Conditionals, Causal Claims and Objectivity

Michał Sikorski

A thesis presented for the degree of
Doctor of Philosophy

Dipartimento di Filosofia e Scienze dell'Educazione
Universita degli Studi di Torino
Italy

22.1.2020

Contents

Chapter 1

Introduction

In my thesis, I develop two distinct themes. The first part of my thesis is devoted to indicative conditionals and approaching them from an empirically informed perspective. In the second part, I am developing classical topics of philosophy of science, specifically, scientific objectivity and the role of values in science, in connection to recent methodological developments revolving around the Replication Crisis.

1.1 Conditionals and Causal Claims

Indicative conditionals are important and widely used expressions. They are central for reasoning, explanation or prediction. Still as mentioned by some authors (see e.g., Douven 2016) despite the long history of studies dedicated to conditionals almost everything about them is controversial. For example, the truth conditions of conditionals are controversial and the main competitors are the material implication, three-valued semantics, possible world semantics, and inferential semantics. Material implication is the oldest theory, it was developed as part of the project of the mathematization of logic at the beginning of the twenty century. It is a truth-functional proposal, it defines the truth-conditions of conditionals in terms of the truth values of its arguments. According to the theory, a conditional is false if its antecedent is true and its consequent is false and it is true in any other case. Since it was proposed, many problems with the material implication were diagnosed, for example, the fact that the falsity of antecedent is sufficient for the truth of the conditional is re-

garded to be unintuitive and therefore considered to be one of the paradoxes of material implication. Similarly, Edgington 1995 shows that the implications of the theory for the probability of conditionals are problematic. Despite all that the material implication still has some proponents. Typically such proponents defend the core truth conditions by including additional auxiliary assumptions (see e.g., Johnson-Laird and Byrne 2002 or Jackson 1987). The mentioned paradox of the material implication motivated another truth-functional approach, the three-valued semantics (see e.g., Finetti 1936, Cooper 1968 or Belnap 1970). Similarly as in the case of material implication these semantics recognize a conditional as false if its antecedent is true and its consequent is false and as true if both arguments are true. On the other hand, if the antecedent is false a conditional has the third value. The standard interpretation of the third value is undefined, the conditional in such case lacks the truth value. This difference makes the theory more intuitive which explains its popularity. It is popular among psychologists (see e.g., Baratgin et al. 2018) philosophers (see e.g., Cantwell 2008 or Egré, Rossi, and Sprenger 2019) or computer scientists (see e.g., Dubois and Prade 1990 or Goodman, Nguyen, and Walker 1991). The different approach is taken by proponents of possible world and inferential semantics. Both theories are based on the assumption that an adequate semantics for conditionals cannot be truth-functional. The possible world semantics claims that a conditional is true if the consequent is true in the most similar possible world in which antecedent is true. The original idea was presented in Stalnaker 1968 while related theories where proposed in Lewis 1973b or Kratzer 1989. Finally, the inferential semantics starts with the intuition that the connection between arguments is a necessary condition for the truth of a conditional. This connection can be conceptualized as an inferential connection, consequently, inferential semantics claims that a conditional is true if there is an inferential connection between the arguments of conditionals. Versions of inferential semantic were defended in Krzyżanowska, Wenmackers, and Douven 2014 and Douven et al. 2019. In contrast to all of these proposals some authors claim that conditionals are not truth-apt expressions and therefore do not have truth conditions at all. This view is called the non-truth value view (NTV). A big part of the motivation for it is dissatisfaction with the proposed semantics but some other arguments were proposed to support it, see for example Bennett 2003 or Edgington 1995 and Douven 2015 for a critical discussion.

Proponents of such a view have to explain why natural language users tend to systematically use some conditionals and not another. In the case of truth-apt sentences this is explained by truth values but how can we explain it if we deny that conditionals are truth-apt? The proponents of NTV typically use the notion of acceptability. Acceptability roughly means aptness to be accepted in conversation. Two types of acceptability are discussed in the context of conditionals, the categorical acceptability which typically plays the role of truth and graded acceptability which may be used as a substitute for a probability. Similarly, as in the case of truth conditions, the acceptability of conditionals is controversial. In the case of the graded acceptability, the standard view is Adams' Thesis (Adams 1975):

AT $ac(A \rightarrow B) = p(B|A)$

which equals the acceptability of a conditional with conditional probability of the consequent given antecedent. It was defended for example, by Jackson 1987. On the other hand, it was recently criticized from both theoretical (e.g., Hájek 2012) and empirical perspective (e.g., Skovgaard-Olsen, Singmann, and Klauer 2016) and alternative proposals were proposed (e.g., Crupi and Iacona 2019a). It is similar in the case of categorical acceptability. The standard view of categorical acceptability of conditionals is:

QAT An indicative conditional "If A, B" is assertable for/acceptable to a person if and only if the person's degree of belief in B given A, $Pr(B|A)$ is high.

An alternative proposal was proposed in Douven 2008 (see chapter 2). Moreover, Douven and Verbrugge 2012 present an experiment that shows that the alternative theory fits intuitions of natural language speakers better.

Other controversial topics include probability of conditionals (see e.g., Evans and Over 2004 and Skovgaard-Olsen, Singmann, and Klauer 2016) or updating with conditionals (see e.g., Eva, Hartmann, and Rad 2019). Finally, the best way to approach all these issues was also until recently controversial. Traditionally, philosophers based their theories concerning conditionals on their own intuitions elicited in consideration of the case by case studies. For example, that is how Jackson (1987) argues for Adams' Thesis which equates assertability of conditionals with the conditional probability of its consequent given antecedent:

> "Take a conditional which is highly assertible, say, 'If unemployment drops sharply, the unions will be pleased'; it will invariably be one whose consequent is highly probable given the antecedent. And, indeed, the probability that the unions will be pleased given unemployment drops sharply is very high."

Recently an alternative, experimental approach gained a lot of popularity. Conditionals received a lot of attention from psychologists, all of the issues mentioned above were empirically tested. The truth-conditions were tested for example, in Douven et al. 2019 or Krzyżanowska, Collins, and Hahn 2017 while the probability of conditionals were tested for example, in Evans, Handley, and Over 2003, Over et al. 2007 or Skovgaard-Olsen, Singmann, and Klauer 2016. As pointed out by Douven 2016, the results of such experiments are typically more generalizable and robust than the intuitions of a single author and therefore they should provide a firmer basis for new theories. In effect, this turn toward experiments should make progress more likely. So after almost twenty years of extensive empirical studies dedicated to conditionals was this promise fulfilled? Is there any more consensus because of all the collected empirical evidence? The answer seems to be a tentative yes. In recent years a dominant paradigm of thinking about conditionals emerged. Inspired by the work of Ramsey, de Finetti, and Adams, the paradigm is built around The Equation:

Equation $P(A \to B) = P(B|A)$

which equates the probability of conditionals with conditional probability of the consequent given the antecedent. It was supported by a number of empirical studies (e.g., Evans, Handley, and Over 2003, Over et al. 2007, Fugard et al. 2011 or Pierre and Gauffroy 2015) and in a big part because of that it become the core of the dominant approach to conditionals (see e.g., Over and Cruz 2018). The Equation is typically combined with a version of three-valued, de Finetti's semantic (e.g., Baratgin et al. 2018) or with rejecting the idea that the conditionals are truth-apt (e.g., Edgington 1995).

In the part of my thesis devoted to conditionals, I will follow the empirical approach to the study of conditional in basing my conclusion on the available empirical evidence rather than my intuitions. I will also present an original

experiment conducted (together with my co-authors Jan Sprenger and Noah van Dongen) to shed some light on the aspect of the semantics of conditionals which was not yet extensively studied, the relation between them and causal claims. On the other hand, I will argue against the dominant paradigm partly by appealing to newer experimental evidence provided for example in the work of Igor Douven, Niels Skovgaard-Olsen or Karolina Krzyżanowska.

1.2 Objectivity, Replication and Values

The second theme of the thesis is the Replication Crisis and its implication for some of the traditional issues of philosophy of science like the role of values in science or scientific objectivity.

Experimental science is undeniably successful. Just as in the case of conditionals, in countless other cases, the experimental approach proved to be fruitful. On the other hand, recently several problematic issues were diagnosed. Firstly, some of the practices which are commonly used by scientists were diagnosed to be detrimental for the reliability of the experiments through simulations or empirical studies (e.g., Simmons, Nelson, and Simonsohn 2011; Wicherts et al. 2016). Scientists rarely decide to out-rightly fraud their results (for a clear example see e.g., Stapel 2012). On the other hand, there seems to be a class of widely used practices which while not being fraudulent are credited with inflating the rate of false-positive results. Evidence suggest that these Questionable Research Practices are prevalent in science (e.g., Simmons, Nelson, and Simonsohn 2011). They include hypothesizing after results are known (e.g., Murphy and Aguinis 2019), data dredging (e.g., Head et al. 2015) or optional stopping (e.g., Montori et al. 2005). Similarly, the reliability of an experiment can also be compromised by biased methodological decisions (e.g., Wilholt 2008). Partly in effect of those practices estimated rates of replicability of experiments are disappointingly low (e.g., Open Science Collaboration 2015). This constitutes the Replication Crisis which was extensively discussed in methodological literature. The causes of the crisis were discussed (e.g., Ioannidis 2005) but also a way to remedy it and if it is problematic in the first place (e.g., Stroebe and Strack 2014). Surprisingly, the Crisis is not widely discussed.[1]

[1] A notable exception is Romero 2016.

How does all this connect to the issues like the role of values in science or objectivity? The philosophical literature devoted to both of these issues is built in opposition to a value-free ideal of science and value-free objectivity. A value-free ideal attributed to neo-positivists claims that scientists should not use their non-epistemic values when they justify their hypotheses. The value-free objectivity inspired by the ideal claims that a scientific practice to be objective must not involve choices motivated by non-epistemic values. The value-free ideal was undermined by several influential arguments like the underdetermination argument (e.g., Longino 1990) or inductive risk argument (e.g., Rudner 1953) and become unpopular. Today the main topic of discussions seems to be which of the non-epistemic values should be used or how can we use them, rather then if we should use them. For example, Douglas 2009 claims that only indirect uses of values are acceptable which roughly amount to values guiding the methodological choices made for the sake of the experiment.

In the case of the objectivity the dissatisfaction with value-free ideal was transmitted to the value-free objectivity which resulted in a multitude of alternative conceptualizations of objectivity (see e.g., Douglas 2004 or Koskinen 2018). Some of them are based on the new ideals of scientific conduct. For example, Longino 2004 proposes an alternative ideal of scientific conduct which recommends the inclusion of all values present in a given society combined with a critical discussion concerning the admissibility of all of them. At the same time, she proposes the corresponding conceptualization of scientific objectivity (Longino 1990) which claims that a scientific practice becomes more objective the more distinctive values were discussed and included for its purpose.

