: Oxford University Press.

CHALMERS, D. (2010): The Two-Dimensional Argument against Materialism. In: Chalmers, D. (ed.): *The Character of Consciousness*. Oxford: Oxford University Press, 141-205.

CHURCHLAND, P. S. (1986): *Neurophilosophy*. Cambridge (Mass.): MIT Press.

CHURCHLAND, P. (1996): The Rediscovery of the Light. *The Journal of Philosophy* 93, No. 5, 211-228.

CROOK, S. & GILLETT, C. (2001): Why Physics Alone Cannot Define 'Physical': Materialism, Metaphysics, and the Formulation of Physicalism. *Canadian Journal of Philosophy* 31, No. 3, 333-359.

CUTTER, B. & TYE, M. (2011): Tracking Representationalism and the Painfulness of Pain. *Philosophical Issues* 21, No. 1, 90-109.

DENNETT, D. (1991): *Consciousness Explained*. Boston: Little, Brown and Co.

DESCARTES, R. (1985): The World or Treatise on Light. In: Cottingham, J., Stoothoff, R. & Murdoch, D. (eds.): *The Philosophical Writings of Descartes*. Vol. 1. Cambridge: Cambridge University Press, 81-98.

DIAZ-LEON, E. (2008): Defending the Phenomenal Concept Strategy. *Australasian Journal of Philosophy* 86, No. 4, 597-610.

DOVE, G. (2016): Redefining Physicalism. *Topoi*, online first. DOI 10.1007/s11245-016-9405-0.

DOWELL, J. (2006): The Physical: Empirical, Not Metaphysical. *Philosophical Studies* 131, No. 1, 25-60.

GRAHEK, N. (2007): *Feeling Pain and Being in Pain*. 2nd ed. Cambridge (Mass.): MIT Press.

GREENE, B. (2004): *The Fabric of the Cosmos*. New York: Vintage.

HEMPEL, C. (1980): Comments on Goodman's Ways of Worldmaking. *Synthese* 45, No. 2, 193-200.

JACKSON, F. (2006): The Knowledge Argument, Diaphanousness, Representationalism. In: Alter, T. & Walter, S. (eds.): *Phenomenal Concepts and Phenomenal Knowledge*. Oxford: Oxford University Press, 52-64.

JUDISCH, N. (2008): Why 'non-mental' won't work: on Hempel's dilemma and the characterization of the 'physical'. *Philosophical Studies* 140, No. 3, 299-318.

KIRK, R. (1974): Sentience and Behaviour. *Mind* 83, No. 329, 43-60.

KIRK, R. (2006): Physicalism and Strict Implication. *Synthese* 151, No. 3, 523-536.

KIRK, R. (2008): The Inconceivability of Zombies. *Philosophical Studies* 139, No. 1, 73-89.

KLEIN, C. (2007): An Imperative Theory of Pain. *Journal of Philosophy* 104, No. 10, 517-532.

KOSTIĆ, D. (2016): The Topological Realization. *Synthese*, online first. DOI 10.1007/s11229-016-1248-0.

KRIPKE, S. (1980): *Naming and Necessity*. Cambridge (Mass.): Harvard University Press.

KUHN, T. (1962): *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

KUHN, T. (1970): Reflections on my Critics. In: Lakatos, I. & Musgrave, A. (eds.): *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press, 231-278.

KUHN, T. (1974): Second Thoughts on Paradigms. In: Suppe, F. (ed.): *The Structure of Scientific Theories*. Urbana: Urbana Illinois Press, 459-482.

LADYMAN, J. (2011): The Scientistic Stance: the Empirical and Materialist Stances Reconciled. *Synthese* 178, No. 1, 87-98.

LAKATOS, I. (1978): Falsification and the Methodology of Scientific Research Programmes. In: Worrall, J. & Currie, G. (eds.): *The Methodology of Scientific Research Programmes*: Philosophical Papers. Vol. I. Cambridge: Cambridge University Press, 8-102.

LEIBNIZ, G. W. (1997): [First] explanation of the New system of the communication between substances, in Reply to what was said of it in the Journal for 12 September 1695. In: Woolhouse, R. & Francks, R. (eds.): *Leibniz's 'New System' and Associated Contemporary Texts*. Oxford: Clarendon Press, 47-52.

LEUENBERGER, S. (2008): Ceteris Absentibus Physicalism. In: Zimmerman, D. W. (ed.): *Oxford Studies in Metaphysics*. Oxford: Oxford University Press, 4-145.

LEWIS, D. (1983): Mad Pain and Martian Pain. In: *Philosophical Papers*. Vol 1. Oxford: Oxford University Press, 122-130.

MCGINN, C. (1977): Anomalous Monism and Kripke's Cartesian Intuitions. *Analysis* 37, No. 2, 78-80.

MONTERO, B. (1999): The Body Problem. *Noûs* 33, No. 2, 183-200.

NEURATH, O. (1983): Physicalism: The Philosophy of the Viennese Circle. In: Cohen, R. & Neurath, M. (eds.): *Otto Neurath: Philosophical Papers 1913-1946*. Dordrecht: D. Riedel Publishing Company, 48-51.

NEWTON-SMITH, W. (1981): *The Rationality of Science*. London: Routledge & Kegan Paul, Ltd.

NEY, A. (2008a): Defining Physicalism. *Philosophy Compass* 3, No. 5, 1033-1048.

NEY, A. (2008b): Physicalism as an Attitude. *Philosophical Studies* 138, No. 1, 1-15.

PAPINEAU, D. (2001): The Raise of Physicalism. In: Gillett, C & Loewer, B. (eds.): *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, 3-36.

PLACE, U. T. (1956): Is Consciousness a Drain Process? *British Journal of Psychology* 47, No. 1, 44-50.

PRELEVIĆ, D. (2017): Physicalism as a Research Programme. *Grazer Philosophische Studien*, online first. DOI 10.1163/18756735-000023.

QUINE, W. V. O. (1951): Two Dogmas of Empiricism. *The Philosophical Review* 60, No. 1, 20-43.

RYLE, G. (1949): *The Concept of Mind*. New York: Barnes and Noble.

ROSENTHAL, D. (2011): Exaggerated reports: reply to Block. *Analysis* 71, No. 3, 431-437.

ROWBOTTOM, D. (2011): Stances and Paradigms: A Reflection. *Synthese* 178, No. 1, 111-119.

SMART, J. J. (1959): Sensations and Brain Processes. *The Philosophical Review* 68, No. 2, 141-156.

STOLJAR, D. (2005): Physicalism and Phenomenal Concepts. *Mind & Language* 20, No. 2, 469-494.

TELLER, P. (2004): Discussion: What Is a Stance? *Philosophical Studies* 121, No. 2, 159-170.

TYE, M. (2009): *Consciousness Revisited: Materialism without Phenomenal Concepts*. Cambridge (Mass.): MIT Press.

VAN FRAASSEN, B. (2002): *The Empirical Stance*. New Haven & London: Yale University Press.

VAN FRAASSEN, B. (2004): Replies to the Discussion on The Empirical Stance. *Philosophical Studies* 121, No. 2, 171-192.

WALSH, D. M. (2015): Variance, Invariance and Statistical Explanation. *Erkenntnis* 80, Supplement 3, 469-489.

WILLIAMSON, T. (2007): *The Philosophy of Philosophy*. Oxford: Blackwell Publishing.

WILSON, J. (2006): On Characterizing the Physical. *Philosophical Studies* 131, No. 1, 61-99.

WITTGENSTEIN, L. (1958): *Philosophical Investigations*. Oxford: Basil Blackwell.

WOOLHOUSE, R. (1985): Leibniz's Reaction to Cartesian Interaction. *Proceedings of the Aristotelian Society* 86, 69-82.

Organon F 24 (4) 2017: 511-525

The Role of Priors in a Probabilistic Account of "Best Explanation"

ANTON DONCHEV[1]

ABSTRACT: In this paper, I argue that the notion of "best explanation", as it appears in the Inference to the Best Explanation (IBE), can be defined in terms of explanatory power (EP) (i.e. the best explanation among a set of possible explanations is the one having the highest EP), if we employ a probabilistic measure of EP, which takes into account both the likelihoods and the prior probabilities of the compared explanatory hypotheses. Although the association between the EP of a hypothesis and its likelihood is largely uncontroversial, most of those working on EP do not see an association between EP and the prior probability of an explanatory hypothesis. I provide three examples (two toy examples and one from real scientific practice), in order to show that the role of priors in decisions about the best explanatory hypothesis deserves a serious consideration. I also show that such an explication of "best explanation" allows us to compare IBE and Bayesian confirmation theory (BCT) in terms of the probabilities they assign to two competing hypotheses, and thus to elicit the conditions under which both IBE and BCT lead to the same conclusion and are in this sense compatible.

KEYWORDS: Bayesianism – Bayesian confirmation theory (BCT) – Inference to the Best Explanation (IBE) – explanatory power – prior probabilities.

[1] Received: 31 January 2017 / Accepted: 31 August 2017

✉ Anton Donchev
Department of Philosophy and Sociology
New Bulgarian University
21 Montevideo Street, 1618 Sofia, Bulgaria
e-mail: donchev.anton@gmail.com

1. Introduction

The present state of the literature on Inference to the Best Explanation (IBE) reveals two problems of considerable importance. On one hand, there is no clear explication of the term "best explanation". The main idea driving the need for such an explication is that, in order to "infer to the *best explanation*" out of a set of competing explanations, we need a clear way of comparing and/or grading the explanations within that set. However, one of the best-known accounts of a mechanism of comparing explanations, i.e. Lipton's (2004), uses the term "loveliest explanation", where "loveliest" is an umbrella term for a set of undefined number of explanatory virtues, such as simplicity, unification or scope, most of which also lack clear formal explications (cf. Norton 2016). Another example of the same problem is the classical (or textbook) form of IBE, which is often expressed by the following rule:

> Given evidence E and candidate explanations $H_1, \ldots H_n$ of E, infer the truth of *that* H_i which best explains E. (Douven 2011)

The above rule fails to answer the crucial question at the heart of IBE – how do we find which is the best explanation of the evidence, out of a set of competing explanatory hypotheses. Its failure in that respect has in fact prompted Schurz to claim that IBE "is epistemically rather uninformative" (Schurz 2008, 204).

On the other hand, there is the issue of IBE's compatibility with Bayesian Confirmation Theory (BCT). Incompatibilist philosophers of confirmation claim that IBE and BCT are two irreconcilable methods of confirmation, of which only one is rational. Here belong arguments such as van Fraassen's, who claims that any probabilistic formulation of IBE should either: be equivalent to Bayes' rule, and is thus redundant; or, if it is not equivalent to Bayes' rule, it has to provide a satisfactory answer to the Dutch book argument (cf. van Fraassen 1989). Another argument for IBE – BCT incompatibilism is the claim that "confirmation is *logically* independent of explanation" (Salmon 2001, 88), and so explanatory considerations, such as those driving IBE, should not enter into a method of confirmation, such as BCT. In the end, most incompatibilists' accounts are skeptical towards the role of IBE as a genuine rule of non-deductive inference (see also Iranzo 2008; Norton 2016).

However, both incompatibilists and compatibilists tend to view the issue of IBE – BCT compatibility as a two-sided matter. Either these approaches to confirmation are deemed completely incompatible – as one is rational and the other is not, or they are deemed completely compatible and assumed to work in tandem. The problem with such views is that there are different possible explications of IBE, and different models of BCT. Some of these may turn out to be compatible, while others may not. What is more, an IBE explication may be compatible with a specific Bayesian model of confirmation only *under certain conditions*.