My assumption in working on both those issues was that the philosophy of science should be informed by current development in the methodology of science (a large part of which is preoccupied with Replication Crisis). In line with that, I will be arguing that legitimizing the use of non-epistemic values will legitimize types of problematic behavior that contribute to the crisis and therefore the rejection of value-free ideal may be premature and should be rethink. In the case of objectivity, this approach together with the conviction that concepts developed in the philosophy of science should be useful for scientists lead me to develop (together with my co-author Noah van Dongen) a scientifically useful notion of objectivity based on the methodological consideration

surrounding the Replication Crisis.

1.3 Outline of the Thesis

Part I Conditionals and Causal Claims

Chapter 1, discuss an empirical and theoretical standing of The Equation, Adams' Thesis, and Qualitative Adams' Thesis. I will try to show that despite their popularity all three theses are not well-supported by the available evidence and their role in the future studies of conditionals should be re-think. I will show that neither of the theses is supported by the results of empirical studies which includes irrelevant and negatively relevant conditionals and that theoretical consideration (e.g., triviality proofs) made accepting them, at the very least, theoretically costly.

In **Chapter 2**, I will show that empirical research concerning the acceptability of the irrelevant conditionals points to counterexamples to the Ramsey Test, widely accepted as the procedure for deciding if a given conditional is acceptable. I will also present a version of the test able to accommodate the problematic cases.

Chapter 3, defends a distinction to actual and tendency causal claims made in Hitchcock 2001 against the criticism made in Jakob 2006 on the ground of the Minimal Theory of Causation. I show that the theory is problematic and the alternative distinction proposed on its basis is unintuitive. I extend my criticism to INUS theory, the predecessor of the minimal theory of causation.

In **Chapter 4**, I (with my co-authors Jan Sprenger and Noah van Dongen) present a hypothesis concerning the relation between conditionals and causal claims. Roughly the hypothesis claims that a true tendency causal claims support the truth of the corresponding indicative conditionals. In the paper, we present experimental results of which supports our hypothesis and discuss consequences for theorizing about conditionals and causal claims.

Part II Values, Objectivity and Replicability

Chapter 5, argues for the value-free ideal of science by showing that its rejections force us to accept as cases of legitimate scientific conduct some of the

problematic scientific practices like biased methodological decisions or Questionable Research Practices. I will also point to a plausible way of making the ideal realizable against the famous argument from inductive risk and connect these issues to the Replication Crisis.

In **Chapter 6**, I (together with my co-author Noah van Dongen) propose a new conceptualization of objectivity. It builds on the negative theories of objectivity and recent developments in scientific methodology. The proposal aims to be, in opposition to the traditional notions, useful and testable conceptualization of objectivity.

Part I

Conditionals and Causal Claims

Chapter 2

Re-thinking the Acceptability and the Probability of the Indicative Conditionals

2.1 Introduction

Indicative conditionals, like:

(1) If you press this button, the fire alarm goes off.

are an important part of our language. We use them for example, to express our prediction or generalizations. Partly because of their importance, conditionals are interesting for philosophers and psychologists. They are interested, for example, in truth conditions[1] of conditionals or updating our beliefs with them.[2] Two other issues which received a lot of attention are probability and acceptability of indicative conditionals.

In the case of the probability, reasons for all this attention are clear. For instance, if we were able to define probabilities of conditionals we could incorporate reasoning which use conditionals into the popular and successful framework of Bayesian epistemology.[3]

[1] See e.g., Baratgin et al. 2018 or Jackson 1987.
[2] See e.g., Eva, Hartmann, and Rad 2019.
[3] See e.g., Talbott 2016 or Sprenger and Hartmann 2019.

In the case of the acceptability, the attention is a bit harder to explain. The acceptability conditions of other complex expressions are not so widely discussed. They are, to be sure, studied as a part of the pragmatics or epistemology but it seems that there is no, for example, a special problem of the acceptability conditions for conjunction. What is different in the case of conditionals? It seems to me that, it is an influence of a very popular philosophical position called the non-truth value view (NTV). It claims that conditionals do not have truth values.[4] The proponents of NTV have to deal with at least two problems. Firstly, we systematically recognize that some conditionals are appropriate to utter in some situations while others are not. In the case of other sentences it can be often explained by the difference between truth and falsity. So how can we explain that without postulating truth values for conditionals? Secondly, if we claim that conditionals are not truth-apt it seems natural to assume that they are not probability-apt. The probability of the sentence is the probability of that sentence being true and if a sentence is not truth-apt (think for example about command or question) it makes little sense to ask about its probability. If it is so, we even in principle cannot incorporate the conditionals to the Bayesian framework. The answer to both challenges is provided by the notion of acceptability. We can use graded acceptability as a substitute for probability and categorical acceptability as a substitute for truth.

The discussion concerning the probability and the acceptability of conditionals is mainly organized around two influential theses. The first of them is so fundamental for the currently dominant paradigm of thinking about conditionals (see e.g., Over and Cruz 2018) that it usually just called The Equation:[5]

Equation $P(A \to B) = P(B|A)$

The second thesis is called Adams' thesis:

AT $ac(A \to B) = P(B|A)$

where "ac" indicates acceptability. AT in this form is not a good substitute for truth conditions. It does not provide us with a threshold of acceptability above which a conditional would be acceptable. Such a threshold is provided by another version of AT; the Qualitative Adams' Thesis:

[4]For the details and motivation of the view see e.g., Bennett 2003 or Edgington 1995. For the critical discussion see: Douven 2015.

[5]See e.g., Edgington 1995.

(QAT) An indicative conditional "If A, B" is assertable for/acceptable to a person if and only if the person's degree of belief in $P(B|A)$ is high.[6]

All three theses were evaluated from both empirical and theoretical perspectives. In my article, I will examine both of these approaches and show that there are no convincing reasons to accept any of them and therefore we should re-think their role in the future study of conditionals. In the second section, I will discuss the experiments dedicated to all three theses. Then I will discuss the theoretical considerations for and against them. In the last section, I will conclude and point to some alternative conceptualizations of the probability of conditionals.

2.2 Empirical support

In this section, I will discuss the empirical experiments concerning three theses. Before that, I will make a distinction useful in this context.

Conditionals can be divided into positively relevant, irrelevant and negatively relevant. The positively relevant conditionals are conditionals whose antecedents are positively probabilistically relevant for their consequents. If a sentence is positively probabilistically relevant for another one, then the truth of the first sentence makes the second one more probable. The negatively relevant conditionals are conditionals whose antecedents are negatively probabilistically relevant for their consequents, which mean that the truth of the antecedent decreases the probability of the consequent. Irrelevant conditionals are the conditionals whose antecedents are probabilistically irrelevant for their consequents. The concept of relevance can be mathematically represented in at least two ways. Firstly we can use $\Delta P = P(B|A) - P(B|\neg A)$ proposed in Spohn 2012. If the value of ΔP is 0 the corresponding conditional is irrelevant, when it is higher then it is positively relevant and when it is lower the conditional is negatively relevant. Secondly, the relevance can be conceptualized as a difference measure $(P(B|A) - P(B))$. As in the case of ΔP when the value of difference measure is 0 the conditional, is irrelevant, if it is lower it is negatively relevant and if it is higher it is positively relevant. Both conceptualizations classify conditionals in the same way but the exact level of relevance

[6]The source of this formulation is Douven and Verbrugge 2012.

will differ in some cases.[7] Both notions were used in experiments concerning conditionals and the difference will not matter for our conclusions.

The example of intuitively irrelevant conditional is:

(2) If I eat an apple today, I will not inherit 1000000$ today.

and negatively relevant is:

(3) If he smokes, he will not develop a lung cancer.

Going back to our three theses, all of them have been traditionally regarded as descriptively true.[8] Philosophers generally found all of them confirmed by their introspective case by case studies. Many such case were presented for example, in Bennett 2003, Edgington 1995 or Jackson 1987:

> "Take a conditional which is highly assertible, say, 'If unemployment drops sharply, the unions will be pleased'; it will invariably be one whose consequent is highly probable given the antecedent. And, indeed, the probability that the unions will be pleased given unemployment drops sharply is very high."(Jackson 1987)

Systematic experimental studies were, firstly, directed toward The Equation. Results of most of these experiments support it. For example: Evans, Handley, and Over 2003, Over et al. 2007 or Oberauer and Wilhelm 2003 found significant correlation between participants responses concerning probability of conditionals and conditional probability while using different types of conditionals. For example Over et al. 2007 uses so-called causal conditionals,the conditionals which express causal relations. While Oberauer and Wilhelm 2003 use conditionals which describe relations between frequency distributions. A number of studies used moderators, for example, Fugard et al. 2011 showed that the proportion of responses in accordance with The Equation increases

[7]For a detailed discussion of the difference between the two notions and an experiment indicating that ΔP predicts intuitive relevance better than the difference measure, see Skovgaard-Olsen, Singmann, and Klauer 2017.

[8]For example McGee 1989: "Ernest Adams (1965, 1975) has advanced a probabilistic account of conditionals, according to which the probability of a simple English indicative conditional is the conditional probability of the consequent given the antecedent. The theory describes what English speakers assert and accept with unfailing accuracy, yet the theory has won only limited acceptance. "

with the time spent on practising the task. Similarly, Pierre and Gauffroy 2015 found that this proportion increases with the age of children participants and Evans et al. 2007 found that it increases proportionally to the mental capacities of participants. Finally, Cruz et al. 2016 found that The Equation correctly predicts participant judgments about negative conditionals.

Results of all of those studies support The Equation and together with similar results convinced many philosophers and psychologists that The Equation is a correct theory of how people reason with conditionals and made it and probabilistic theory based on it a dominant paradigm of thinking about conditionals.[9]

Both AT and QAT did not receive so much attention. AT was first tested in Douven and Verbrugge 2010. The authors use in the experiment inferential conditionals and they grouped them into inductive, abductive and deductive conditionals. Inferential conditionals are conditionals which express inferences. Inductive conditionals express inductive inferences, deductive express deductive and abductive express abductive inferences. In the first experiment, the authors tested Adams' Thesis and four weaker versions of it:

(WAT1) $Ac(A \to B) \approx Pr(B|A)$

(WAT2) $Ac(A \to B)$ is high/middling/low iff $Pr(B|A)$ is high/middling/low.

(WAT3) $Ac(A \to B)$ highly correlates with $Pr(B|A)$.

(WAT4) $Ac(A \to B)$ at least moderately correlates with $Pr(B|A)$.

The theses were tested by comparing their prediction with responses given by participants to the question concerning acceptability and probability of a given conditional. A sixty-seven students took part in the experiment. Questions about acceptability were manipulated between subjects and the type of inferential conditional was manipulated within the subjects.

Surprisingly, only a weak correlation between the conditional probability and the acceptability of conditionals was found. The correlation was especially weak in the case of inductive conditionals. It was not enough to support AT or even few weaker versions of it. Just the weakest version (WAT4) was supported for all kinds of conditionals (inductive, deductive and abductive). In the

[9]See e.g., Over and Cruz 2018 or Evans and Over 2004.

third experiment presented in the paper, participants were asked to judge the conditional probability of the consequent given the antecedent and the probability of the conditional. The results of the first experiment and the third experiment were compared. The comparison showed a significant difference between participants judgments concerning the acceptability and the probability of conditionals. I will discuss this issue later on.

QAT was, also, tested the first time by Igor Douven and Sara Verbrugge. The experiment was presented in Douven and Verbrugge 2012. The authors tested the predictions of QAT and the so-called Evidential Support Theory presented in Douven 2008:

"EST An indicative conditional "If A, B" is assertable/acceptable if and only if $Pr(B|A)$ is not only high but also higher than $Pr(B)$."

The idea behind EST is that the high conditional probability is not enough for a conditional to be acceptable and positive relevance has to be included as an additional condition. Results show that QAT predicted judgments of speakers worse than EST, and especially poorly in the case of irrelevant and negatively relevant conditionals. This result was replicated in Krzyżanowska, Collins, and Hahn 2017.

The similar idea, of using irrelevant and negatively relevant conditionals, was adopted by Skovgaard-Olsen, Singmann, and Klauer 2016. The authors tested The Equation and AT. The items include positively relevant and, crucially, irrelevant and negatively relevant conditionals. The results showed a significant correlation between the conditional probabilities and the probabilities of the positively relevant conditionals. At the same time, it was not the case for irrelevant and negatively relevant conditionals. There the probabilities of conditionals were much lower than the conditional probabilities. The results for the acceptability were almost the same. The failure of AT is not that surprising if we take into consideration the failure of its qualitative version and the results from Douven and Verbrugge 2010. On the other hand, the poor performance of The Equation is unexpected given the rich history of experiments which supported it. This result was replicated in experiments with different experimental designs. For example, results of Krzyżanowska, Collins, and Hahn 2017, Skovgaard-Olsen, Singmann, and Klauer 2017, Vidal and Baratgin 2017 and Fugard, Pfeifer, and Mayerhofer 2011 all suggest The

Equation does not correctly predict the probability of conditionals in the case of irrelevant and negatively relevant conditionals.

How should we explain this difference in results? The authors of Skovgaard-Olsen, Singmann, and Klauer 2016 claim that previous studies do not systematically include irrelevant or negative relevant conditionals and therefore cannot support the unrestricted version of Equation. For example, most of the conditionals considered in Over et al. 2007 seem to be intuitively positively relevant one.[10] It is similar in the case of Oberauer and Wilhelm 2003 or Hadjichristidis et al. 2001. The successful replications and the lack of the irrelevant and negatively relevant conditionals in the stimuli used in the earlier experiments strongly suggest that the effect of the relevance on the assessment of the probability or acceptability is robust and the support for The Equation provided by those experiments should be re-evaluated.

A defender of the Equation may claim that the effect of the relevance of conditionals is pragmatic and therefore the unrestricted version of The Equation can still be preserved. This solution is somehow supported by the results of Skovgaard-Olsen et al. 2017 where the authors showed that the effect of relevance on the assessments of truth is weaker than its effect on the acceptability or probability of conditionals. On the other hand, results of different experiments suggest that the relevance influence the truth assessments, for example, Krzyżanowska, Collins, and Hahn 2017 or Douven et al. 2019.[11] The hypothesis that the effect is pragmatic was also tested directly in Skovgaard-Olsen et al. 2019. The authors tested three hypotheses describing different pragmatic mechanisms generating the reason-relation part of the content of indicative conditionals responsible for the effect. Firstly they checked if it is cancelable in the way the conversational implicatures are, secondly, they tested if its projection behavior resembles one of the presuppositions and finally, they tested if it is treated as not-at-issue content which is believed to be one of the characterizing features of the conventional implicature. Surprisingly, the results of all three experiments were negative which suggests that the reason-relation part of the content is not conversational implicature, presupposition nor conventional implicature. The authors in discussing their results point that the features of conventional implicature (including it being not-at-issue content) are still very

[10] E.g.,"If Adidas get more superstars to wear their new football boots then the sales of these boots will increase" or "If the cost of petrol increases then traffic congestion will improve".

[11] For the discussion see: Douven 2017.

controversial. Because of the results of Skovgaard-Olsen et al. 2017, it is, still most likely that the reason-relation part of the content is conventional implicature. This is the opinion of authors does not make it a part of the pragmatic content given the conventional implicatures are typically classified as semantic content. In light of that, it seems that the pragmatic origin of the effect of relevance on probability or acceptability of conditionals is at least implausible.

Finally, we may wonder if it is possible to restrict The Equation to make it consistent with the available evidence? It seems possible. A version of The Equation restricted to the positively relevant conditionals seems to be in line with the results of all the mentioned experiments. Such a version can look for example like:

Equation+ If $\Delta P > 0$ then $P(A \rightarrow B) = P(C|A)$

At the same time, it puts all the theses in a somehow similar position. All of them were initially regarded as intuitive and supported by introspective case by case examination. In light of the available empirical evidence, both QAT and AT seem to be empirically inadequate. QAT performs poorly (Douven and Verbrugge 2012) in comparison to an alternative theory. AT was disconfirmed by results of Skovgaard-Olsen, Singmann, and Klauer 2016 which show that it fails in the case of the irrelevant conditionals and by results of Douven and Verbrugge 2010 which show that it is not supported in the case of the inductive conditionals. Similarly, the results which were considered to be evidence for The Equation seem to be undermined by the results of Skovgaard-Olsen, Singmann, and Klauer 2016 and consideration concerning the conditionals used in the earlier experiments.

2.3 Theoretical arguments

The theoretical studies concerning The Equation, AT and QAT has a longer history than the empirical ones. Still, it seems that there is not much of theoretical justification for the three theses. Even some of their defenders seem to agree. For example, Douven 2015 about The Equation:

> "While there is no known argument for this thesis showing that it has any normative force, to many the proposal does ring true, at least prima facie."

In this section, I will discuss the theoretical considerations presented for and against The Equation, ST and QAT. I will start by discussing the Ramsey Test, which is commonly used to argue for The Equation or AT. Then I will move toward trivialization proofs. I will discuss them with special attention dedicated to the two most popular ways to block it: denying that conditionals are propositions and postulating that the meaning of a conditional depends on the beliefs of the speaker. Finally, I will discuss the relationship between the semantics of conditionals and its probability.

Ramsey test

The Ramsey test was presented by Ramsey 1990 as a procedure for evaluating the acceptability of indicative conditionals:

> "If two people are arguing 'If p will q' and both are in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; so that in a sense 'If p, q' and 'If p, $\bar{q}$' are contradictories. We can say that they are fixing their degrees of belief in q given p."(Ramsey 1990)

The test is very popular among philosophers and psychologists,[12]and many cases in which its predictions are correct were considered and presented.[13] Because of this intuitiveness but also simplicity the procedure served as a direct inspiration for three successful research programs: belief revision theory, possible world semantics for conterfactuals and suppositional theories of indicative conditionals. The theories from the last group are typically committed to The Equation or AT. The Equation is a probabilistic reinterpretation of Ramsey test, therefore, the argument from the one to another is straightforward: If you accept the Ramsey test then you have to accept The Equation which is just its probabilistic reformulation.[14]

There are two problems with this argument. Firstly, the intuition behind the plausibility of both Ramsey test and The Equation seems to be exactly the same one. The second is merely a reformulation of the first and in all cases in

[12]E.g., "Most theorists of conditionals accept the Ramsey test thesis for indicatives."(Bennett 2003).

[13]See e.g., Evans and Over 2004 p 21-22.

[14]See e.g., Bennett 2003 or Evans and Over 2004.

which Ramsey test delivers a correct result, The Equation will give us just as satisfying answer. Therefore it seems that by appealing to the test we do not provide any independent evidence for The Equation.

Secondly, the close parallel between The Equation and Ramsey test and empirical results which established limits of The Equation point toward possible limits of the test. As we have seen in the previous section The Equation seems to fail for the irrelevant and negatively relevant conditionals. It seems to be similar in case of the Ramsey test, considers once again a negatively relevant conditional:

(3) If he smokes, he will not develop a lung cancer.

Let us say that the lifestyle of the person in question is perfect and he does not have any genetic predispositions to developing cancer so even in case he smokes the probability that he will develop cancer is really low for example 1%. In such case, if we conduct Ramsey test on (3) we will get the conditional probability of 99% and therefore we should believe in (3). Still because antecedent of (3) is negatively relevant for its consequent, 3) is hard to accept. This deficiency of the Ramsey test was considered and the revised version of the test was proposed in Rott 1986.

To sum up, it seems that the intuitions behind the Ramsey test are the same intuition which drives The Equation, therefore, appealing to the former does not provide any independent justification for the later. Secondly, the plausibility of the Ramsey test may be restricted to positively relevant conditionals.

Triviality proofs

Triviality proofs show that accepting The Equation leads to unacceptable conclusions. For example, the first proof from Lewis 1976 showed that we can infer from The Equation that $P(A \rightarrow B) = P(B)$ which is generally false:

(4) $P(A \rightarrow B)$

(5) $P(A \rightarrow B|B)P(B) + P(A \rightarrow B|\neg B)P(\neg B)$

(6) $P(B|A,B)p(B) + P(B|A,\neg B)P(\neg B)$

(7) $P(B)$[15]

As we have already mentioned the conclusion is clearly unacceptable. The two most popular ways to block the proof is to deny that conditionals are propositions (e.g., Bennett 2003 or Edgington 1995) or to postulate that the meaning of conditionals depends on beliefs of the speakers (e.g., Douven 2015 or van Fraassen 1976).

The first option involves accepting NTV. It claims that the conditionals are not propositions and are therefore not truth-apt. If conditionals are not propositions they cannot occur in Boolean combinations therefore, for example, we cannot use the law of total probability on conditionals therefore Lewis' proof is blocked.

But how plausible is NTV? There are a few presented arguments for this conclusion, I will discuss one of them later on and all of them were, in my opinion convincingly, countered in Douven 2015. On the other hand, the rejection of propositional view seems to be really costly and these costs are rarely acknowledged.

First of all, one of the consequences of NTV is that conditionals no longer have a probability. Probability of a sentence is typically understood as the probability of this sentence being true, therefore if a sentence is not truth-apt it is also not probability-apt. Because of that, we have to replace The Equation with AT. It describes the acceptability of conditionals, and therefore, does not require them to have probabilities.

Secondly, the NTV has a problem with explaining the way conditionals are regularly used as premises in reasoning. Typically we understood the validity of reasoning as the preservation of truth. If one of these premises is not truth-apt there is nothing to be preserved. Therefore NTV makes reasoning involving conditionals unexplainable if one understand validity as truth preservation. This is an instance of so-called Frege-Geach problem[16] and to solve it one would have to propose an alternative, revisionary way of understanding the validity of reasoning. One such proposal, p-validity was presented in Adams 1975 in which AT was also defended:

[15]Steps from (4) to (5) and from (6) to (7) are instances of probability rules, $P(x) = P(x|y)P(y) + P(x|\neg y)P(\neg y)$ and $P(x|y, \neg x) = 0$.