Ultimately, the question whether IBE and BCT are compatible or not will depend upon a future investigation into these conditions; and in order to enable such an investigation, IBE and BCT should first be translated into the same language. The key to such a translation is to find an adequate probabilistic explication of the term "the best explanation". There have been several attempts in the literature to give such explications of "the best explanation", in the form of measures of explanatory power (EP).[2] However, no direct measure of EP that I know of accounts for the prior probabilities of the explanatory hypotheses. As I show in the next part, insensitivity to priors may lead to very counterintuitive conclusions in certain cases. Therefore, it seems that an adequate probabilistic explication of "the best explanation" should take into account not only likelihoods, but priors as well. Interpreting IBE probabilistically in this way has three distinct advantages. First, it provides a mechanism for actually finding the best explanation. Second, it can account for cases in science, which can be accounted for by neither classical IBE, nor explications of "the best explanation" insensitive to priors. Third, it enables comparisons between IBE and specific Bayesian models of confirmation, in order to investigate the conditions under which these two approaches to confirmation may turn out to be compatible.

2. A probabilistic measure of EP should account for priors

A viable approach to solving the first problem outlined in the introduction – IBE's lack of mechanism of comparing competing explanations – is

[2] For lists of such measures see e.g., Schupbach (2011) and Glymour (2015).

to seek a probabilistic explication of the key term "*best explanation*". In other words, we may strive to provide a probabilistic mechanism to compare competing hypotheses in terms of their EP. The question "which one is the best explanation", would then receive the answer "the one that has the highest explanatory power according to such-and-such probabilistic measure". There are a few direct measures of EP in the literature, such as the one by Schupbach & Sprenger (2011):

$$Ep_1(E,H) = \frac{P(H|E) - P(H|\neg E)}{P(H|E) + P(H|\neg E)}$$

Crupi & Tentori (2012) have proposed another measure:

$$Ep_2(E,H) = \begin{cases} \dfrac{P(E|H) - P(E)}{1 - P(E)} & if\ P(E|H) \geq P(E) \\ \dfrac{P(E|H) - P(E)}{P(E)} & if\ P(E|H) < P(E) \end{cases}$$

In addition, there are several more direct measures of EP, which have been created from different measures of confirmation (see Schupbach 2011).[3]

So far, all proposed direct measures of EP share a common characteristic – they are not influenced by prior probabilities. For example, the measure of Schupbach & Sprenger (2011) is, at first glance, dependent on posterior probabilities and thus on the priors which form them, yet calculations reveal that this is not the case and the priors actually cancel each other out.

However, there are examples, which seem to show that prior probabilities play a major role in our decisions about the best explanation. Consider the following simple case:

---

[3] There are also some probabilistic measures of unification or coherence, which have been proposed as indirect measures of EP, e.g., Myrvold (2003), Fitelson (2003), Glass (2007), Wheeler (2009). These fall outside the scope of the current argument, which focuses on direct measures of EP.

Waking up in the morning you see that the grass is wet (E). You form two hypotheses explaining this fact:

H_1: "It rained tonight."

H_2: "The gardener watered the lawn earlier in the morning."

These hypotheses have the same likelihoods, i.e. $P(E|H_1) = P(E|H_2)$, because both H_1 and H_2 entail E (the wet lawn). Suppose, however, that you made this observation in July and you live in a place where rains in July are extremely rare. Based on this background knowledge, you assign a very low prior probability to H_1. What is more, intuitively, the gardener watering the lawn is a far better explanation of the wet grass, than the extremely improbable rain in July. In order to capture that intuition, a measure of EP should account for prior knowledge. In other words, it should reach the intuitive conclusion, that even though the likelihoods of the two hypotheses are the same, the one with the higher prior better explains the evidence you have observed. The likelihood-only based measures cannot reach that result, and are forced to conclude that H_1 and H_2 are of equal EP, which is highly unintuitive in that case.

The above example is quite simplistic, so let us push the argument for the importance of priors in EP with a second, more complex example:

Patient X (aged 45) has paresis (E). This could be the result of various medical conditions, but for the sake of simplicity, we will take into account only two:

H_1: "X had untreated syphilis."

H_2: "X had a stroke."

By previously consulting X's medical record, as well as various other sources of medical information, his physician brings into the case the following information:

i) X was diagnosed with syphilis, but not treated for it: $P(H_1) = 0.9$;
ii) About 25% of those who have untreated syphilis get paresis in later age: $P(E|H_1) = 0.25$;
iii) About 80% of stroke survivors get some kind of paresis: $P(E|H_2) = 0.8$;

iv) The physician does not know whether X had a stroke, but she knows that about $0.2\% - 0.4\%$ of the population of his age are at a high risk of stroke, i.e. $P(H_2) \approx 0.004$.

Although the likelihood of the stroke hypothesis is greater than the likelihood of the paresis one (i.e. $P(E|H_2) > P(E|H_1)$), most physicians, given the information in i) – iv) would assign higher EP to H_1. In other words, most physicians would explain the paresis with the untreated syphilis. What this example aims to illustrate again is that priors may play an important role in some decisions about the best explanation, so much so, that they may overturn a large difference in likelihoods. A probabilistic measure of EP, which is not sensitive to priors, and depends solely on likelihoods, will not be able to account for such cases, i.e., if we applied such a measure to these kind of cases, we would be led to counterintuitive results.

In summary, an adequate probabilistic interpretation of IBE should aim at a probabilistic explication of the key term "best explanation", thus giving IBE the means to answer the question "how do we find *the best* amongst competing explanations". This is a clear advantage over classical IBE, which is silent on this crucial question. The explication of the "best explanation" would be in the form of a probabilistic measure of EP; however, the measure should be influenced by the prior probabilities of the evaluated hypotheses, in contrast to the direct measures of EP proposed in the literature. If the measure does not account for priors, it runs into two kinds of problems. On the one hand, it will provide counterintuitive results in certain cases, as illustrated by the above toy-examples. On the other hand, it will fail to account for real cases in science, as will be shown in the next part.

Providing and defending the prescribed new measure of EP are aims for future research, which fall outside the scope of the current paper. The focus here is on arguing for the important role of priors in a probabilistic explication of the "best explanation". We will now turn towards a third example for their importance, this time from real scientific practice.

3. The role of priors in scientific practice: the case of Planet Nine

In January 2016, two Caltech astronomers – Konstantin Batygin and Michael Brown – inferred the existence of a still unobserved planet in the outer Solar System (cf. Batygin & Brown 2016). This conjecture was made in order to provide the best explanation of certain peculiarities in the otherwise stable orbits of a set of trans-Neptunian objects. It was observed that six objects in the scattered disk of the Kuiper Belt (Kuiper Belt Objects or KBOs), which had perihelia greater than the orbit of Neptune, and semi-major axes greater than 150 AU ($a > 150$ AU), exhibited a strange clustering of their arguments of perihelion ($\omega \sim 0$). In other words, the perihelion of every one of these objects lied on the ecliptic, and their ascending nodes coincided with their perihelia, i.e. they all shared the same orbital direction – from south to north. Batygin and Brown calculated that orbits with $a > 50$ AU, clustered this tight, occur only 0.007% of the time, which means a probability of only 0.00007 that the clustering is due to chance. They stated that, considering the age of our Solar System, such groupings are expected to randomize, unless held together by some physical mechanism.

At that point in 2016, the above peculiarities in the six KBOs' orbits have already been noted, and there were three contending explanatory hypotheses. The first one was proposed by Trujillo & Sheppard (2014). They concluded that the observed perihelia, which librated around $\omega = 0$, might be held by a massive body on an outer orbit, about five times the mass of Earth. The second explanatory hypothesis was that the observed phenomenon was due to a self-gravitational instability of the scattered disk population of the Kuiper Belt (see Madigan & McCourt 2015). The third one was Batygin and Brown's own model, predicting the existence of an unobserved planed. Batygin and Brown also systematically criticized the other two explanations.

As for the first one, the mechanism employed to explain the data in Trujillo & Sheppard (2014) had certain assumptions that would require the existence of several massive bodies, on orbits exactly tailored in order to explain the peculiarities in the orbits of the six KBOs. Furthermore, the same mechanism could not explain by itself why we observe objects clustered at $\omega \sim 0$, but there is no such observed clustering in $\omega \sim 180$ (see

Batygin & Brown 2016). This explanation required the assumption that our Solar System had a strong stellar encounter in the past – an assumption that does not fit with the rest of our knowledge about the Solar System. All of these assumptions reduce the quality of Trujillo & Shepard's (2014) explanation of the KBOs' orbits, making it more ad hoc.

As for the second explanation by Madigan and McCourt (2015), which employed a so called "inclination instability", it assumed the scattered disk of the Kuiper Belt was once much more massive than current estimations, and stayed that way for a prolonged period of time, or it could not have provided enough gravity for the proposed mechanism of instability. Not only do we lack sufficient evidence for such an assumption, but also it is highly unlikely for theoretical reasons. Most of the mass that the disk might have contained in the past was most probably ejected from the Solar System due to interactions with the gas and ice giants. As Batygin & Brown (2016) noted, such interactions usually end up in hyperbolic trajectories for the less massive objects.

The best available explanation of the clustering of the six KBOs is that there is a massive body of about or above ten Earth masses, with a semi-major axis $a \sim 700$ AU and an estimated perihelion of about 200 AU, and an aphelion of about 1200 AU, which has eluded observation so far (Batygin & Brown 2016). It is speculated that this "perturber" of the orbits of the KBOs would likely be an ice giant, formed by an ejected giant planet core during the early phases of development of our Solar System. It probably has an orbital period in the range of 10 to 20 thousand years, and most of the time it is too far from Earth to be observed without very high-resolution equipment, which would explain why it has not yet been discovered. If its existence is confirmed by observation, the planet will receive an official name, but in the meantime, it has been called "Planet Nine". Batygin & Brown (2016) point out that their explanation of the clustering of the six KBOs by the existence of Planet Nine also has implications about other features of the Kuiper Belt, which not only increase the scope of their explanation, but also provide "a direct avenue for falsification of our hypothesis" (Batygin & Brown 2016, 2).

In summary, we have three competing explanatory hypotheses, all of which entail the evidence, i.e., the observed clustering of KBOs. These are: Trujillo & Shepherd's (2014) hypothesis, whose model requires the existence of several undiscovered massive bodies; Madigan & McCourt's

(2015) hypothesis, which presupposes that the scattered disk of the Kuiper Belt was much more massive and for a longer period of time, than current estimations indicate; and Batygin & Brown's (2016) hypothesis, which presupposes the existence of Planet Nine. When interpreting IBE probabilistically, if we explicate the "best explanation" through any of the measures of EP sensitive only to the likelihoods, we would be forced to the conclusion that the above three hypotheses are equally good explanations of the observed evidence. This conclusion, however, will go against the expert opinion of those astrophysicists who believe that Batygin and Brown's hypothesis is the best available explanation (e.g. see opinions by Rodney Gomes, quoted in Lovett 2012, and by Alessandro Morbidelli, quoted in Achenbach & Feltman 2016). If we include in our explication of the "best explanation" the differences in prior probabilities of the competing hypotheses, then this controversy is resolved. Trujillo and Shepherd's hypothesis requires the existence of several massive bodies, whereas Batygin and Brown's hypothesis requires just one. According to the rules of classical probability, the probability for the existence of a single massive body would always be higher than the combined probabilities for the existence of several massive bodies. Madigan and McCourt's hypothesis requires that the Kuiper Belt was once much more massive, and for a longer period, than current estimations indicate. Furthermore, it is unlikely for theoretical reasons – most of that mass would have been quickly ejected out of the Solar System due to planetary encounters. Ceteris paribus, a hypothesis, which is not in agreement with current estimations, and is unlikely from a theoretical point of view on top of that, cannot receive a higher prior than a hypothesis that does not run into such problems. Based on these considerations, Batygin and Brown's hypothesis seems to have the highest prior probability of the three competing explanations thus far. If we include that consideration in our decision about which one of them is the best explanation of the evidence, we would reach a conclusion in accordance with the opinions of the experts. In other words, if our decision about the "best explanation" takes the prior probabilities of the assessed hypotheses in consideration, then it could adequately account for scientific cases, such as the one with Planet Nine.