[16]See e.g., Kolbel 1997.

> "...an inference to be *probabilistically valid* (abbreviated p-valid) if and only if the uncertainty of its conclusion cannot exceed the sum of the uncertainties of its premises." (Adams 1998 p. 131)

This proposal on its own will not help us with our problem. As we have seen above one of the consequences of NTV is that the conditionals cannot have probability, or at least not in the sense the truth-apt sentences do.[17] It seams that acceptability cannot be used in computing p-validity given that it is typically believed to has different properties than probability. Those differences cause a probabilistic version of Frege-Geach problem, a probabilistic framework (e.g., Bayesianism or p-validity) cannot accommodate conditionals which does not have probability. In effect even if we use p-validity it still not clear how to understand the reasoning with mixed conditional and non-conditional premises.

Thirdly, accepting NTV makes it hard to make sense of conditionals embedded in truth-functional contexts like disjunction or conjunction, for example:

(8) Either he is in Rome, if he is in Italy, or he is in Bordeaux, if he is in France.[18]

According to NTV, the conditionals are not type of things which can occur in such contexts. The evaluation of the whole sentences requires its arguments to be true or false but according to NTV conditionals are not. The defenders of AT developed elaborate ways of explaining away such sentences (see e.g., Edgington 1995), at the same time others come up with new examples harder to explain away (see e.g., Kolbel 2000).

All these problems seem to suggests that conditionals behave as truth-apt propositions. It is also suggested by the reaction of participants of the experiment asked to asses truth values or probability of conditionals. They perfectly well understand both questions about truth-values (see e.g., Douven et al. 2019

[17] In fact Adams 1975 claims that this natural interpretation of probability is not applicable to conditionals. His seems to be aware of how problematic consequences of NTV are for example:

> "The author's very tentative opinion on the 'right way out' of the triviality argument is that we should regard the inapplicability of probability to compounds of conditionals as a fundamental limitation of probability, on a par with the inapplicability of truth to simple conditionals."(Adams 1975 p.35)

[18] Example from Kolbel 2000.

or Krzyżanowska, Collins, and Hahn 2017) and probabilities of conditionals (e.g., all the articles which test The Equation) and not seem to be confused by neither of them. This is, once again, unexpected if conditionals are not propositions, consider for example asking somebody about truth-value of a question. In light of that, denying that the conditionals are propositions is both unintuitive and costly.

The second popular way to dodge the triviality was explored in Douven 2015 (after van Fraassen 1976). The prove to use a generalized version of the Equation, GSH:[19]

GSH $p(A \rightarrow B|C) = p(B|A,C)$

It was used to infer (6) from (5). Lewis derives GSH form three assumptions. The first assumption claims that the considered class of probability functions is closed under conditionalization. The second assumption is The Equation and the third is that the interpretation of the natural language indicative conditionals does not depend on the belief states of the speaker. I will refer to this assumption as the independence assumption or IA. Both Douven 2015 and van Fraassen 1976 argue against the assumption in order to save the Equation.

Van Fraassen believes that the source of Lewis' assumption is his metaphysical view, so-called modal realism. According to modal realism, possible worlds are real and objective in the sense in which the actual world is. If we combine modal semantics which defines the meanings of conditionals in terms of the properties of possible worlds with the modal realism, the meanings of conditionals do not depend on our beliefs but on the objective properties of possible worlds. Van Fraassen claims that if we adopt a less realistic notion of possible worlds, the assumption loses its appeal. If possible worlds are not objective and in some sense depend on our beliefs then the meanings of conditionals will also depend on them. Douven 2015 discusses the IA in more detail. He gives three arguments against it and attacks some of the arguments, which were presented for it. I will start by discussing his three arguments:

Firstly, some of the popular and promising semantic theories proposed for conditionals suggest that IA is false. The two theories mentioned by the author are Stalnaker–style modal semantics which uses the notion of similarity between possible worlds and inferential semantic.

[19]The Equation is sometimes called Stalnaker hypothesis, therefore its generalized version is called Generalized Stalnaker Hypothesis(GSH).

The Stalnaker semantics can be interpreted in a way in which it supports IA. The realistic interpretation held, according to Van Fraassen, by Lewis is an example of such interpretation. More importantly, Stalnaker semantics is inconsistent with The Equation (see e.g., Stalnaker 1976). Therefore appealing to it in order to attack IA and defend The Equation not seems to be a convincing strategy.

The inferential semantics presented in Krzyżanowska, Wenmackers, and Douven 2014 seems to be a very promising theory. Its main claim is:

Definition 1 A speaker S's utterance "If p, q" is true iff (i)q is a consequence—be it deductive, abductive, inductive, or mixed-of p in conjunction with S's background knowledge, (ii) q is not a consequence—whether deductive, abductive, inductive, or mixed—of S's background knowledge alone but not of p on its own, and (iii) p is deductively consistent with S's background knowledge or q is a consequence (in the broad sense) of p alone.

If we consider this formulation, it is not clear why the inferential semantic supports rejection of IA. The meanings of conditionals are here relative to the knowledge but not to the beliefs of the speaker. Authors explain that it would be counter-intuitive to treat as true, conditionals whose consequence were inferred from antecedents with use of false beliefs.

Douven 2015 presents a different version of the theory:

Definition 2 "a conditional is true in a given context iff the consequent follows via a number of steps from the antecedent, possibly in conjunction with contextually accepted background premises where, first, the steps are valid in deductive, inductive or abductive sense, and second the consequents does not follow (in the same generalized sense) from the premises alone."

According to him the belief-sensitivity of conditionals is imposed by this version of the semantic because the acceptability of potential background premises depends on the beliefs of the speaker or evaluator. This dependence causes the second formulation of inferential semantic to collide with IA but it also makes the proposal vulnerable to the problem which motivated the phrasing of the first formulation.

If the speaker or the evaluator is liberal in accepting the background premises for example, he accepts as premises all beliefs of the speaker then his false beliefs can be a basis for true conditionals. For example, let us assume that

I believe that the moon is made of cheese and all my beliefs are acceptable premises for my conditionals. It is known to all of my interlocutors that I share this preposterous belief. It is easy to see that according to the Definition 2 a conditional:

(9) If we bring the moon to the surface of the earth, we will end the world hunger.

uttered by myself is true. Still, it seems to me that none of my sane interlocutors would agree to it. The fact that they know that I believe that the moon is made of cheese seems to make no difference for their assessment of (9) (uttered by me). This seems to suggest that the Definition 2 is too permissible in the way it relates the truth of a conditional to the beliefs of the speaker or evaluator.

Secondly, Lindström 1996 proposed rejecting IA as a way out of so-called Gärdenfors' Paradox (Gärdenfors 1986). The paradox shows that no non-trivial belief system can at the same time satisfies both the Ramsey Test and the Preservation Condition:

"(P) If a proposition B is accepted in a given state of belief K and A is consistent with the beliefs in K, then B is still accepted in the minimal change of K needed to accept A."(Gärdenfors 1986 p. 82)

(P) seems to be a very natural assumption while the Ramsey Test, as we have seen, is a popular procedure for testing conditionals. Lindström shows that we can have both if we drop IA. As we have already noted it seems that appealing to Ramsey test, of which The Equation is a probabilistic reformulation, to defend the thesis seems to not give us a lot of additional independent evidence. Secondly the empirical evidence concerning the effects of relevance on the probability of conditionals suggests that the intuitiveness of Ramsey test may be limited so despite its popularity it may not be worth preserving. As an independent justification for rejection of IA Lindström presents the certeris paribus cases. These are cases in which we cease to accept a conditional after we learned some additional evidence. An example of such case is:

(10) If I pass today's exam, I will go for a beer afterward.

which is true, or at least acceptable, about me. But it ceases to be the case if I learn that I have another, very hard exam tomorrow. Lindström claims that when I learn about the second exam, (10) changes its meaning. If (10) conveys the second meaning it is false while if it has the first meaning (the meaning it had before I learned about the second exam) it is, still, true. This explanation of the ceteris paribus cases seems to have an unintuitive consequence. Let us consider a discussion between me and my friend, she knows about the second exam of which I am still unaware. We disagreed about (10). According to Lindström's proposal, we talk past each other because each of us means different things by (10). This is unintuitive.

Finally, Douven 2015 points that similar proposals were made for different expressions (e.g., taste predicates, modal operators). This is undoubtedly true but as far as I know, neither of these proposals is uncontroversial (see e.g., Hirvonen, Karczewska, and Sikorski 2019). Even if it was the case that was an uncontroversial one, it is not clear why their success should tell us anything about conditionals.

It seems that the postulated relativity should be reflected by the way we use conditionals. As far as I know, the only reported phenomenon which can suggest it is so-called Gibbard phenomenon. Consider the following story:

"Sly Pete and Mr. Stone are playing poker on a Mississippi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good, and signals its content to Pete. My henchman Jack sees both hands, and sees that Pete's hand is rather low, so that Stone's is the winning hand. At this point, the room is cleared. A few minutes later, Zack slips me a note which says "If Pete called, he won," and Jack slips me a note which says "If Pete called, he lost." I know that these notes both come from my trusted henchmen, but do not know which of them sent which note. I conclude that Pete folded." (Gibbard 1981, p. 231.)

Now according to Gibbard if both conditionals are true they would together with so-called conditional non-contradiction rule:

CNC $\neg((A \to \neg B) \wedge (A \to B))$

lead to inconsistency. Both conditionals are based on true beliefs and the support for them seems to be symmetrical. Therefore there is no reason why we should ascribe them different truth values or judge any of them as false. Gibbard concludes that both conditionals are acceptable and the existence of such

pairs is an argument for NTV. There seems to be a problem with this argument. The observation that in this situation both conditionals are acceptable is in tension with The Equation (and even more so with QAT).[20]

It is easy to see that according to the thesis it cannot be the case that both $(A \to B)$ and $(A \to \neg B)$ are highly probable at the same time. Therefore it is the case of two acceptable conditionals which cannot have at the same time a high probability $(< 50\%)$. That seems to show that we cannot use the example to argue for NTV to defend The Equation or AT.

The phenomenon is very controversial, many different interpretation were proposed. For example, Lycan 2003 denies that the support for both conditionals is symmetrical and therefore claims that just one of them is true. Finally, following Krzyżanowska, Wenmackers, and Douven 2014 one can claim that the meaning of a conditionals depend on the beliefs of the speaker. In case described by Gibbard it is clear that both Zack and Jack based their conditionals on different beliefs based on different evidence. Because of that both conditionals despite their superficial form are not in any tension and therefore not inconsistent even when combined with CNC, they are based on different beliefs and therefore they express different relations. According to this interpretation of the phenomenon, it in fact, supports rejections of IA.

It seems to me that it is unclear if the natural language speakers are willing to accept the Gibbard-like pairs of conditionals. Even If they do it is even less clear how to interpret this phenomenon. In light of that, it not seems to be the case that this argument makes IA significantly less plausible.

At the same time it should be noted that the rejection can have potentially unwelcome consequences. For example, it was noted by Lewis 1976 that it is not clear if we are able to explain a disagreement about conditionals if they meaning is relative in the proposed way (in line with our discussion of (10) above). It was countered by Douven 2015 that it is not necessary for the dis-

[20]It is also discussed in Jackson 1987: "When A is consistent, there is something quite generally wrong with asserting both $(A \to B)$ and $(A \to \text{not-}B)$. We cannot assert in the one breath 'If it rains, the match will be cancelled' and 'If it rains, the match will not be cancelled'. This conforms nicely with [AT]; for, by it, we have $As(A \to B) = 1 - As(A \to \text{not-}B)$, from the fact that $P(B/A) = 1 - P(\text{not-}B/A)$. Thus, the fact that $(A \to B)$ and $(A \to \text{not-}B)$ cannot be highly assertible together when A is consistent is nicely explained by [AT] as a reflection of the fact that $P(B/A)$ and $P(\text{not-}B/A)$ cannot both be high when A is consistent. Indeed, [AT] explains the further fact that $(A \to B)$ and $(A \to \text{not-}B)$ have a kind of 'see-saw' relationship. As the assertibility of one goes up, the assertibility of the other goes down."

agreement that the arguing parties interpret the proposition in question in exactly the same way. On the other hand, it seems that we should agreed with Lewis that it may be hard to account for disagreement on the basis of theory which makes the meaning of conditionals relative to opaque features of speaker (her beliefs). It is not clear if such explanation which not falls into other problems is available.

Finally, it seems that rejecting IA would be in tension with The Equation. The Equation claims that the probability of a conditional depends just on the conditional probability of its antecedent given its consequent and not on any other factors. If we reject IA we claim that the meaning of a conditional and therefore its truth condition as depends on some other factors namely beliefs of the user. If we assume that the probability of a sentence is determined by it truth condition, which seems to be a natural assumption, then it seems that meaning relativized to beliefs does not correspond well to the probability which is not relativized.

A number of other triviality proofs were proposed for example, Carlstrom and Hill 1978, Milne 2003 or Fitelson 2015.[21] As far as I know all of this proofs are blocked by NTV but not by rejecting IA. For example, in order to block a triviality proof from Hájek 1989 Douven has to claim that no finite model can represent a rational agent belief states (Douven 2015). Discussing plausibility of this assumption goes beyond the scope of the paper.

It is hard to consider the triviality proofs to be the conclusive arguments against The Equation. The two discussed ways to block the proofs, despite its discussed problematic consequences, are available and they are hardly the only one (see e.g., Bradley 2000.). On the other hand, as far as I know none of this ways can be consider especially attractive and therefore the triviality proofs shows, at the very least, that sticking to The Equation is costly.

Hájek 2012 argued that AT is also susceptible to a triviality proof analogous to one he presented in Hájek 1994 against The Equation. He points there that a plausible conceptualization of the acceptability has to share with the probability features which made it susceptible to his argument.

[21] For discussion see: Hájek and Hall 1994.

Truth conditions and Probability

What is the relation between the truth conditions of a sentence and its probability? Let us start by considering a sentences which are not truth-apt and therefore have no truth conditions. In such cases attributing probability to such sentences seems to be a category mistake. As we have already seen it seems nonsensical to ascribe probabilities to questions or commands, uncontroversial and prototypical examples of not truth-apt sentences. If a sentence in question is truth-apt as I already hinted a natural and straightforward interpretation seems to be:

SP The probability of x is the probability of it being true.

This interpretation of the relation between the semantic and probability seems to be uncontroversial to the point that, as far as I know, no alternative was proposed.[22] It captures the relation between probabilities of complex sentences and its arguments, for example general probability rule for disjunction: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ reflects its truth conditions: $(A \text{ or } B)$ is true iff (A) is true or (B) is true. The disjunction is true if one of the disjuncts is true therefore to get the probability of disjunction we have to sum up probabilities of the disjuncts. If the conjunction of the disjuncts has non-zero probability it has to subtract. The truth of the second disjunct does not make disjunction any more true if the first one is already true. At the same time, this subtraction assures that the axioms of probability will not be violated. Is the relation the same in the case of conditionals? It seems so. If we adopt the NTV view we are in the first case and, as we have already shown, we have to retreat from The Equation to the AT which does not postulate anything about the probability of conditionals, as is done for example in Adams 1975. Therefore SP is trivially fulfilled, no truth and no probability. Otherwise, we have to explain how it is possible that conditionals do not have truth values but have probabilities.

Propositional semantics also adhere to SP. For example, the authors of Byrne and Johnson-Laird 2009 defend the mental model theory according to

[22] Adams 1975 reject SP for conditionals but as far as I understand, he does not provide an alternative. At the same time, his theory is usually interpreted as describing acceptability of conditionals rather than their probability.

which the truth conditions of natural language conditionals are those of material implication: $(A \rightarrow B)$ is true iff (A) is false or (B) is true. Consequently they propose a fitting probability definition: $P(A \rightarrow B) = P(\neg A \text{ or } B)$. So, the relation between semantic properties and probability of conditionals conforms to SP and therefore the theory, despite it other well-described shortcomings, provide a coherent picture of truth and probability.

In light of that, it is interesting to see if there is a semantic theory which can provide a basis for The Equation or conversely what semantic properties are suggested by it.

A semantic theory which is, together with SP, consistent with The Equation was provided by de Finetti.[23] He presented it as part of his more general subjective Bayesian theory of reasoning. In his Finetti 1980 he proposes three levels of knowledge. *Level 0* describes the objective knowledge and is well described by the binary logic. *Level 1* describe categorical knowledge as possessed by humans and therefore it includes the third logical value *uncertain* which represents a given individual being uncertain about a given sentence. Finally, *Level 2* is human knowledge represented in a graded numerical way. De Finetti three-valued semantics for conditionals is a part of a description of *Level 1*. According to it, a conditional is true if both antecedent and consequent are true, is false if the antecedent is true and consequent is false and it is uncertain or void if the antecedent is false. The semantics is supported by many experiments in which participants tend to produce so-called defective truth table and is popular among both psychologist and philosophers.[24]

What do these truth conditions tell us about the probability of conditionals? In the words of Over and Cruz 2018:

> "The probability of the conditional if p *then* q for is the probability that $p\&q$ holds given that the conditional makes a non-void assertion, that p holds, and this probability is of course the conditional probability of q given p, $P((p\&q)|p) = P(q|p)$."

So as we see, this semantic theory implies The Equation which makes it well supported. It was also not an accident that the two levels fit well together. It was an intention of De Finetti to present a unified system (composed of

[23] See Baratgin et al. 2018 for discussion.

[24] See e.g., Egré, Rossi, and Sprenger 2019. For discussion of defective truth tables see: Baratgin et al. 2018 or Over and Baratgin 2016.

all three levels) describing the single phenomenon, in this case, the indicative conditionals. So is The Equation satisfactory justified by De Finetti semantics? Sadly, there are two problems with that justification. Firstly, it seems that the semantics assumes that the conditionals are propositions. Conditionals according to the theory are sometimes true (or false) and typically that is enough for something to be a proposition. On the other hand, de Finetti's conditionals are sometimes described as not expressing "(full) proposition".[25] It is not exactly clear to me how to interpret that, but as we have seen De Finetti's conditionals have probabilities. Therefore the way to block the Lewis' proof by declaring conditionals not to be propositions and therefore denying them having probabilities is not available anymore.

At the same time, the semantics does not make space for the relativisation of truth conditions so the second way to avoid trivialization seems to be closed. In light of that, I see no way in which one can accept the semantics, a probability conditions that it implies and still avoid trivialization.

Secondly, as we have seen the empirical support for The Equation is restricted to the positively relevant conditionals. The semantics supports the unrestricted Equation. It is unclear if and how it can be modified in order to support the qualified version of Equation. Therefore it seems that we deal here with a curious situation in which empirical and theoretical considerations seem to pull in opposite directions. The unrestricted version of The Equation is theoretically justified by the corresponding semantics but not supported by the empirical results, while it is not clear if the restricted version supported by the totality of empirical evidence is supported by any semantic theory.

The situation is a bit more complicated in the cases of QAT and AT. It is so because it is not clear what is the relation between the truth and the acceptability of a given sentence. In light of that, it seems that if we have any theoretical justification for QAT or AT it will come from their relation to The Equation.

Probability and Acceptability

In this section, I will discuss the possible conceptual relation between all three theses.

[25] See e.g., Wijnbergen-Huitink, Elqayam, and Over 2015.

The relation between probability and acceptability is a well-discussed topic in philosophy. The most straightforward way to relate the two notions is the Lockean Thesis:[26]

LT A proposition φ is acceptable iff the probability of φ is high.

From The Equation and LT we can deduce QAT. The intuition behind LT seems, also, to supports AT. If categorical acceptance coincides with high probability then it seems natural that, if there is something like graded acceptability, it will coincide with probability. But what if we accept the NTV and therefore deny that the conditionals have probabilities? It seems that in such case we have to reject LT in order to be still able to claim that conditionals have acceptability at all. If we endorse any other theory of acceptability[27] it seems that we are losing the theoretical basis for QAT and AT.

In this place, we should also point out another controversial issue, namely the differences in our intuitions concerning the acceptability and the probability of conditionals. Results from Skovgaard-Olsen, Singmann, and Klauer 2016 found no significant differences between assignments of acceptability and probability to conditionals made by participants. This suggests that $P(A \rightarrow B) = ac(A \rightarrow B)$. On the other hand, Douven and Verbrugge 2010 found a significant difference in the case of the inductive and abductive conditionals. Possible explanation is that Skovgaard-Olsen, Singmann, and Klauer 2016 used causal conditionals while Douven and Verbrugge 2010 used inferential conditionals. If so it may be the case that there is a difference between our intuitions concerning acceptability and probability is restricted to the inferential conditionals. It seems that more evidence should be collected in order to settle this issue. Replicating both experiments may be a good first step.

2.4 Conclusion

In this section, I will try to conclude by judging how the theses stand against the presented evidence then I will discuss the proposed and possible alternatives to the three theses.

[26] LT seems to be quite popular, see e.g., Foley 2009.

[27] Alternative theories are usually more complex see e.g., Proust 2012.

How the three theses stand against the presented evidence? Let us start with the theoretical considerations. All these seem to be in a similar situation. There seems to be no strong theoretical arguments for any of them. The intuitions behind the Ramsey test seems to be the same intuitions which initially make the theses plausible. Therefore appealing to it not give us additional reasons to believe it. The Equation is supported by de Finetti three-valued semantics consideration but adopting the semantics make two most popular strategies to dodge trivialization unavailable. QAT is supported by The Equation if we agreed on LT and unsupported otherwise. AT seems to be, to some degree, supported by QAT.