4. A method for exploring the conditions of IBE – BCT compatibility

Another advantage of the described probabilistic explication of "best explanation" is that it makes investigating the conditions under which IBE and BCT are compatible relatively straightforward.[4] As was already mentioned in the Introduction, the problem of IBE – BCT compatibility depends on the particular explication of IBE and the particular Bayesian model of confirmation we want to compare. Furthermore, an IBE explication and a Bayesian model may turn out to be compatible only if certain conditions hold. How do we know which conditions should hold? We may find that out through a method of comparing inequalities, and we will consider an example demonstrating how the method works.

For the sake of argument, let us take as a probabilistic explication of "best explanation" the following simple measure of EP:

(1) $Ep(E,H) = P(E|H) \times P(H)$

In other words, let us assume that *the best explanation*, out of a set of competing explanatory hypotheses, is the one that has the highest value of Ep. We interpret IBE to mean that this hypothesis is also *the most confirmed one*.

This specific measure of EP was chosen because it cannot have values above 1 or below 0, which would be hard to interpret and would violate the axioms of classical probability. It also accounts for prior probabilities, as has already been argued. Nevertheless, it is introduced for the purposes of this example and should not be taken as a proposal to measure EP in real life.

Using the above measure, we will define the EP of two competing explanatory hypotheses H_1 and H_2, both of which explain some empirical evidence E:

(2) $Ep(E,H_1) = P(E|H_1) \times P(H_1)$

(3) $Ep(E,H_2) = P(E|H_2) \times P(H_2)$

[4] By "compatibility", I will understand the ability to provide equal results, when applied to the same case. Although there could be other meanings of the term, these will not be addressed here.

We would like to know under which conditions our interpretation of IBE may turn out to be compatible, in the sense of providing compatible results, with the measure of confirmation proposed by Eells (1982) and defended by Jeffrey (1992):

(4) $C(E,H) = P(H|E) - P(H)$

In other words, according to Eells and Jeffrey, confirmation is an increasing function of the difference between posterior and prior probabilities. We will define the measures of confirmation of our two hypotheses as:

(5) $C(E,H_1) = P(H_1|E) - P(H_1)$

(6) $C(E,H_2) = P(H_2|E) - P(H_2)$

We start comparing the two methods by exploring the scenario in which H_1 is better confirmed by E than H_2. According to our interpretation of IBE, H_1 is better confirmed than H_2 if the following condition is satisfied:

(7) $Ep(E,H_1) > Ep(E,H_2)$

(8) $P(E|H_1) \times P(H_1) > P(E|H_2) \times P(H_2)$

And according to our chosen Bayesian measure of confirmation, H_1 is better confirmed than H_2 if:

(9) $C(E,H_1) > C(E,H_2)$

(10) $P(H_1|E) - P(H_1) > P(H_2|E) - P(H_2)$

We may immediately notice that, as both H_1 and H_2 explain the same evidence, we may transform (8) by dividing both sides of the inequality by $P(E)$, assuming $P(E) > 0$:

(11) $\frac{P(E|H_1)\times P(H_1)}{P(E)} > \frac{P(E|H_1)\times P(H_1)}{P(E)}$

(12) $P(H_1|E) > P(H_2|E)$

We may transform (10) into:

(13) $P(H_1|E) - P(H_2|E) > P(H_1) - P(H_2)$

After which we may transform (12) into:

(14) $P(H_1|E) - P(H_2|E) > 0$

Now, if we assume that:

(15) $P(H_1) - P(H_2) \geq 0$

From (13) and (15) we can infer:

(16) $P(H_1|E) - P(H_2|E) > 0$

As (14) and (16) are obviously equivalent, we may argue that (13) and (14) are also equivalent, given that (15) is satisfied. In other words, both IBE and BCT would conclude that H_1 is better confirmed by E than H_2 if $P(H_1) \geq P(H_2)$, and that when IBE and BCT lead to different predictions, this is due to a violation of condition (15).

Now that we know the above condition, one way to proceed would be to rationalize why it should hold. However, if the result is deemed indefensible or strongly counterintuitive, another way to proceed is to seek a different measure of explanatory power, or a different Bayesian measure of confirmation, and employ the method again to find if they are compatible and under what conditions.

The bottom line is that investigating the conditions of IBE – BCT compatibility, by employing the method presented above, lends itself naturally to an interpretation of IBE, which uses a probabilistic explication of "best explanation" (as outlined in section 2). This is an advantage of the probabilistic interpretation of IBE over classical IBE, as it allows us to explore the issue of IBE – BCT compatibility in much higher detail (i.e., on a model-by-model basis), rather than just announcing them completely compatible or incompatible.

5. Conclusions

A probabilistic interpretation of IBE should aim at providing an adequate probabilistic explication of the term "best explanation" – in this way it would be able to complete IBE with a formal mechanism for finding the best out of a set of explanations. An *adequate* probabilistic explication of

the "best explanation" should be a measure of explanatory power influenced by the prior probabilities of the explanatory hypotheses. There are cases, which show that priors have a key role in decisions about the best explanation, and that measures, which are sensitive only to the likelihoods of the assessed hypotheses, would lead to counterintuitive conclusions, when applied to these cases.

Such a probabilistic interpretation of IBE has three distinct advantages. The first one is the above-mentioned formal mechanism for finding the best explanation, whereas classical IBE lacks such a mechanism. The second advantage is that it can account for cases in science as the one with Planet Nine. Interpretations of IBE, which use probabilistic measures of explanatory power that are not influenced by priors, fail to account for such cases. They would consider all competing hypotheses as equally good explanations of the evidence, against the experts' better judgment, whereas a priors-sensitive measure would be able to explain why one of the hypotheses is considered a superior explanation.

The third advantage is that such a probabilistic interpretation of IBE allows for investigation of the particular conditions, under which specific explications of IBE and specific Bayesian models of confirmation give compatible results. The issue of IBE – BCT compatibility is more nuanced than outright compatibility or incompatibility: there are different explications of IBE and different models of BCT. Some of these may turn out to be compatible, but only under certain conditions. In order to resolve the issue, these conditions will have to be explored in future research, which may be done in a straightforward way by employing the method presented in the previous part.

There are also several other topics, which remain open for further research. The main one is to provide a probabilistic measure of EP that accounts for priors and test it against the already proposed measures of EP. There is also a decision to be made whether competing hypotheses should have EP above a certain threshold, in order to be considered "good enough" in Lipton's term, and avoid van Fraassen's (1989) "argument from a bad lot". Last but not least, introducing priors in EP may give rise to objections against making EP "subjective", similar to the objections against BCT. These objections will have to be addressed, and one way to do it is to argue that priors may be formed by considerations about simplicity, unification, scope and other explanatory virtues.

Acknowledgments

I would like to thank Lilia Gurova for her helpful comments and ideas on earlier drafts of this paper.

References

ACHENBACH, J. & FELTMAN, R. (2016): New Evidence Suggests a Ninth Planet Lurking at the Edge of the Solar System. *The Washington Post*. [Accessed 17 June 2017] Available at: https://www.washingtonpost.com/news/speaking-of-science/wp/2016/01/20/new-evidence-suggests-a-ninth-planet-lurking-at-the-edge-of-the-solar-system/.

BATYGIN, K. & BROWN, M. (2016): Evidence for a Distant Giant Planet in the Solar System. *The Astronomical Journal* 151, No. 22, 1-12.

CRUPI, V. & TENTORI, K. (2012): A Second Look at the Logic of Explanatory Power (with Two Novel Representation Theorems). *Philosophy of Science* 79, No. 3, 365-385.

DOUVEN, I. (2011): Abduction. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy.* (Summer 2017 Edition). Available at: https://plato.stanford.edu/archives/sum2017/entries/abduction/

EELLS, E. (1982): *Rational Decision and Causality*. Cambridge: Cambridge University Press.

FITELSON, B. (2003): A Probabilistic Theory of Coherence. *Analysis* 63, No. 3, 194-199.

GLASS, D. H. (2007): Coherence Measures and Inference to the Best Explanation. *Synthese* 157, No. 3, 275-296.

GLYMOUR, C. (2015): Probability and the Explanatory Virtues. *The British Journal for the Philosophy of Science* 66, No. 3, 591-604.

IRANZO, V. (2008): Bayesianism and Inference to the Best Explanation. *Theoria* 23, No. 1, 89-106.

JEFFREY, R. (1992): *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

LIPTON, P. (2004): *Inference to the Best Explanation.* 2nd ed. London: Routledge.

LOVETT, R. (2012): New Planet Found in Our Solar System? *National Geographic News*. [Accessed 17 June 2017]. Available at: http://news.nationalgeographic.com/news/2012/05/120511-new-planet-solar-system-kuiper-belt-space-science/

MADIGAN, A. M. & MCCOURT, M. (2015): A New Inclination Instability Reshapes Keplerian Discs into Cones: Application to the Outer Solar System. *Monthly Notices of the Royal Astronomical Society* 457, 89-93.

MYRVOLD, W. (2003): A Bayesian Account of the Virtue of Unification. *Philosophy of Science* 70(2), 399-423.

NORTON, J. D. (2016): Inference to the Best Explanation: The General Account. In: *The Material Theory of Induction*. Manuscript [accessed 11 June 2017]. Available at: http://www.pitt.edu/~jdnorton/papers/material_theory/8. Best Explanation General.pdf.

SALMON, W. (2001): Explanation and Confirmation: A Bayesian Critique of Inference to the Best Explanation. In: Hon, G. & Rakover, S. (eds.): *Explanation: Theoretical Approaches and Applications*. Berlin: Springer, 61-91.

SCHUPBACH, J. (2011): Comparing Probabilistic Measures of Explanatory Power. *Philosophy of Science* 78, No. 5, 813-829.

SCHUPBACH, J. & SPRENGER, J. (2011): The Logic of Explanatory Power. *Philosophy of Science* 78, No. 1, 105-127.

SCHURZ, G. (2008): Patterns of Abduction. *Synthese* 164, No. 2, 201-234.

TRUJILLO, C. & SHEPPARD, S. (2014): A Sedna-like Body with a Perihelion of 80 Astronomical Units. *Nature* 507, No. 7493, 471-474.

VAN FRAASSEN, B. (1989): *Laws and Symmetry*. Oxford: Clarendon Press.

WHEELER, G. (2009): Focused Correlation and Confirmation. *The British Journal for the Philosophy of Science* 60, No. 1, 79-100.

Organon F 24 (4) 2017: 526-545

Reason, Science, Criticism

JOSEPH AGASSI INTERVIEWED ON HIS 90TH BIRTHDAY BY ZUZANA PARUSNIKOVÁ[1]

Joseph Agassi was born in 1927 in Jerusalem. In 1956 he received his PhD in logic and scientific method in the London School of Economics under the supervision of Karl Popper. Between 1954 and 1957 he worked as Popper's assistant and importantly participated in the preparation of the English translation and the revisions, and the *Postscript*, of *Logik der Forschung*. Joseph Agassi developed his own version of critical rationalism and has always endorsed Popper's suggestion that philosophy should not be sectarian but should apply the open-minded critical attitude to other subjects beyond science. In his books he has addressed a broad range of issues from history and philosophy of science to aesthetics, politics, education, psychiatry, medicine and the relation of science to society and culture. Joseph Agassi was appointed a Professor of Philosophy at Boston University, York University, Toronto and Tel-Aviv University. His rich bibliography includes among other books *Science in Flux, Popper and his Popular critics*, *Towards an Historiography of Science, Faraday as a Natural Philosopher*, *The Continuing Revolution: A History of Physics From the Greeks to Einstein*, *Science and Society, Science and Culture, A Critical Rationalist Aesthetics* (with Ian Jarvie). One should

[1] ✉ Joseph Agassi

37, Levi Eshkol Street
Herzlia 46745 Israel
e-mail: agass@post.tau.ac.il

✉ Zuzana Parusniková

Institute of Philosophy
Czech Academy of Sciences
Jilská 1, 110 00 Praha 1, Czech Republic
e-mail: parusnikova@volny.cz

also mention his intellectual autobiography *A Philosopher's Apprentice: In Karl Popper's Workshop*.

Zuzana Parusniková (**ZP**): You received your secondary education at a theological school. After finishing it you decided to give up further religious (rabbinical) training and entered the university to study physics. That seems to me like a suicidal decision, given that you had no background in science or mathematics. Why physics?

Joseph Agassi (**JA**): I knew that it would be difficult, but not how much. I knew I could study the humanities and social sciences on my own. I also knew I could not do so with physics but I believed that to be serious one had to know some physics beyond popular physics. It is easy for me to explain it now but then it was very difficult, hazy indeed – the popular physics of the time was a mix of inductivism and instrumentalism, both of which sounded to me fishy and some idealism that I knew was rubbish. I had read Eddington and that made this clear to me: he was a philosopher and his philosophy was wrong; I found his idea that metaphysics is private ridiculous. (Popper did, too.)

ZP: Don't you find the Talmudic culture close to rational discourse based on criticism, and thus to science? For instance, Menachem Fisch (in his *Rational Rabbis* and elsewhere) compares Popper's critical rationalism with the Talmudic discourse and argues that the anti-traditionalist camp encourages the critical scrutiny and even revision of the halakhic tradition.

JA: Yes, I found the Talmud rational. But its rationality was insufficient for me. Fisch is a former student of mine yet I often fail to comprehend him. His idea in his *Rational Rabbis* is right but he is apologetic about it. I consider the term "apologetics" in the clear and traditional meaning. Look at Maimonides, one of the most rationalist thinkers of all times. He said, it is permitted to argue rationally in favor of the commandments, but not against them. This is apologetic. Yet Talmud is staunch on one thing: always let serious science have the last word.

ZP: When you came to London you were trying to find your own philosophical way. As you say, meeting Popper was a crucial, eye-opening

moment in this search; could you describe what appealed to you most, in Popper as a person and in critical rationalism? At that time, *The Logic of Scientific Discovery* had not yet been published, and you had not read *The Open Society*.

JA: I met Popper in January 1953. I decided to be his student after I had read a paper of his. And when I listened to a lecture of his I was sold. I became his assistant in fall 1954 to fall 1957. His logic lectures showed logic as goal directed – erotetic – and his epistemology was utterly skeptical but utterly optimistic. He chatted with students in class as equals. I read *The Open Society and Its Enemies* in Easter vacation 1953 and was very impressed. Edited the translation of the book into Hebrew a few years ago and was more so. I read drafts of the translation of *The Logic of Scientific Discovery* in 1955 and corrected them without having the original. Worked on it till it appeared, including the index.

ZP: It seems you were attracted mainly by two things in Popper – by his open-minded attitude to philosophy and by his willingness to discuss a wide variety of issues beyond science.

JA: Yes, Popper's broad cast annoyed Eva Cassirer. She said: "How dare he mention Mozart in a logic lecture?" It delighted me. I had much dialectical background before having met Popper and I left the Jewish dialectic as too apologetic. When I heard Popper present logic as dialectic I was thrilled. I realized at once that his view of science was the view of scientific method as dialectic. That delighted me, too.

I was attracted by his genuine freedom from convention. He was ready to discuss logic and music together if the discussion went this way, he allowed ignorant people to express their opinion freely as long as they were easy about it, he zeroed in on interest without bothering why it was there, as long as it was not fleeting and taken seriously by the one who exhibited it. He did not pull rank when he was at a loss, and he admitted being at a loss with no hesitation. Of course, these are qualities that we all have and that none of us has enough of. He stood out. The most impressive thing I found about him when we first met was his genuine interest in science and lack of power-worship. I was a physics student very frustrated by my teachers. Popper was very interested in my observations and also shared some of them. That was a boon for me.

ZP: Interesting that you use the term "dialectic", given Popper's dismissal of dialectic on the grounds that it violates the law of contradiction?

JA: You use the word "dialectic" in Hegel's sense. The default sense is the Socratic: criticism by the book.

ZP: For both Popper and you, Hegel is the deterrent example of "bad" dialectic. But whatever your view on Hegel is, don't you think that Popper is unfair in accusing Hegel of "confusing dialectic and logic"? Surely Hegel would never claim that formal logic should be redefined dialectically!

JA: It surprises me that you think well of Hegel's dialectic. It is all sheer rubbish. Nevertheless, I do think Popper undermined Hegel. (He could not say a good word on a charlatan, and this is a serious mistake.) Popper said, the only valuable thing in Hegel's output is his critique of the Enlightenment movement, and this Burke did earlier and better. The fact is, Hegel did influence much valuable biological research and revolutionized historical writing. True, it is mostly reactionary and contributed positively to the rise of militarism and thus to the disasters of the 20th century. Nevertheless, he did revived historiography.

ZP: I meant that Hegel never claimed that the law of contradiction should not be valid in formal logic (as Popper accuses him). Hegel's Logic is not logic but metaphysics. I can understand that Popper, you and many others dismiss Hegel's metaphysics. But Popper makes the mistake of transposing metaphysical terms (of Logic) to logic and the philosophy of science. For Hegel, the dialectical logic reflects the ontological structure of reality. But he recognizes the authority of logic in the sciences – in the sphere of understanding (*Verstand*). In science, contradictions are detected errors stimulating the further growth of knowledge.

JA: Making error is regrettable, though unavoidable. Detecting error is the motor of intellectual progress. This we agree about and this should do. I do not know of any modern philosopher who said it except Einstein and perhaps also Russell, and much later Bunge, Wisdom, Watkins and Gellner; but it is the cornerstone of Popper's philosophy and no one did it half as well as he did in his *Logik der Forschung*. There is nothing like this in Hegel or even in Marx or Engels.

ZP: OK, let us then define "good" dialectic as Socratic critical debate in a detachment from belief (*dogma*). These characteristics also define rationality. Yet, not every critical debate is rational. Can we debate metaphysics critically? How can we then determine which critical debate is rational?

JA: Your definition of dialectic and of rationality is agreeable to me, provided we remember that they are an ideal only approximated at rare moments. Even in science, the stronghold of rationality, much is fashion and authority and the love of mysticism.

ZP: But here exactly is Popper's point! Let scientists be mystics, as long as they can formulate a coherent, interesting and empirically testable hypothesis. The rationality of science lies, for Popper, in the methods of the refutation-aimed testing.

JA: There are no conflicting opinions here: Popper – and Russell before him – preferred not to argue about mysticism and encourage the mystics to talk clearly. They do not. You ask, how can we determine which critical debate is rational. Who says we can? The right question is not yours (how can we determine which critical debate is rational) but this: which critical debate is rational? Or, what makes some critical debate rational? Popper said, all critical debate is rational; not so, in my opinion. My rubberstamp refutation is scholasticism. The Roman Catholic Church just as the Socialist quasi-Church advocated the latest defense of the True Doctrine, but without insisting on it. When the latest defense is refuted, they accept the refutation and advocate the next defense. This is the dogmatic use of dialectic. Popper and Quine both viewed Carnap rational because he always accepted criticism. To me he seems a caricature of the rational, a kind of buffoon. He repeatedly offered a new variant of Wittgenstein's anti-philosophy. Wittgenstein seems to me even worse, yet he was a powerful critic and his criticism, at least of Russell when they cooperated, was admirably rational, his narrow-minded dogmatism notwithstanding.

ZP: In which aspect was Wittgenstein dogmatic?

JA: Wittgenstein said there are no philosophical problems and no proper philosophical assertion. Yet he considered his philosophy – including the

morality and politics that he absorbed from Tolstoy – "unassailable and definitive": not open to debate!

ZP: It seems we agree that criticism is the heart of rationality, that scientific rationality has an additional requirement for theories to have empirically falsifiable predictions, and that the growth of (any) knowledge is facilitated by new problems and hypotheses that are pressing and engaging (daring and risky). Further, critical debate is rational if it genuinely aims at the revision (even at the cost of a refutation) of the proposal.

JA: You say, rightly, since rationality is the characteristic not only of science, then criticizability is not a sufficient characteristic of science. What is, then? The answer seems to me to be criticizable explanation. This too may be too wide. Possibly an explanation is scientific if it invites criticism by relatively new experiments. It is hard for me to adjudicate.

ZP: You have often discussed the relation between science and metaphysics; Popper did not follow the positivist judgement that metaphysics is meaningless but you have always given metaphysics more importance than Popper did. Popper later allowed for a positive role of metaphysics in stimulating science. If I understand you well, you say – unlike Popper – that the interaction between physics and metaphysics works both ways. Metaphysics provides a certain context in which science (all knowledge) operates. Our criticism, then, not only moves knowledge forward but also alters the metaphysical framework.

JA: The interaction between science and metaphysics described here is optimal; most people and even most cultures are not science oriented at all and even the best scientific societies are steeped with superstitions and with positivism. My description of metaphysics and science is idealized: it is the logic of the researches of the best researchers in town. Popper said in his *Logik der Forschung* that he was concerned only with a part of their researches, not with heuristic, for example. My effort is to complement Popper's work by discussing a part of heuristic. It is what I learned from Einstein (who fully approved of Popper's theory and even found it trivial and unsatisfactory due to the omission discussion of the need that research has for some metaphysics). Late in life Popper was drawn into heuristic and into metaphysics, but not sufficiently and with some inconsistencies that are easy to remove.

ZP: Should critical rationalism thus not view metaphysics as a servant for science?

JA: Yes, AND *vice versa*. The idea of cooperation then is better than that of service. Metaphysics is thus not the queen of science and not an outcast. Still, science is in a stronger position – metaphysics without science stagnates but science without metaphysics is not quite blind, as some claim.

ZP: Popper says that metaphysical research programmes contain a possible framework for testable scientific theories. He described Democritus' atomism as metaphysics that was useful for science (in contrast to Hegelian dialectics). He also showed that later atomism became testable and he saw that as the optimal development. However, there is good and there is bad metaphysics, depending on what framework it provides for science. Within the good metaphysics he favoured rationalism, indeterminism, realism and later Darwinism. How do you exactly enrich this picture?

JA: I agree what you say about Popper although with a proviso: in Popper's first book he preferred to overlook this as there he discussed only given theories and their testability. He used atomism as a refutation of the philosophy of Wittgenstein and the "Vienna Circle", explicitly refusing to include in his methodology.

You ask a question that I wish to word thus: how do we render an untestable theory testable? Both in Plato and in Galileo there is a procedure: you ask more detailed questions; the more detailed a theory is the more we can hope to find a way to test it. I differ from Popper only in MINOR ways. There are a few differences that you have not mentioned. As to metaphysics and science, I do discuss the interaction between scientific theories and the metaphysical theories that they abide by. I also say that researchers often face a discrepancy between a scientific and a metaphysical theory and they try to reduce it. Try, not necessarily succeed. Popper says, a satisfactory (scientific) theory is a testable explanation. I say, it is more satisfactory when it integrates with other scientific theories within a metaphysical system. I say, the question what is a metaphysical theory depends on a metaphysics. After all, the problem of demarcation between science and metaphysics is metaphysical, not scientific.

ZP: Why is the problem of demarcation metaphysical?

JA: Obviously: the theory "a theory is scientific iff it is empirically testable" is not empirically testable. Popper sticks to deductive explanation that is testable, which is a positivist idea. Consider this historically. The two central metaphysical systems are realism and idealism. The conflict was also methodological: between the view that science rests on *a priori* truths and the view it rests on experience. When we give up certainty, the dispute remains: does science begin with *a priori* or *a posteriori* assumptions? Why not both? Because they may be in conflict with each other. Indeed, Kant employed both but first he divided science between the *a priori* and the *a posteriori* given, so as to insure the absence of conflict. Popper required this very conflict.