At the same time, we have strong arguments against The Equation in the form of the triviality arguments. Neither of them seems to be conclusive, given the possible way to dodge it. On the other hand, they convinced some philosophers to abandon The Equation (e.g., Stalnaker 1976) and showed that sticking to The Equation is costly. For example, we have to abandon IA which, as I tried to show in the third section, is plausible. A triviality argument of similar strength was also presented against AT. On the other hand, I am not aware of any comparable theoretical arguments against QAT.

As we have seen, all three theses were traditionally regarded as descriptively true but the results of the empirical studies seem to suggest a different picture. The situation is more complicated in the case of The Equation than in the case of AT and QAT. QAT and AT attracted much less attention than The Equation but, as far as I know, they were not supported by results of any of the relevant studies. AT was disconfirmed by Skovgaard-Olsen, Singmann, and Klauer 2016 which showed that it fails in the case of the irrelevant and negatively relevant conditionals and Douven and Verbrugge 2010 which showed that it is not supported in the case of the inductive conditionals. QAT performs poorly (Douven and Verbrugge 2012) in comparison to EST.

The Equation has a long tradition of good performance in empirical studies. On the other hand, the results of Skovgaard-Olsen, Singmann, and Klauer 2016 strongly suggest, that it fails in the cases of irrelevant and negatively relevant conditionals. The result was conceptually replicated by a few subsequent studies. At the same time, as is point out in Skovgaard-Olsen, Singmann, and Klauer 2016 the experiments which confirmed The Equation did not include irrelevant on negatively relevant conditionals and therefore did not use

a representative sample of conditionals. This seems to undermine them and together with results of Skovgaard-Olsen, Singmann, and Klauer 2016 suggests that overall the unrestricted Equation is not empirically adequate.

In light of all that, it seems that we have neither theoretical nor empirical reasons for accepting the theses beyond their initial intuitiveness. In light of that, it seems that their role in the future study of indicative conditionals should be re-thought.

On the other hand, I did not show that any of the theses is false. Conclusive arguments against them, as far as I know, do not exist and maybe never will. Specifically, someone impressed with the intuitiveness of any of the theses may treat it as a desideratum to be satisfied by a successful theory of conditionals. Even in such case the tension between them and some of the empirical findings and involved theoretical costs should remain clear.

Now we can discuss alternative proposals. I will start with the Evidential Support Theory proposed by Douven 2008. As we have seen, the core of the theory is the Evidential Support Thesis(EST):

"EST An indicative conditional "If A, B" is assertable/acceptable if and only if $Pr(B|A)$ is not only high but also higher than $Pr(B)$."

It is a counterproposal to QAT. In subsequent Douven and Verbrugge 2012 it was shown that EST predicts intuitions of natural language users much better than QAT. This is a clear advantage of EST and a good reason to prefer it over QAT. On the other hand, as it stands now, this approach also lacks theoretical justification.

EST is not supported by The Equation in a way in which QAT is and, as far as I know, it is not supported by any proposed semantics for conditionals. Perhaps further work on inferential semantics can provide a theoretical basis for EST.

As we have seen, EST is empirically more successful than QAT because it classifies irrelevant and negatively relevant conditionals as not acceptable. In light of that, it seems natural that users of language will judge the acceptability and the probability of conditionals as lower in such cases. Skovgaard-Olsen, Singmann, and Klauer 2016 showed that this is true. If so, maybe we can restrict The Equation and AT to be more in line with this finding. As we have seen, a restricted version of both may look, for example:

Equation+/AT+ If $\Delta P > 0$ then $P/ac(A \rightarrow B) = P(C|A)$

The Equation+ and AT+ are more consistent with available empirical evidence than the original theses. Because of the restriction, they are not undermined by the results of Skovgaard-Olsen, Singmann, and Klauer 2016, but AT+ is still undermined by the results of Douven and Verbrugge 2010.

What about their theoretical position? Once again we lack any theoretical motivation for both theses. The situation is even worst in the case of The Equation+. There is nothing in it which would block a triviality proof analogous to Lewis one restricted to the positively relevant conditionals. The result of the proof will be that for all positively relevant conditionals $P(A \rightarrow B) = P(B)$. This is just as unacceptable as the original unrestricted result. The bottom line here seems to be that if The Equation is proposed for any kind of conditionals we can make the Lewis-like proof for these conditionals. $P(A \rightarrow B) = P(B)$ is true for irrelevant conditionals but The Equation restricted just to them would be both uninteresting and empirically inadequate (as suggested by the results of Skovgaard-Olsen, Singmann, and Klauer 2016).

Let us move to theoretical considerations concerning conditionals. Can they point us toward a new definition of probability (or acceptability)? Triviality proofs do not give us clear advice concerning probability and acceptability of the conditionals. They provide us with a purely negative lesson concerning The Equation (and AT) and it is hard to predict which of the alternative proposals will be susceptible to analogical arguments.

Perhaps more promising and natural approach is to start with truth conditions proposed by some of the plausible semantics and on the basis of that work out a corresponding probability conditions. Most of the popular semantic theories postulate complex and subtle truth conditions which translate into similarly complex definitions of probability.[28] For example, if we combine, the already presented inferential semantics with SP we will get:

IP The probability of a speaker S's utterance "If p, q" is the probability that (i)q is a consequence—be it deductive, abductive, inductive, or mixed-of

[28] As we have seen the material implication theory is an exception. It provides us with truth conditions which can be easily translated into the definition of probability. Sadly, both the definition of probability and truth conditions proposed by the material implication theory seems to be unintuitive.

p in conjunction with S's background knowledge, (ii) q is not a consequence—whether deductive, abductive, inductive, or mixed—of S's background knowledge alone but not of p on its own, and (iii) p is deductively consistent with S's background knowledge or q is a consequence (in the broad sense) of p alone.

It is easy to see that, IP is less elegant and harder to test than The Equation. At the same time, it is directly justified by the Inferential Semantic. That alone puts IP in better theoretical position than The Equation and perhaps it is enough to make it worth further studies. Can it accommodate the existing evidence concerning the probability of conditionals? Can we construct trivialization arguments against it or perhaps show that it is impossible? Answering those questions goes well beyond the scope of the paper. On the other hand, I hope that this example shows that there are promising alternatives to The Equation and therefore we are not stuck with it.

Chapter 3

The Ramsey Test and Evidential Support Theory

3.1 Introduction

The Ramsey Test (RT) was presented by Ramsey 1990 as a procedure for evaluating the acceptability of indicative conditionals:

> "If two people are arguing 'If p will q' and both are in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; so that in a sense 'If p, q' and 'If p, $\bar{q}$' are contradictories. We can say that they are fixing their degrees of belief in q given p."(Ramsey 1990, p. 155.)

Even today, RT is widely discussed. Some authors criticize it. For example, Fuhrmann and Levi 1994 presents a class of conditionals that are assertable but RT judges them as not assertable and proposes a different version of the test. On the other hand, because of its intuitiveness but also simplicity the procedure served as a direct inspiration for three successful research programs: belief revision theory, possible world semantics and suppositional theories of indicative conditionals, and it is still considered to be the default test for the acceptability of conditionals (see e.g., Bennett 2003, Over and Cruz 2018 or Evans and Over 2004.).

In my article, I will present a new argument against RT. It is based on the idea that one of the conditions of acceptability of conditionals is that the an-

tecedent is positively probabilistically relevant for the consequent. This idea was introduced in Douven 2008. In the article, the author presents the Evidential Support Theory which incorporates the probabilistic relevance as one of the conditions of the acceptability of conditionals and was confirmed by the results of the experiment presented in Douven and Verbrugge 2012. I will show that the RT does not incorporate the probabilistic relevance requirement and therefore it is not an adequate test for the acceptability of conditionals. Then, I will present an improved version incorporating the relevance condition.

In the second section, I will introduce the notion of probabilistic relevance, then I will present the Evidential Support Theory and the Relevance Hypothesis. In the third section, I will discuss alleged counterexamples to the Relevance Hypothesis. In the fourth section, I will present the counterexamples to RT inspired by the Relevance Hypothesis and a version of RT which incorporates the Relevance Hypothesis. Finally, in the last section, I will draw some conclusions.

3.2 Evidential Support Theory and Relevance Hypothesis

Douven and Verbrugge in their Douven and Verbrugge 2012 presented the Evidential Support Theory.[1] The main claim of the theory is the Evidential Support Thesis(EST):

"EST An indicative conditional "If A, B" is assertable/acceptable if and only if $Pr(B|A)$ is not only high but also higher than $Pr(B)$."

The theory is supported by the results of an experiment described in the paper. During the experiment sixty-two participants were presented with 18 items. Each item consists of there questions. The first two are about the probabilistic relation between two events and the last one is about the acceptability of indicative (and concessive) conditionals with these events as antecedents and consequents. The authors compare the answers of the participants with the predictions of EST and the Qualitative Adams' Thesis(QAT):

[1]The theory was first presented in Douven 2008, I will use the version from Douven and Verbrugge 2012.

"(QAT) An indicative conditional "If A, B" is assertable for/acceptable to a person if and only if the person's degree of belief in $Pr(B|A)$ is high."

The results clearly support EST and to a much lesser extent QAT. According to the authors, the results show that QAT identifies only a part of the acceptability conditions of indicative conditionals. The missing part is the condition of positive relevance added in EST. The probabilistic relevance can be conceptualized in at least two ways.

Firstly, we can use $\Delta P = P(B|A) - P(B|\neg A)$ proposed in Spohn 2012. If the value of ΔP is 0 the corresponding conditional is irrelevant, when it is higher then it is positively relevant and when it is lower the conditional is negatively relevant. Secondly, the relevance can be conceptualized as a difference measure $P(B|A) - P(B)$. As we have seen this notion was used by Douven to express the additional requirement of positive relevance in EST. As in the case of ΔP when the value of difference measure is 0 the conditional, is irrelevant, if it is lower it is negatively relevant and if it is higher it is positively relevant. Both conceptualizations classify conditionals in the same way but the exact level of relevance will differ in some cases.[2] Both notions were used in literature and the difference will not matter for our conclusions.

With the notion of probabilistic relevance in hand we may single out the addition made in EST in comparison to QAT:

RH A positive relevance ($\Delta P > 0 \leftrightarrow P(B|A) - P(B) > 0$) is a necessary condition for a indicative conditional $A \rightarrow B$ to be acceptable.

I will call it Relevance Hypothesis (RH). As we have already seen EST is supported by the results of the experiment presented in Douven and Verbrugge 2012. Given the fact that the only difference betwen QAT and EST is that EST incorporates RH it seems that the experiment indirectly support RH. It is also directly supported by the results of other experiments for example, Krzyżanowska, Collins, and Hahn 2017 or Douven et al. 2019.

[2]For a detailed discussion of the difference between the two notions and an experiment indicating that ΔP predicts intuitive relevance better than the difference measure, see Skovgaard-Olsen, Singmann, and Klauer 2017.