ZP: Let us look at realism now. We cannot prove that realism is true but we accept it because it provides the best framework for the development of science. Yet, Popper says, we should not accept realism dogmatically. If there is ever a more science-friendly metaphysical theory we should be ready to consider it as a matter of principle.

JA: You say Popper accepted realism. If you mean, he deemed realism true, then yes, of course he did. He considered every scientific hypothesis realist: he considered science cosmology. Idealism is a crazy theory that wise people like Berkeley advocated with ingenious arguments. Popper noted that as a Christian believer Berkeley could not believe in idealism, and believe that Jesus Christ went up to Heaven bodily. It was for him a methodological necessity, not an ontology. It is true of all versions of idealism. They are all obsolete. Reichenbach said, when we look at an object twice we have no right to assert that it existed between the two events. He calls this existence inter-phenomenon. I think he was a fool as he did not think the object disappears and re-appears. The assertion that the object does not disappear between the two observations is realist.

ZP: Yes, science is about real nature so let us declare true, but not dogmatically, immune from a revision (though that is very unlikely).

JA: You say, a revision of realism is very unlikely, as a matter of principle, and this makes me laugh. What is unlikely? What principle are you talking about? The only principle I know is the decision to be willing to consider alteration due to some criticism. I do not see that we need or even

can decide these things, but if so, my decision is to remain open to criticism as much as is within my (intellectual) powers.

ZP: Popper explicitly proposes "to accept realism as the only sensible hypothesis – as a conjecture to which no sensible alternative has ever been offered". It is a conjecture open to criticism and it is our decision to accept it as true (because it gives point to our search for truth). But let us turn to the decision to accept rationalism. It must be done prior to any argument. Hence, as Popper says, rationalism is not self-contained.

JA: Suppose irrationalism is an attitude, the attitude of distrust in reason. Then, to cite *The Open Society and Its Enemies*, "We could then say that rationalism is an attitude of readiness to listen to critical arguments and to learn from experience. It is fundamentally an attitude of admitting that '*I may be wrong and you may be right, and by an effort, we may get nearer to the truth.*'" This is not to decide that attitudes are or are not open to criticism but to draw attention to the fact that criticizing attitudes differs from criticizing theories, (and Bartley overlooked this fact).

ZP: Popper says that neither logical argument nor experience can establish the rationalist attitude. Therefore, we have to accept is as a matter of an "irrational faith". Some philosophers (e.g. Bartley) criticized Popper for fideism. What is your opinion?

JA: This is an odd question. Fideism is a qualified uncritical rationalism. Critical rationalism therefore cannot possibly be fideism. Bartley called Popper a fideist in a sense that is an extension of the initial sense of the word: as a qualified critical rationalism, not as qualified uncritical rationalism. Now fideism comes as a result of a criticism of uncritical rationalism. Bartley's criticism of Popper comes to show that there is no such criticism of critical rationalism. But Bartley said, as Popper said he qualified his critical rationalism, he was a fideist in the extended sense of the word. The fideist says, endorse an axiom uncritically and with no justification, and then you can demand that every later step be justified. This initial axiom is known by Descartes' term "the Archimedean point". The fideist says, the Archimedean point is not given so you have to choose one arbitrarily. Popper said, if you want to choose an Archimedean point arbitrarily, it is advisable to endorse the minimal assumption. And the minimal assumption is that criticism is beneficial. This brilliant idea Bartley

foolishly called fideism. To be more specific, the word "faith in reason" is not to be taken too seriously. Perhaps "life of reason" is more accurate.

ZP: I think that this problem can only be understood in the social context. Popper says that accepting rationalism is a moral choice. Applying criticism and rationalism in society (social theory and politics) is a defence against totalitarianism, utopianism and fanaticism. Social changes should be done by piecemeal steps that can be revised and reversed relatively painlessly.

JA: Yes, to be uncritical is irresponsible and even childish. Traditional social thinking was largely conservative: make as little change as possible, because reform – social or political – is very difficult and very unpredictable and fraught with undesired unintended consequences, especially weakening the social fabric. The exception was almost all utopian: ignore current social settings and start afresh. The last option was, ignore all social systems and do not replace them: live a-socially and take good care of yourself. Around 1900 the industrial revolution led to a new attitude and to a broad development: solve social problems. The first such move, it seems to me, came earlier: it was the cooperative movement discussed and summed up by Sam Smiles and by George Jacob Holyoake. It was developed by the Fabians and the London School of Economics and Political Science (especially Beatrice and Sydney Webb, Bernard Shaw and Bertrand Russell); their discussion between the choice of evolution or revolution made them famous. Popper made this the principle of piecemeal social engineering. This was read as his advice to undertake minor projects before major ones. This is not true. His idea was to tackle specific problems and not the ills of society as such, as the later leads to an overhaul of society which is utopian engineering, and he wanted social engineers to be aware of the possibility of big errors that make the cure worse than the illness. In my opinion all this is right but outdated by the need to save humanity from extinction. This need demands rethinking Popper's ideas.

ZP: You regard Popper's concept of corroboration as too rigid. Can you explain why?

JA: The idea that criticism is valuable is just terrific, it is Socratic. Nevertheless, my greatest dissent from Popper is my claim that we can admire different parts or aspects of science for different reasons.

ZP: What other different aspect of science do you have in mind? Popper – as I understand him – allows all sorts of practices in science. Valuable are those which are bold (improbable), provocative, offer new ideas, have high empirical and informative content. Criticism consists (in science) ONLY in ruthless falsification attempts. Popper's originality lies for me in his radicalism. No other method of testing is allowed or needed. This rules out any form of justification – for instance a weak appeal to criticism (after which the theory acquires some evidential support that makes it "stronger" than it was before the testing). Do you consider this too strict?

JA: Before answering your question, let me add a few points to what you say of Popper, which is very nice indeed. First, William Whewell, who was a justificationist, of course, outlined the hypothetico-deductive method. Popper's view of it is in agreement with it, even though Whewell was not as clear and emphatic. (Quine, it seems, did not like Popper's emphasis although he had no criticism of his methodology). In Whewell's view refuted theories play no positive role in science. Acknowledging that Newton's theory of light had been rejected (in the year 1818), he declared it worthless. Popper could not possibly declare worthless Newton's theory of gravity, and so he had a better challenge and showed better results. Popper's rejection of all justification and all justification surrogates is his greatness. So he had a problem: what is the value of corroboration? He said in the first place it is the information that a search was a failure. This is obviously true. Also, the corroborated hypothesis has an increased explanatory power and thus fewer competitors for the time being. Popper wanted more, and there he faltered; demanding more he made a concession to the inductivists: it is, to use his words, his admission that the view includes a whiff of inductivism. He took back this concession at once. And so, to answer your question, Popper's additional inductivist requirement is not too strict but too lax. He should not have made it and I am glad he withdrew it.

ZP: Do you mean the whiff of inductivism that you found in Popper's "third requirement for the growth of knowledge"? (Theories should not only be refutable but they should pass some tests; this makes it possible to get nearer to the truth). I thought Popper said that a corroborated theory simply stays in the game – it advanced some interesting problems and new

solutions and if it is not refuted (it is corroborated) it is conjecturally true. Nothing else is gained by corroboration.

JA: What you say of Popper on corroboration is true. It is his third requirement that is puzzling, the requirement that a theory be corroborated before it be refuted. This is new and redundant at best, perhaps also refuted by cases like that of the Bohr-Kramers-Slater 1924 theory that is very important and that was refuted upon its very first test. In any case, the question why some theories are corroborated is scientific, not methodological or epistemological. Thus, according to Einstein, Newton's theory of gravity is highly corroborated because the sun is so much heavier than the planets and the size of the solar system is not too small. Popper later said, the "whiff of inductivism" in his theory is the hope that the history of science shows an increased verisimilitude. This is more interesting.

ZP: I know that you demand some positive role of corroboration alongside with Watkins, Worrall and Zahar, especially in practice. But rationally, good reasons are not available or necessary. Do we need more?

JA: Yes, we do: the practice that the inductivists discuss is psychological as you suggest. The practice that we live is social, including the need to license practices. My study of technology rests on the fact – a scientific fact – that the laws of civilized countries require the corroboration of certain hypotheses before launching them in the market is permitted. These hypotheses are guarantees that the promised performances of the technologies in question are valid and that there are no harmful unintended consequences. The law often specifies what consequences are empirically excluded and how severe the tests should be to exclude the undesired effects. In other words, some justifications of technologies are required by law. They differ from induction, from the philosophical justifications that are impossible: they are valid only within certain restrictions. When these restrictions are violated, these technologies may fail. The failures of the technologies in question are then absolved as act of God (*Force majeure*).

ZP: Yes, this is the notorious practical problem of induction. For Popper, there is no guarantee of success. He says: "in spite of the rationality of choosing the best tested theory as a basis of action, this choice in NOT

rational in the sense that it is based upon good reasons for expecting that it will in practice be a successful choice".

JA: The practical problem of induction is whether the technology we successfully use today is promising in the long run. The theories used in such technologies are insufficient – many techniques have no theoretical background – and usually false – we still use Galileo's theory and Pasteur's theory, both very well refuted. And indeed, there is no guarantee that the sun will rise tomorrow. Science was invited to prove this and it proves the opposite instead: the sun is a nuclear furnace so it can explode any day. In my view although there is no guarantee that the sun will rise tomorrow it is rational to assume that it will and not that it will not. For, if the sun will not rise tomorrow, then it does not matter what we do today but if it will it does.

ZP: However, in Popper's intention corroborated theories do not promise anything (in the sense of reliability) from the rational point of view; they may cause some subjective psychological reassurance, that's all. On a smaller scale it is the same as the dilemma whether to jump from the window or take the lift.

JA: We take it for granted that jumping through the window is disastrous, and that taking the elevator is not sufficiently safe either. The discussion of this case in the literature is the lie that the elevator is safe.

ZP: Back to our question of why corroboration matters. You know I am a supporter of Miller's version of critical rationalism and thus I would say that corroboration does not matter. The question rather stands: "why NOT accept a corroborated theory?" Questions about the reasons FOR accepting a corroborated theory beg a justificationist answer.

JA: What is the good of corroboration? Firstly, it is the refutation of a competitor; when the competitor was the consensus, the public-relations spokespeople of science did not like to notice refutations and so they stressed corroboration. This holds for the vulgar. Secondly, technology needs corroboration, and by law, and in order to show responsibility. Look at history: the theory of rational belief rests on the great discovery (Bacon, Galileo) that observations are theory-laden. Bacon recommended to stop believing in any theory and to rely on naïve realism (commonsense). Galileo refuted this idea

(you might very well see the moon jumping from rooftop to rooftop when strolling down the moon-lit street). He recommended using mathematics, a recommendation that is a cornerstone of Kant's theory of science. Now all this is past history. There is no reason to suppose that we can be free of erroneous theories in research situations or in any other situation. Rather, we can try different theories and see which functions better under which conditions. In sum, we should not worry about acceptance of theories. Rather we should always seek explanations and, when possible, competing ones.

ZP: Let me close the discussion on corroboration with this question: your view on corroboration entails the possibility of empirical support of a theory. In that case it entails induction. Do you allow induction in methodology?

JA: Induction as one way to generate testable hypotheses is fine. Induction as justification is silly. Corroboration as a crucial experiment is a refutation. Corroboration as the increase of explanatory power is fine. What other option is there?

ZP: The option is that a proposed theory deserving our research interest already entails new information, new explanations and predictions. It is either refuted or retained (corroborated); the increase of explanatory power is not due to corroboration. If you claim it is, do you, then, allow induction in methodology?

JA: Yes, Popper did so already (in his third requirement).

ZP: But you said that he withdrew it! So you propose an inductive-critical model of knowledge?

JA: No. Heavens forbid! It does not deviate from Popper's hypothetico-deductive model, since that model does not apply to heuristic. It was never meant to apply there: the model does not say how a hypothesis is generated. And so it may be inductive although this is unlikely.