3.3 Counterexamples to the Relevance Hypothesis

RH is still controversial despite growing empirical support for it. Two strategies of arguing against it are present in the literature. Firstly, one can look for an example of an acceptable conditional which clearly does not involve a positive relevance. Secondly, one can find an example of valid reasoning from premises that do not involve relevance to conclusions which includes a true conditional.

Let us start with the first strategy. A clear example of two conditionals which was judged to be acceptable but cannot both be positively relevant is provided in Gibbard phenomenon:

> "Sly Pete and Mr. Stone are playing poker on a Mississippi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good and signals its content to Pete. My henchman Jack sees both hands and sees that Pete's hand is rather low so that Stone's is the winning hand. At this point, the room is cleared. A few minutes later, Zack slips me a note which says "If Pete called, he won," and Jack slips me a note which says "If Pete called, he lost." I know that these notes both come from my trusted henchmen, but do not know which of them sent which note. I conclude that Pete folded." (Gibbard 1981, p. 231.)

According to Gibbard's original interpretation, the evidence the narrator has for both conditionals are equally strong. On the other hand they cannot be both true, this would lead together with a plausible conditional non-contradiction rule:

CNC $\neg((A \to \neg B) \wedge (A \to B))$

to contradiction. Therefore he concludes that both conditionals are acceptable and sees the phenomenon as an argument for a popular philosophical position called the non-truth value view which claims that conditionals are not truth-apt (see e.g., Bennett 2003 or Edgington 1995).

The phenomenon is very controversial and alternative interpretations were proposed. Firstly, Lycan 2003 denies that the support for both conditionals is

symmetric and therefore claims that one of them is true, and therefore acceptable, while the second one is false, and therefore not acceptable.

Secondly, following Krzyżanowska, Wenmackers, and Douven 2014 one can claim that the meaning of conditionals depend on the beliefs of the speaker. In the case described by Gibbard, it is clear that both Zack and Jack based their conditionals on different beliefs based on different evidence. Because of that, both conditionals despite their superficial form are not in any tension and therefore are not inconsistent even when combined with CNC. To put it otherwise, according to this interpretation $A \rightarrow B$ and $A \rightarrow \neg B$ are not a proper formalization of the conditionals from the story, they are based on different beliefs and therefore they express different relations.

How that is connected to RH? According to both conceptualization of relevance, it is not possible for $A \rightarrow B$ and $A \rightarrow \neg B$ to be both probabilistically relevant at the same time. Therefore, if both conditionals are acceptable and their logical form is $A \rightarrow B$ and $A \rightarrow \neg B$ we would have a clear counterexample to RH. At the same time, the example is equally problematic for QAT, two sentences of such form cannot have both high conditional probability at the same time. Still, as we have seen the phenomenon is very controversial. The intuition of speakers in such cases was, as far as I know, never tested so it is not clear if the Gibbard's intuitions are generalizable.

A different kind of counterexamples are sentence like:

(1) If it will not rain tomorrow I will go to the beach and if it will rain tomorrow I will go to the beach.[3]

In the case of this and similar conjunctions of conditionals, both conjuncts, are supposed to be acceptable for the person which is sure that the consequent (beach trip) of both conditionals will happen no matter the state of the antecedent (presence or absence of rain). If such conditionals are in fact acceptable they constitute a clear counterexample to RH, both $\neg A \rightarrow B$ and $A \rightarrow B$ cannot be positively relevant at the same time. On the other hand, as in the case of the Gibbard phenomenon, there are no empirical studies that show that the users of the natural language are willing to accept such conjunction. Even if such expressions are systematically acceptable in some contexts it is

[3] A similar counterexample was implied in Stalnaker 1968: "If the Chinese enter the Vietnam conflict, the United States will use nuclear weapons and if the the Chinese will not enter the Vietnam conflict, the United States will use nuclear weapons."

not obvious that they express a conjunction of two indicative conditionals. For example, it may be the case that by means of them speakers express something like:

(2) If it will not rain tomorrow I will go to the beach and even if it will rain tomorrow I will go to the beach.

In such case despite its superficial structure (1)-like utterances are not a conjunctions of two indicative conditionals but conjunction of an indicative conditional and a concessive one (conditional which involve "even if" clause) and therefore are not a counterexamples to RH. Such reading seems to be plausible. To show that the conjunction like (1) are conjunction of indicative conditionals one would have to show that in contexts in which speakers are willing to assert them they are also willing to assert each of the indicative conditionals combined in it on their own. This would show that both $\neg A \rightarrow B$ and $A \rightarrow B$ are acceptable in indicative form at the same time and constitute evidence against RH.

Two other counterexamples to the RH were discussed in Skovgaard-Olsen et al. 2019. The authors discuss counter-examples to the idea that the truth of an indicative conditional requires the existence of a connection between antecedent and consequent. The connection of some kind is necessary for the positive relevance, therefore, a convincing example of acceptable conditional which does not involve any connection will be an example of conditional which does not involve relevance. The first discussed example was originally presented in Johnson-Laird and Byrne 2002:

> "We do not deny that many conditionals are interpreted as conveying a relation between their antecedents and consequents. However, the core meaning alone does not signify any such relation. If it did, then to deny the relation while asserting the conditional would be to contradict oneself. Yet, the next example is not a contradiction: If there was a circle on the board, then there was a triangle on the board, though there was no relation, connection, or constraint, between the two—they merely happened to co-occur." (p. 651)

Skovgaard-Olsen and co-authors claim that Johnson-Laird and Byrne are mistaken in claiming that there is no connection involved in the described example. The co-occurrence mentioned in the quote is a connection that justifies

the utterance of the conditional. That seems to be a convincing response and considering probabilistic relevance makes it even clearer, the correlation between two shapes makes the occurrence of the circle positively relevant for the occurrence of the triangle.

The second counterexample was suggested to Skovgaard-Olsen and his coauthors by an anonymous reviewer:

> "Detective interviewing shopkeeper:
>
> D: We need to know what Mr. Smith bought today, can you help us out?
>
> S: I'm sorry, I didn't find out about any customers' names today.
>
> D: Well, he was carrying a large polka-dotted umbrella.
>
> S: If he carried a polka-dotted umbrella, then he bought a gold watch."

as in the previous case, the authors claim that this example involves a connection. It is established by detective recognizing Mr. Smith as the man who was carrying a polka-dotted umbrella. This addition makes the arguments of the conditional connected in this context and, at the same time, secures the positive relevance.

An example reasoning which is credited with being a counterexample to the is the Conjunctive Sufficiency also called centering:

CS A and $B \models A \rightarrow B$

As we see, CS is an inference that takes us from a conjunction of to the conditionals from one conjunct to another one. Relevance is not required for the truth nor the acceptability of conjunction so if the inference from conjunction to a conditional is valid then relevance cannot be a part of the acceptability condition for conditionals. The CS is validated by most of the popular semantics of indicative conditionals. It is validated by the possible world semantics (see e.g., Stalnaker 1968), three-valued semantics (see e.g., Baratgin et al. 2018 or Egré, Rossi, and Sprenger 2019) or popular suppositional theory (see e.g.,Over and Cruz 2018). On the other hand, some authors regard it to be un-intuitive and defend the semantic or pragmatic theories which do not validate it. The

example of such theory is already discussed EST or promising inferential semantics defended for example in Douven 2015 or Krzyżanowska, Wenmackers, and Douven 2014.

The results of the empirical experiments concerning CS are mixed. Cruz et al. 2016 supports it by showing that the way participants react to instances of CS is more in line with how they typically react to valid rather than invalid inferences. At the same time, the results of Krzyżanowska, Collins, and Hahn 2017 and Douven et al. 2019 goes against the CS.

In light of the above, it seems that the validity of CS is still a controversial issue and therefore it may be premature to reject RH on this ground. At the same time, as far as I know, no conclusive counterexample against it was yet proposed. Therefore, it seems that given it strong empirical standing the relevance hypothesis should be regarded at least as very plausible.

3.4 Relevance Hypothesis and Ramsey Test

What all that has to do with RT? As we have seen, the QAT a probabilistic reformulation of RT does not include the requirement of positive relevance amongst the acceptability conditions of conditionals. Unsurprisingly it the same for RT. It does not include the probabilistic relevance as a condition of acceptability which in light of strong standing of RH seems to be problematic. To see that, consider:

(3) If I eat an apple today, I will not inherit 1000000$ today.

Let us assume that in case of (3) the consequent is very probable and the antecedent is probabilistically irrelevant for consequent.[4] RT will judge (3) as acceptable; we add the antecedent to our stock of beliefs our subjective probability of the consequent will be high. At the same time, EST will not judge it as acceptable because the antecedent is not relevant for the consequent. We can multiply similar examples;[5] it seems that in all of them our intuitions go

[4]Obviously, we can fix the probability of the consequent as high as we want without making the antecedent probabilistically relevant for it.

[5]Similar examples were used in an experiment described in Douven and Verbrugge 2012 or in Skovgaard-Olsen, Singmann, and Klauer 2016. The results of the experiment presented in the second paper suggest that the acceptability of conditionals generally does not correspond to the conditional probability of a consequent given an antecedent.

together with the verdict of EST. This advantage is, as we have seen, confirmed by the results of empirical studies. All that together constitute a powerful argument against RT as a procedure for judging the acceptability of indicative conditionals.

After establishing our negative results two questions remain: what does RT really test? And, what would be a better test for the acceptability of conditionals?

The answer to the first question goes beyond the scope of this article. At the same time, it seems very plausible that RT provides the interpretation for conditional degrees of beliefs as proposed by Edgington 1995 or Sprenger 2015. This interpretation is also supported by the original Ramsey's formulation:

> "We can say that they are fixing their degrees of belief in q given p"

If that is the case and if a high conditional probability of a consequent given an antecedent is not enough for a conditional to be acceptable, as is predicted by EST, it is natural that RT, which tests just a conditional probability, is not a reliable test for acceptability of conditionals.

The answer to the second question is easier. EST suggests a way in which we can upgrade RT to prevent it from accepting irrelevant conditionals. It is enough to add a clause where the subject checks if the acceptance of the antecedent raises the probability of the consequent. The upgraded test looks like this:

RT+

1. Add p hypothetically to your stock of beliefs and update the rest of your beliefs in order to make them consistent with the acceptance of p. Is your subjective probability of q high?
2. Compare your degree of belief in q now and before you added p. Is it higher now?

If the answers to both questions are positive the conditional $p \longrightarrow q$ is acceptable. RT+ preserves the intuitions behind RT and should be treated as an improved version rather than a new test. In any case, RT+ corresponds well to EST. Therefore the presented evidence which supports EST also supports the new version of the test as well.

3.5 Conclusion

In my article, I described a new argument against RT. I argued that a positive probabilistic relevance requirement is one of the conditions of the acceptability of indicative conditionals. It is both supported by empirical evidence and there is no uncontroversial counterexample to it. RT does not incorporate the requirement and therefore is not a successful procedure for judging the acceptability of indicative conditionals. At the same time, it can be easily augmented in a presented way.