ZP: Lakatos argued that it is impossible to falsify an isolated theory since theories are interlinked in a research program. He draws on the Duhem-Quine thesis. Do you think we can test an isolated theory?

JA: The Duhem-Quine thesis says that verification is not possible, not even by crucial experiments. The argument for it is the observation of Duhem (1906, 1954): crucial tests are no verifications as they employ

unverified working hypotheses. What this has to do with Popper I have no idea. Nevertheless, I expanded on it in my *Popper and his Popular Critics*; the popular critics all thought the Duhem-Quine problem/thesis/argument was a fatal objection to Popper's characterization of scientific character as falsifiability. Roughly, it relates to a brief rider of Duhem to his discussion of his thesis: a crucial experiment is no proof; at most it is a disproof. But, he added as a rider, even that is not clear-cut since we can always rescue the refuted hypothesis from refutation by blaming the working hypothesis for it – logic always allows for this move.

ZP: Irrespective of Duhem, do you think it is possible to apply a crucial falsifying test to an isolated theory?

JA: Of course when we use the same instrument it is harder to say that the instrument mislead against one theory without also saying that it mislead against the other. And if it misleads against both yet the experiment goes one way, it is a challenge to examine the situation afresh. When a situation looks challenging, it seems to me a Good Thing and to Elie Zahar a Bad Thing. Duhem said all tests involve working hypotheses about the test's environment. This does not change in cases of crucial tests. This, incidentally, is why before testing a hypothesis it is wise to test the working hypothesis involved: a part of it is the calibration of the test's instruments. Every experimenter knows this. My trouble then is in the question, what is it about Duhem's argument that some people say it refutes Popper? Also, why do these people do not say that Popper's *Logik der Forschung* discusses this argument at some length? I have a conjecture: people may think that it is hard to persuade people, and so Popper says, better dissuade them. And then a shock: it is hard to dissuade people, too. Popper said so repeatedly and this is why the Vienna Circle people disliked his views. They wanted to impose the scientific worldview and for this they tried hard to fight dogmatism. Popper said, this cannot be done. Dogmatism is unwise, but it is logically permissible.

ZP: Popper's concept of verisimilitude was proved wrong (Tichý, Miller) because false theories cannot be compared for verisimilitude. Many philosophers of science thought that this was the end of Popper's non-inductive account of the progress in science (Newton-Smith talks about a "full blown storm" of inductivism). What is your opinion on this matter?

JA: The idea of verisimilitude is Einstein's, not Popper's. Popper offered a MEASURE of verisimilitude. He never explained what it is good for. It backfired because he did not examine it carefully as he usually did with his innovations. It was, incidentally, a booboo. I offer a minimal definition in order to save verisimilitude:

Popper has offered his late view of scientific progress (L) in addition to his early view (E),

(E) Progress is empirical success;

(L) Progress is verisimilitude increase;

I have offered a minimal definition of verisimilitude increase: a theory is more verisimilar than its predecessor if and only if all crucial evidence concerning the two goes its way.

ZP: You appreciate Popper's emphasis on criticism and say that he elevated criticism "from hors d'oeuvre to the main dish". I tend to see Popper as elevating criticism to the only (rational) dish; I speak of testing. His new conception of reason – *ratio negativa* – allows NO justification. Don't you betray this legacy when you claim that "when an attempt at empirical criticism misfires the result is positive evidence"? Why not just say that the result is the absence of empirical refutation?

JA: It surprises me very much that you ask why not call it "the absence of empirical refutation", as I do that a few times. Yet the received name is "positive evidence" and there is never a good reason against a received name.

ZP: The name evokes the justificationist interpretation. Popper would say (as he does in the case of "good reasons") – call it positive evidence if you must but it does not involve any support of the theory (any increase of probability, credibility, reliability, certainty etc.).

JA: The claim that seems to bother you is that positive evidence is easy to find if that is what you want. I do not see why. Most unsophisticated people seek positive evidence, and they usually find it. Of course, they seek it in the hope that it makes them feel secure. This feeling is often illusive. At times the illusion that it gives them is dangerous. All this is well known and it is not clear to me what troubles you about it. What you incite me to

say is that the positive evidence that comprises failed criticism, corroboration, is different. And it is: it does not matter whether corroboration strengthens our belief or not; it matters that it is valuable information – even though refutation is more valuable. Consider the corroborations that the science textbooks cite. It is often significant information and it is often enlightening. Popper wanted to know why and he offered a theory of it. We can and should put it to critical assessment. That is all that there is to it. Corroboration takes place also in everyday life. Most philosophers of science say it is probability and Popper has refuted this.

ZP: Well, justificationism troubled Popper. From the point of logic he dismissed positive evidence as flawed (induction). But from the psychological point of view he knew that we (sophisticated or not) have inborn dogmatic tendencies – we want our expectation to be fulfilled. Therefore, we find "positive evidence" reassuring and are after it. This tendency is, for Popper, not rational, does not encourage bold and risky conjectures and is hostile to criticism. You yourself defend the value of positive evidence and claim it is needed.

JA: This is not the way it seems to me. No positive evidence supports any theory. Nevertheless, there are different kinds of positive evidence that require different treatments. First, the positive evidence that is inductive (Hempel calls it instantiation) Popper rightly dismissed; the positive evidence that is the result of tests and that science textbooks cite is valuable, though not as a support and less than as refutations. The question, what is the value of positive evidence, is where my view differs from Popper's, not the previous points.

ZP: Let me elaborate on Popper's attitude to dogmatism. Dogmatic tendencies are strongly ingrained in our nature; they are in our genes – or, as Popper says – biologically *a priori*. Popper urgently felt the danger and the cunning of dogmatism – we could continue to practice dogmatism while proclaiming criticism (dogmatism can sneak in through the back door, as you say): hence his radicalism in identifying the rational approach with criticism. In other words, it is criticism that needs the boost.

JA: The disposition for dogmatism is ubiquitous, but once we learn how to doubt and to criticize, it becomes an unshakeable habit. Popper said – although hardly wrote – that the risk of dogmatism is permanent and we

must keep vigil. This seems to me exaggerated. His criticism of Neurath – for his permission to ignore refutations with no qualification – seems to me exaggerated. It is nice to have Popper's qualification, but that qualification seems to me not necessary and at times even excessive. For example, it is a censure of Faraday for his having refused to accept the verdict of experience when he failed to find what we call today the Kerr effect. Dogmatism is stagnation and boredom. Those who like it are welcome to it. The healthy response is to prefer innovations and the Popper who teaches us that dogma and novelty conflict does better than the Popper who warns us against dogmatism. The refusal to be dogmatic seems to me – not to him – to be possible to take for granted in some contexts.

ZP: My point was that Popper proposed a strictly negative methodology because of his *horror dogmatis*. He saw this as the only way to keep dogmatism behind the door.

JA: My appreciation is to Popper's avoidance of dogmatism without despair in reason. Yes, his view of dogmatism is reasonable; his putting pressure to prevent it is not. Popper's negative philosophy appeals to me very much. His insistence on it is superfluous. It even betrays a measure of distrust that is unbecoming.

ZP: Well, putting pressure to prevent dogmatism is the only effective strategy. You and Ian Jarvie provocatively argue for the rationality of dogmatism. You base this argument on Popper's repeated claim that dogmatism is the necessary initial stage in assessing a hypothesis. The hypothesis must show its strength and prove its mettle before it is critically attacked. I consider this dangerous since dogmatism could then become uncontrollable.

JA: The paper of Jarvie and myself works on the assumption that there are degrees of rationality, so that the rationality of dogmatism is limited, but that we are all dogmatic to this or that extent, and usually unknowingly. The great switch of Popper in the study of rationality, including scientific rationality, of course, is the shift from psychologism to sociologism, from 'How do I learn/know?' to 'How do we learn/know?'. This is not to exclude psychology, of course, but to limit it to social settings. The discussion that Bartley developed concerns the faith in reason of an individual (of first person singular). It was psychological and its sociological aspect is lost.

ZP: Yes, but back to dogmatism. If we allow some degree of dogmatism into methodology we cannot control the extent of dogmatism – it can spread indefinitely, insisting that the theory has not yet proved its mettle. How are we to determine when the dogmatic protection of the theory from criticism should cease? As a radical Popperian I think that in order to keep dogmatism behind the door we must apply radical measures and separate rationality from dogmatism. Dogmatism is inherent in all forms of justificationism; strategies seeking the confirmation of a theory are not only logically flawed and thus irrational, but tend to immunize theories against criticism and thus suppress the growth of knowledge.

JA: Yes, I do not fear dogmatism, even though I agree with you about dogmatism being the default option. Most people I know are dogmatists, including philosophers of course and including scientists to my surprise. I consider the critical attitude incurable, or else dogmatists would have won the day long ago. To follow Popper's theory of tradition, it is possible to destroy the critical approach (as the middle ages illustrate), but for this it is necessary to fight criticism on a large scale and systematically, something like what was experienced in the leading early-twentieth-century non-democracies. They failed because they could not bar the import of criticism, no matter how little. See how isolated Popper was in Vienna, and how powerful Schlick was, yet miraculously Schlick is dead and Popper is not. Think of the amount of radio/TV religious propaganda in the USA and look at its yield. It makes one optimistic.

ZP: Well then, you are more optimistic than I am about the natural willingness of the human species to practice criticism. I hope you are right.

Finally, what do you consider the most valuable and original in Popper's contribution to philosophy? I would mention criticism as the basis of his original negative conception of reason; the encouragement not to be intimidated by any authority; the appeal to appreciate the positive value of erring and to welcome disagreement. The weak point is tying rationality too tightly to logic.

JA: Popper's weak points are his Protestant work ethics (it is misanthropic) and his Kantianism. His anti-metaphysics is Kantian and is to be dismissed as a part of Kant's dismissed justificationism. Kant's categorical imperative that won admiration is absolute and so not applicable. Popper's

vision of science is contrary to Kant. His great achievements are his view of science as a Socratic dialogue with the result of making room for the Einsteinian revolution, and his linking democracy with science thus presenting it as limited but able to progress.

ZP: Joseph, thank you for this interesting conversation.

JA: Thank you, Zuzana, for your interest in my opinions, for your interrogations and for bringing my responses to the public. I hope readers find the discussion interesting and helpful, since it dispels the popular misconceptions of Popper's ideas and enables the interested to pursue the problems that Popper's philosophy raises.

Organon F 24 (4) 2017: 546-558

Axel Gelfert: *How to Do Science with Models: A Philosophical Primer*
Springer, 2016, 135 pages[1]

"Models (…) are all around us, whether in the natural or social sciences, and any attempt to understand how science works had better account for, and make sense of, this basic fact about scientific practice" (p. v).

Over the past twenty years scientific modeling has become a booming topic in philosophy of science. Axel Gelfert's book *How to Do Science with Models: A Philosophical Primer* is an up-to-date introduction to a number of hot topics as well as an original contribution to the literature. First two chapters function as an overview of the debates on the nature of models and about the way in which models represent their target systems. Anyone interested in general philosophical debates on modeling will profit from reading it as it serves as much needed coherent introduction. The remaining three chapters are different in that they offer a detailed analysis of a number of examples of actual scientific practice (chapter 3), an exciting analysis of a neglected topic, exploratory models (chapter 4), and an interesting take on the issue of a material and a theoretical dimension of models (chapter 5).

Conceptually, the book can thus be divided into two segments or approaches, one that addresses more general philosophical issues that have been vigorously debated in the literature in the recent years, the other rather specific with an eye on particular detailed examples taken from (mostly but in no way exclusively) physics. As has been common in recent years, the book is written in a style that values and pays attention to actual scientific practice. This pragmatic turn allows Gelfert to present the reader with a vast number of strategies that appear in the scientists' modeling practices. All of this makes Gelfert's book a valuable contribution to otherwise vast and disparate literature on scientific modeling.

[1] ✉ Martin Zach
Department of Philosophy and Religious Studies
Faculty of Arts, Charles University
Nám. Jana Palacha 2, 116 38 Prague, Czech Republic
e-mail: m_zach@seznam.cz

In the remainder of this review I will summarize the content of the chapters, focusing on both novel and interesting insights provided by Gelfert and on certain problems.

In the first chapter, Gelfert poses the question of what scientific models are. We get a nice summary of all the main contending positions which offers a great introduction for anyone new to the subject. First of all, there is a number of ways to classify different kinds of models. Thus, one can attempt to provide a typology of models (e.g. scale models, analogue models, mathematical models, theoretical models), or focus on a functional characteristics of models, for instance, on the representational aspects. Gelfert devotes some space to reviewing the models-as-analogies account of Mary Hesse, and to the syntactic and semantic view of models. He then goes on to elaborate on the fiction view of models. With a reference to Thomson-Jones, Gelfert notes that, given idealizations and abstractions, it is often the case that model systems are not instantiated in the real world, hence the "models-as-missing-systems" account. However, the practice of speaking about these kinds of model systems, as if they were instantiated, has been referred to as the face value practice. The question is, then, what is it that we speak of when we speak of a model system?

Some accounts of models take these model systems as akin to characters from novels. Against the view that model systems could be regarded as "imagined physical systems, i.e. as hypothetical entities that, as a matter of fact, do not exist spatio-temporally but are nevertheless not purely mathematical or structural in that they would be physical things if they were real" (Frigg 2010, 253), Gelfert points to models in sociology and cognitive psychology that would not necessarily be 'physical if real'. Indeed, this, for me, brings an interesting question as to what these models would be if they were real.

Although Gelfert does a good job in summarizing the debates and providing some of his own insights, he also errs on at least one occasion. When discussing the so called direct and indirect fiction view of models which is based on Kendall Walton's make-believe approach and according to which model descriptions are taken to be prescriptions to imaginings, Gelfert incorrectly places Roman Frigg into the direct fictionalist camp. Gelfert claims that "recently, more thoroughgoing *direct* views of models as fictions have been put forward, including by Roman Frigg and Adam Toon" (p. 17). But this is mistaken. As Toon says, "Frigg also draws on Walton's theory of fiction, but he advocates an indirect view of theoretical modeling (…)" (Toon 2010, 308). Furthermore, Frigg himself criticizes the direct fiction view while defending the indirect one (see Frigg & Nguyen 2016). The chapter closes with the 'challenge from scientific practice': by seeing how models

are actually used by scientists, one had to either modify the semantic view or leave such a view behind and accept a 'radical heterogeneity of scientific practice'. As I noted above, it is indeed this heterogeneity that Gelfert makes vivid in his book.

Second chapter deals with the problem of scientific representation and other functions of models. Models 'stand in for' their target systems. However, by virtues of what does a model represent its target? In accord with the literature on scientific representation, Gelfert embraces the distinction between 'informational' and 'pragmatic' accounts of representation. The former concerns an objective relation between the model and the world while the latter includes the intentions of agents and the various specific uses for which models are designed.

As Gelfert notes, any adequate account of scientific representation has to be able to account for a number of things, e.g. the fact that models serve as surrogate systems and that they often misrepresent their targets. Goodman's general views on representation are discussed, followed by a review of specific accounts of scientific representation, such as Hughes' DDI account, Suárez's inferentialist account or Contessa's interpretational account. Gelfert then points to a number of other functions that have been discussed, such as the fact that false models and incomplete models are actually epistemically valuable (Wimsatt), or that there might be non-representational uses of models (Grüne-Yanoff).

In the third chapter Gelfert presents several case studies to illustrate the strategies of model-building. Here, Gelfert argues for a middle ground between unitarism and pluralism about model-based science. He recognizes that there are multiple strategies but he also notes that some of them are actually recurring. He discusses three general types of scientific models that, nevertheless, can overlap: phenomenological models, causal-microscopic models, and target-directed models. Each type of model is suited to different purposes and to answering different kinds of questions, and each has its advantages and disadvantages.

To illustrate how these strategies are put to work in actual scientific practice, Gelfert devotes a large chunk of the chapter to providing detailed examples of accounting for the phenomenon of superconductivity. He discusses the phenomenological approach of Ginzburg-Laudau's model, the BCS microscopic model, Hubbard many-body model, and then Lotka-Volterra model for modeling population dynamics. Although the discussion gets rather technical at certain points, and thus it might prove challenging to follow the argument in depth for someone without advanced knowledge of physics, it nevertheless illustrates the main point rather well, i.e. that different modeling strategies are at play in scientific practice.

In the context of strategies of model-building, it has become customary to reference Richard Levins' work (trade-offs between precision, generality, and

realism) and Gelfert is no exception here. The existence of trade-offs has been thought to be a distinctive feature of biological models, however, as Gelfert argues, many models in physics and chemistry exhibit the same trade-offs as well. In concluding remarks to this chapter, Gelfert summarizes the point with an example of climate models (p. 68):

> In other words, rather than aiming for a model that reflects every available detail of the target system, it may be preferable to have a model that makes adequate predictions primarily of those features that matter to us – say, changes in rainfall patterns in agriculturally productive parts of the world – even if it misrepresents other parts of the target system as a whole.

In chapter 4, we are presented with the notion of exploratory models. Gelfert begins by noting the importance of scientific understanding in the form of model-based understanding. This sort of understanding has an important tacit dimension: a 'feeling for' the model and the behavior of its target system which is acquired by simulation or manipulation (physical and/or symbolic). A central notion of this chapter is the notion of exploration, though. Exploration can be either 'specific' in the sense that it "converges upon a specific question, fact, detail, or 'missing link'" (p. 75), or 'diversive' which is not directed at a specific object or a question. Experimenting as well as modeling concerns both senses of exploration which Gelfert well documents on a number of examples.

These exploratory tasks can be aimed at forming and stabilizing certain conceptual frameworks, and in some cases, a tentative proposal of an operational definition is a prerequisite to an intelligible experiment. Based on the last point, Gelfert claims that concepts may play an exploratory role in a similar way to experiments. Although suggestive, it seems to me that this claim would have benefited from further arguments. Be as it may, Gelfert's main interest lies with exploratory models. He takes minimal models (e.g. in ecology, physics, and social sciences) to be instances of this category, the exploratory models. Minimal models, as he sees them, are intended as tools for investigating certain model systems which do not refer to any particular real world systems, nor make precise quantitative predictions. One might object that Gelfert should have at least mentioned the fact that the term 'minimal model' has been used in rather different senses in the literature, however, his main concern is not with the variety of the meanings of the term, but rather with the fact that at least some of the usage illuminates well the notion of exploration.

He then goes on to further illustrate the importance of explorative models by showing four different functions these models may have: they serve as starting

points of research, as proof-of-principle demonstrations, they generate potential explanations, and they explore the suitability of the target. Gelfert thus highlights the exploratory function of scientific models and puts the notion of exploratory models on a par with other important kinds of models such as predictive and explanatory models.

In chapter five, Gelfert first devotes space to differing accounts of scientific models. Models-as-mediators account stresses certain autonomy of models from both theory and data and highlights the fact that models are often constructed by using various tools. Models have also been construed as epistemic tools, as concrete artefacts, built by specific representational means and constrained by their design (i.e. a given design allows answering certain questions and serving certain purposes but not others). Gelfert wants to go a step beyond the 'models as tools' which he sees as too passive – he wants to stress their active role.

He then focuses on yet another important aspect of models when he says that "for a model to be successful, more is required than that it stand in the right sort of objective relationship to its target system" (p. 117). He adds that "a successful model should enable such learning, by making relevant information about its target accessible to us – not only in principle, but in a sufficiently salient way, such that a reasonably skilled user would be able to draw relevant inferences about the target system from interacting with the model via the representational means it employs" (p. 117). In order to capture this relation between models, model users, and their targets, Gelfert suggests distinguishing between degrees of *immediacy*, which concerns "the phenomenology of our interaction with the representational means deployed by a model" (p. 118), and degrees of *directness*, which pertains to the relation between the model and the target.

Following up on the presented distinction Gelfert draws on yet another distinction originally due to Don Ihde (taken from the philosophy of technology), one between embodiment relations and hermeneutic relations. Embodiment relations concern technologies that interact with our perception and body, whereas hermeneutic relations concern the need of interpretation. Examples include: glasses, telescope, or car-parking (embodiment relations); and measuring or computing apparatus, or graphical chart belong properly to the category of hermeneutic relations. The difference is then further clarified in the following way: "Both the printed map and the handheld telescope are visual technologies of sorts; but whereas in the case of a telescope, we can 'become one with' and, through skilled embodied use – as an extended self, we might say – look through it at the world, in the case of the map the representational medium itself occupies the focal point of our attention:

when we read a map, we are looking at the map, not through it at the world" (p. 122).

We have learned that there is an incredible heterogeneity of model-elements, and this heterogeneity "often entails that some parts of a model may be continuous with our ordinary sensory modalities, whereas others require significant interpretation" (p. 124). Gelfert applies this distinction to the context of models, and, furthermore, he highlights that both kinds of relations are often at play at the same time:

> An engineer designing a new type of aeroplane might begin by constructing a model that has the appearance of the full-scale aircraft, including its geometrical proportions, only to find that not all relevant properties (such as drag, weight, friction etc.) scale proportionately with size; in such cases, one would need to suspend immersive engagement with what looked to be a good stand-in for the target system and 'read' the model in a more detached way: for example, by taking measurements, making appropriate modifications (e.g. adjusting the relative wing size), or adding further elements (e.g. additional background assumptions) to it. Working with models often requires such 'switching' between embodied and hermeneutic modes of interaction. This leads to the second modification of my general claim: not only do scientific models support different types of user-model-target relations, but they often *enable* their users to switch back and forth between them. (p. 124).

To further illustrate this point Gelfert presents us with two more examples. The first one is the Phillips machine which is a machine built from water tanks, levers and tubes, and which serves as an analogue model of macro-economy. The materiality of the Phillips machine "is key to how the machine models economic processes" (p. 125), but interpretation is required as well – one needs to have a good grasp of the economical concepts. The second example is that of modeling proteins. Before the dawn of advanced computer technology scientists were constructing material models of proteins to find out about their structure. Figuring out the three dimensional structure of proteins is difficult because it cannot be straightforwardly predicted from a sequence of amino acids because, in Gelfert's words, "a sequence of amino acids will 'fold' into the most energy-efficient three-dimensional structure, yet determining this structure involves running numerically demanding simulations which, in turn, requires the extensive use of computer technology" (p. 126). Thus, we see an important 'hermeneutic' element involved in the process. Nowadays, however, sophisticated programs have been developed that allow their users to manipulate 'virtual' atoms, followed by rendering of the most probable

structure of a protein, all this in real time. As a result, we get more of the 'embodiment' element, a 'feeling for the molecule'. Gelfert closes by noting that "models, then, are not simply neutral tools that we use at will to represent aspects of the world; they both constrain and enable our knowledge and experience of the world around us: models are mediators, contributors, and enablers of scientific knowledge, all at the same time" (p. 127).

From the very beginning Gelfert has argued that searching for a unified account of scientific modeling is a fool's errand. Indeed, as Gelfert argues, given the various roles and uses of scientific models there will not be any such unified account, ever. What we have been given is a plethora of well documented cases of scientific modeling which show how colorful and multifarious the actual practice is. Gelfert is also to be applauded for opening new philosophical issues to work on such as the role of exploratory models. Anyone interested in the up-to-date research on the philosophy of scientific modeling is well recommended to read this book, as well as anyone interested in the scientific practice more broadly.

Martin Zach

Acknowledgments

The work on this book review was supported by the Charles University, project GA UK No. 66217. I would like to thank Sara Green for some useful hints and comments on the previous version of this review.

References

FRIGG, R. (2010): Models and Fiction. *Synthese* 172, No. 2, 251-268.

FRIGG, R. & NGUYEN, J. (2016): The Fiction View of Models Reloaded. *The Monist* 99, No. 3, 225-242.

TOON, A. (2010): The Ontology of Theoretical Modelling: Models as Make-Believe. *Synthese* 172, No. 2, 301-315.

Hugo Mercier and Dan Sperber: *The Enigma of Reason*
Harvard University Press, Cambridge (Mass.), 2017, 396 pages[1]

At one point in their book, Mercier and Sperber present their readers with what looks like a defective chair. However, the authors point out, it looks so obviously defective only till we realize that it is not a chair, but something different, namely a kneeler. And Mercier and Sperber are suggesting that an analogous obstacle has been hampering our assessment of human *reason* – we have been brooding over its apparent shortcomings, or defectiveness, because we have been misconstruing its principal function.

The term *function* of course, by itself, is potentially ambiguous, but Mercier and Sperber are using it in its well defined sense tied to the context of evolution theory; here the function of an organ or an ability of an organism is what this organ or ability has been *selected for*. From this viewpoint, reason and reasoning is usually thought about as an adaptation helping us to solve problems, to accumulate faithful knowledge of the world and to peruse it in a cooperative way.[2] And given this, we humans should be expert reasoners, making errors only when facing problems that are overly complex, or when trying to solve them under significant stress. How come, then, that as a matter of fact, we make systematic errors when solving some *prima facie* simple tasks, such as the Wason task?

There are various ways to explain such spectacular failures of human reasoning. One way is by appealing to the concept of bounded rationality (see, e.g., Morton 2010): human reason is powerful, but not almighty. Failures are to be expected; we should not measure the performances of reason by abstract standards which do not take into account human limits, such as the restricted capacity of human memory or its limited and not completely robust computational powers. But the pointed question remains as to whether this is adequate to fully explain why humans can sometimes predictably fail to solve problems arguably much simpler than those which they can solve easily.

The idea proposed by Mercier and Sperber is that in fact reason is not at heart the kind of adaptation that has usually been assumed. Rather, it has been formed

[1] ✉ Jaroslav Peregrin
Institute of Philosophy, Czech Academy of Sciences
Jilská 1, 110 00 Praha 1, Czech Republic
e-mail: peregrin@flu.cas.cz

[2] Cf., for example, the notion of *cognitive niche* of Pinker (2010).

by selection pressures different from those centred on favoring the most perfect solutions of problems:

> The main role of reasons is not to motivate or guide us in reaching conclusions but to explain and justify after the fact the conclusions we have reached. (p. 121)

How could this be? Is reason not a tool for solving problems? It is quite clear that we do use it to solve problems, and in many cases with superb effects; however, their claim is that this is not the function it has from the viewpoint of evolution, which would make sense of its seemingly unexplainable failures in some simple cases. What, then, would reason and reasoning have been selected for?

> Contrary to the commonsense picture, much experimental evidence suggests that people quite often arrive at their beliefs and decisions with little or no attention to reasons. Reasons are used primarily not to guide oneself but to justify oneself in the eyes of others, and to evaluate the justifications of others (often critically). When we do produce reasons for guidance, most of the time it is to guide others rather than ourselves. While we would like others to be guided by the reasons we give them, we tend to think that we ourselves are best guided by our own intuitions (which are based, we are sure, on good reasons, even if we cannot spell them out). (pp. 122-123)

Hence what Mercier and Sperber are suggesting is that reasoning did not originally come into being as a means of solving problems, but rather as a means of coping with each other within human communities, in particular, as a means of explaining and justifying oneself. As a result, we tend to excel at producing reasons which are impressive and persuasive without necessarily being fully sound; and we also tend to excel at checking the reasons of others for their soundness.

I think this grand picture is extremely interesting and offers us a fresh vista on reason and reasoning. The book also offers a very clear explanation of this picture and it is well documented with background material. Moreover, Mercier and Sperber fill in a lot of the picture's details; some of them again quite novel and interesting; others raising some doubts.

On the ground level, Mercier and Sperber see human minds as essentially modular, where each of the modules does its cognitive work largely independently of others and using nothing like reasoning:

> Modules, in any case, don't need reasons to guide them. They can use representations of facts as input without having to represent, either as a reason or in any other way, the relationship between these facts and the conclusions they derive from them. Modules don't need motivation or guidance to churn out their output. (p. 129)

Thus, though the popularity of the modular theory of mind appears to be declining (cf., e.g., Prinz 2006), Mercier and Sperber still use it as the point of departure for their theory of reasoning. What then, more precisely, is the cognitive work that the individual modules do? The short answer given by Mercier and Sperber is the drawing of *inferences*.

This is one of the points which I find puzzling. It is clear that on the modular theory of mind, each module takes care of coping with some part or aspect of the world; but why should this always be a matter of inference? If I understand the authors properly, their answer is contained in the following assumption:

> A main goal of cognitive mechanisms is to maintain an accurate representation of the organism's environment, or at least of relevant aspects of it. (p. 218)

An inference, then, is a mechanism that enables the module to produce further representations which can be used for prediction. However, if the main goal of cognitive mechanisms is to cope with the organism's environment (which, to be sure, *may sometimes* – or perhaps *often* – be achieved by manipulating representations of the environment), then inferring does not seem to be the common core of the modules' functioning. (Consider, for example, the theories of situated cognition, going back to Brooks (1991) and others: according to these, lots of coping with the environment can be done wholly avoiding representing the environment and manipulating its representations.)

An important question, of course, is what exactly *is* inference? The authors, for example, speak about inferences inherent to our visual perception, which seems to indicate that inferring is not necessarily something which we *do* with our representations, it may be something that *happens* to us. Then of course, it is less difficult to squeeze anything what the modules do into the boxes of *representing* and *inferring*, but then the concepts would not seem to be very useful.

Anyway, reason, on Mercier's and Sperber's view, turns out to be merely one new module. It is a module which, in effect, reflects on some of our inferences and seeks what makes us draw them, what we see as the (real or alleged) *reasons* for their outcomes. It seeks them especially because we might need them to negotiate our position within our society:

> We show, in other terms, how reason fits among other modules of intuitive inference rather than being a towering superpower. Notwithstanding its virtual domain generality, reason is not a broaduse adaptation that would be advantageous to all kinds of animal species. Reasons, we argued, are for social consumption. Reason is an adaptation to the hypersocial niche humans have built for themselves. (p. 339)

However, once we accept that we became inferring creatures without becoming reasoning creatures, the story that takes us to reasoning proceeds quite smoothly. It is, in essence, a story about us coming to reflect upon our inferences. In this way, we come to reflect that we have certain representations or do certain things because we came to have other representations, and we construct the picture of our peers – and of ourselves as *acting for reasons*:

> Reasons are social constructs. They are constructed by distorting and simplifying our understanding of mental states and of their causal role and by injecting into it a strong dose of normativity. Invocations and evaluations of reasons are contributions to a negotiated record of individuals' ideas, actions, responsibilities, and commitments. This partly consensual, partly contested social record of who thinks what and who did what for which reasons plays a central role in guiding cooperative or antagonistic interactions, in influencing reputations, and in stabilizing social norms. Reasons are primarily for social consumption. (p. 136)

Reasoning thus is primarily tied to social contexts and to language – in its primordial shape it is argumentation (and originally not even argumentation in the sense of cooperatively finding an objective truth, but in the sense of competitively negotiating one's position in a society). Reasoning as an inner mental process is parasitic on argumentation – thus it is also an essentially linguistic matter:

> Unlike verbal arithmetic, which uses words to pursue its own business according to its own rules, argumentation is not logical business borrowing verbal tools; it fits seamlessly in the fabric of ordinary verbal exchanges. In no way does it depart from usual expressive and interpretive linguistic practices. (p. 172)

Is the upshot, therefore, that it is merely illusory to believe that reasoning is an extremely useful tool that helps us attain knowledge and solve problems with a sophistication far beyond the ken of animals unable to reason? Is reason merely an

advocate that seeks to find justification for our preconceptions, disguised as an impartial judge seeking the truth? This is not quite the message of the book. Mercier and Sperber agree that reason can lead us to valuable conclusions, only it must be used in the proper way, where using it in this proper way means using it so that it chimes with its primordial function as much as possible.

What should we do if we want to reason with such beneficial effects? The most important thing, Mercier and Sperber argue, is that we should reason interactively:

> We construct arguments when we are trying to convince others or, proactively, when we think we might have to. We evaluate the arguments given by others as a means – imperfect but uniquely useful all the same – of recognizing good ideas and rejecting bad ones. Being sometimes communicators, sometimes audience, we benefit both from producing arguments to present to others and from evaluating the arguments others present to us. Reasoning involves two capacities, that of producing arguments and that of evaluating them. These two capacities are mutually adapted and must have evolved together. Jointly they constitute, we claim, one of the two main functions of reason and the main function of reasoning: the argumentative function. (pp. 207-208)

We are ingenious in coming up with reasons, though they are not always entirely sound. But we are also ingenious at checking the reasons of other people for their soundness, so when we work in coordination, opposing one another's tendency to spout not always very good reasons (or "reasons"), the interaction may yield something not so far from impartial reasoning and homing in on objective truth.

And what holds about reasoning in general, holds equally about what is often thought about as the quintessence of reasoning, science:

> Scientists' reasoning is not different in kind from that of laypeople. Science doesn't work by recruiting a special breed of superreasoners but by making the best of reasoning's strengths: fostering discussions, providing people with tools to argue, giving them the latitude to change their minds. (p. 329)

What, however, the view of reason and reasoning put forward by Mercier and Sperber does shatter is the recently popular view of man as a *Homo economicus*, who consistently maximizes his gains by means of his reason, which has developed precisely to serve this purpose. This theory, in contrast to Mercier's and Sperber's, sees reason as primarily an individual adaptation for finding objectively best solutions to problems the individual faces.

But does it not follow directly from evolution theory that surviving animals must become experts in solving the problems they encounter? And if so, should this not form also our reason – our principal means of the coping? Their answer is that being the kind of (hyper)social animals we humans are, most of the existential problems that we solve we face *as a society*, and the ability to negotiate one's position in a society is even more important for an individual than to directly handle natural menaces. Though it is probable that reason was, perhaps from the beginning, used *also* to solve problems and to deal with the environment, its social function was so much more important that it was this that shaped it.

Again, I think that in their zeal to revert the reader from the mistaken mainstream view of reasoning, the authors sometimes make dubious claims. Thus they write:

> Does, however, a syllogism that you know to be sound provide you, by itself, with a sufficient argument in favor of its conclusion? It is a common mistake to think so. (p. 166)

Well, I think that undeniably a sound syllogism does provide us with a sufficient argument; or if we doubt its premises, it moves us a step towards such an argument. What the authors probably want to say is that the syllogism often *de facto* does not serve as such an argument, that it can be used and misused in various ways and that we may have alternative means of becoming convinced of its conclusion.

On the whole I think that Mercier's and Sperber's book is extremely interesting and duly thought-provoking. And I suspect that their view of reason and reasoning, path breaking as it is, is largely correct – perhaps not in all details, but surely in the general outline. And I think that its consequences for our studying reason and reasoning are huge.

Jaroslav Peregrin

References

BROOKS, R. A. (1991): Intelligence without Representation. *Artificial Intelligence* 47, 139-159.

MORTON, A. (2010): Human Bounds: Rationality for Our Species. *Synthese* 176, No. 1, 5-21.

PINKER, S. (2010): The Cognitive Niche: Coevolution of Intelligence, Sociality, and Language. *PNAS* 107, 8993-8999.

PRINZ, J. (2006): Is the Mind Really Modular? In: Stainton, R. J. (ed.): *Contemporary Debates in Cognitive Science*. Oxford: Blackwell, 22-36.

Contents

ARTICLES

DISCUSSIONS

INTERVIEW

BOOK REVIEWS

REPORTS