Chapter 4

Minimal Theory of Causation and Causal Distinctions

4.1 Introduction

The Minimal Theory of Causation, presented in Graßhoff and May 2001, aspires to be a version of a regularity analysis of causation able to correctly predict our causal intuitions. I my article, I will argue that it is unsuccessful in this respect. The second aim of the paper will be to defend Hitchcock's proposal concerning divisions of causal relations (Hitchcock 2001) against criticism made, in Jakob 2006 on the basis of the Minimal Theory of Causation. In the second section, I will present the Minimal Theory and Jakob's critique. In the third one, I will critically examine both of them. In the last section I will conclude.

4.2 Minimal theory of causation

The Minimal Theory of Causation (MT) is a version of a regularity theory of causation. It was proposed in Graßhoff and May 2001 and it both builds on and is motivated by shortcomings of the Mackie's INUS theory (Mackie 1980). Therefore, I will start by introducing the INUS theory. According to the theory, a cause of an event is insufficient but necessary (non-redundant) part of unnecessary but sufficient condition of this event, in short, its INUS condition.

A classic example used to present the theory is a case of a fire. When experts judge that the short-circuit is the cause of the fire, they are aware that it was just a part of a set of conditions which together are sufficient for the fire. It includes the presence of oxygen, flammable material etc.. At the same time, the short-circuit was necessary in the described situation, without it there would be no fire. The set of conditions is not necessary for the fire. There are other scenarios in which the house would burn, for example, an arsonist with gasoline and a lighter. Therefore short-circuit is insufficient but necessary (non-redundant) part of unnecessary but sufficient condition or a cause of the fire. This is a good prediction. Because such correct predictions the theory is popular among scientist. For example, Marini and Singer 1988 describes how Mackie's theory is used in social science. Similarly, Warr 2016 argues that INUS theory is the most suitable one for criminology because it is simple and easy to use and, at the same time, able to incorporate multidimensional explanations.

On the other hand, the theory is currently not popular among philosophers, especially in comparison to the interventionists theories (e.g., Woodward 2005) or causal models (e.g., Pearl 2009). One reason for that may be notorious counterexamples, for example, the Manchester factory hooters case. The example was formulated by the Mackie himself:

> "The sounding of factory hooters in Manchester [at 5 p.m.] may be regularly followed by, but does not cause, London workers leaving their work."(Mackie 1980, p.81)

In this case because of common cause of both the sound and leaving of workers the sound is recognized as an INUS condition and therefore cause of workers leaving. This is an unintuitive result and it was recognized as a problem by Mackie himself. The Minimal Theory of Causation, proposed in Graßhoff and May 2001, preserves intuitions behind the INUS theory but does not suffers from the traditional problems. In words of Jakob 2006:

> "At first sight, MT bears a close resemblance to Mackie's (1974) INUS-conditions, but a second glance will show that some important lessons have been learned."(Jakob 2006, p.280)

Surprisingly MT did not get much traction since. It clearly possesses many attractive features. Firstly, it shares already mentioned attractive features of

INUS but is not susceptible to the counterexamples. Additionally, it provides a unified theory of singular and general causation.

In light of all that, it is surprising that the theory did not attracted more supporters. Perhaps the problems I will diagnose in my paper can be a part of explanation why INUS is not popular anymore and MT, despite having attractive properties, never get much popularity. In my presentation of MT I will use a version of the theory presented in Jakob 2006. It is constituted by causal principles, which constrain all deterministic causal relations:

> "*Principle of causal determinism*: The same cause is always accompanied by the same effect."

This principle "defines the essence of deterministic causation".[1] On the other hand, it makes the theory unable to deal with probabilistic causal relations. In most cases of the probabilistic causation, the different effects can be caused by a single cause.

> "*Principle of causality*: If no cause is present, no effect occurs."

It claims that whenever an event type has a cause, then if an event from this type occurred, at least one of his causes had to occur. In other words, it prohibits spontaneous occurrences of effects.

> "*Principle of causal relevance*: Every type of cause is indispensable for the occurrence of an effect in at least one situation."

This principle prevents the theory from accepting as cause an event which just happen to coincide with the effect. In order to be classified as a cause by the theory an event has to be necessary for at least one occurrence of a given effect.

> "*Principle of persistent relevance*: An event type maintains its causal relevance when additional event types are taken into account."

The last principle makes possible for the theory to exclude relations which are not persistent enough.

After he presented the principles, Jakob defines a causal relation which obeys them. He starts by defining a *minimally sufficient conjunction:*

[1] Graßhoff and May 2001 p. 88.

> "**MsC**: A conjunctive sufficient condition ϕ of an event type θ is a *minimally sufficient conjunction* of θ, if and only if no proper part of ϕ is sufficient for θ."

where ϕ and θ designate event types. There can be many different minimally sufficient conjunction for one effect. The disjunction of all of them is, because of the Principle of causality, a necessary condition. We called a necessary condition of x a *minimally necessary condition* if and only if none of its proper parts is a necessary condition of x.

Using previously defined notions, he defines a *minimal theory*:

> "A minimally necessary disjunction of minimally sufficient conditions of E is called a *minimal theory* of E."

Then, with use of the intuition expressed in the Principle of persistent relevance he defines causal relevance, which is at the same time his definition of type-level causality:

> "**CR**: An event type C is *causally relevant* for an event type E if and only if (i) C is part of a minimal theory of E, and (ii) C stays part of this minimal theory across any extensions of its frame of event types."

The final notion is a singular causation which was not defined in Graßhoff and May 2001:[2]

> "**SC**: Two events c and e stand in a singular causal relation if and only if c instantiates a positive event type C which, according to **CR**, is causally relevant for an event type E, such that e instantiates E, and c is coincident with other events that instantiate a minimally sufficient condition CX of E, in which C is contained."

Before I move to the critical part of my paper, I will present one of application of MT, Jacob's critique of Hitchcock's distinctions of causal relations. I will start by presenting the distinctions.

Hitchcock 2001 develops two complementary distinctions in order to clarify and explain one way in which we use causal claims, namely as advices.

[2]Were capital Latin letters denote event types, while small letters denote events.

According to him, the distinctions are conflated in a traditional distinction of causal relations, which distinguishes between singular and general causal relations. (S) is an example of a singular causal claim while (G) of a general one:

(S) David's smoking caused him to develop lung cancer.

(G) Smoking causes lung cancer.

The first of Hitchcock's distinctions distinguishes between actual and tendency causation. Actual causal claims imply that the arguments of a relation actually occur. (S) is an example of an actual causal claim while (G) is a causal tendency claim. Another causal tendency claim is:

(T) David is the sort of person for whom smoking tends to cause lung cancer.

The second distinction divides causal claims into narrow and wide. Narrow causal claims describe relations between single events or between homogeneous populations. Examples of such narrow claims are (S) or (T). Wide causal claims describe relations between "broader, more heterogeneous populations"[3] examples of such claims are (G) or:

(W) Every year, there are thousands of new cases of lung cancer that are caused by smoking.

Hitchcock claims that these are conflated in commonly used distinction into general and singular causation and that the new distinctions cannot be expressed in terms of the old one.

Using the framework of MT, Jacob revises Hitchcock's distinctions. He starts with actual/tendency distinction. According to Jakob, the most relevant cases are the relations which are classified as single causation by the standard division and as tendency causation by Hitchcock's distinction, for example (T). According to MT, a claim is a singular causal claim just if it refers to particular events standing in causal relations. (T) does not refers to any such events and therefore is classified as a general causal claim. The meaning of (T) is, according to MT, captured by the following two clauses:

"(i) Smoking S is causally relevant for lung cancer L.

[3]Hitchcock 2001 p. 220.

(ii) David instantiates a conjunction of event types X together with which smoking is a minimally sufficient condition SX, which is part of a minimal theory of L."

The first clause describes existence of the general causal relation between smoking and cancer. The second one states that David instantiates all parts of a minimally sufficient condition of lung cancer except smoking. In other words, the only thing he lacks to get a lung cancer is smoking. Neither of the clauses imply occurrences of events which are arguments of the relation (David's smoking and cancer) and therefore (T) is a general causal claim.

In opposition to (T), (S) is a singular causal claim. It is so because it refers to two particular events, namely David's smoking and developing lung cancer, standing in the causal relation. According to MT its content can be characterized as:

"(i) Smoking S is causally relevant for lung cancer L, and

(ii') David instantiates a minimally sufficient condition SX_i containing smoking S, and SX_i is part of a minimal theory of L."

The first clause of the meaning of (S) is the same as the first clause of the meaning of (T). Both (S) and (T) describe a causal relation between smoking and lung cancer. The second clause of the meaning of (S) states that David instantiates the whole minimal sufficient condition which includes smoking. (W) is also a singular causal claim, the difference between it and (S) is just the number of persons who instantiate a relevant minimally sufficient condition. Finally, (G) is a typical case of general causal claim, its meaning, according to the author, is exhausted by:

"(i) Smoking S is causally relevant for lung cancer L."

(G) does not imply anything about occurrence of any actual events, which makes it a general causal claim. If a claim is general in according to MT it is not an actual claim in Hitchcock's sense. Therefore according to Jakob the actual/tendency distinction does not add anything.

Then Jakob discusses the division into wide and narrow causal claims. Hitchcock classifies (S) as an example of wide causal claims while (W) as a narrow one. According to Jakob's analysis, both claims share (i) as part of

their meaning and differ just in the attribution of smoking-involving minimal necessary condition. (S) attributes it to David while (W) to thousands of cases. Therefore, both sentences differ just in a number of instantiations they describe, not in the nature of the expressed causal relations.

Jakob concludes that his distinction classifies the presented cases just as well as two Hitchcock's distinctions, and for sake of methodological parsimony we should use it.

4.3 Critique

In this section, I will present some problematic features of the Minimal Theory of Causation and Jakob's critique. Firstly, I will raise some general worries concerning MT. Afterwards, I will show how they translate into problems in Jakob's analysis of meaning of causal claims and his reformulation of singular/general causal claim distinction. Lastly, I will present one way in which Hitchcock's distinctions are superior over Jakob's one.

MT does not seem to be a promising theory of causation. One of the purposes of such theory is to predict causal intuitions of natural language speakers. MT fails to deliver in this respect in at least two different ways:

It is too restrictive. Firstly, a causation by omission is excluded from the analysis. Moreover, it seems that we would have to significantly change the theory in order to incorporate omissions. For example Principle of causal determinism seems to be generally false for causation by omission. We can say that a lack of seat belts is a cause of a serious damage in the case of an accident, but it will not have the same effect in different situation. Secondly, as Jakob's himself admits, MT excludes probabilistic causation: "If there are irreducibly probabilistic relations the MT-analysis does not apply to them."[4] The situation seems to be worst than Jacob admits. The theory not only excludes cases of indeterministic causation but also does not give us any answer in cases in which we do not know the deterministic mechanism behind the connection. Arguably, (G) is one of such cases. This seems problematic if we consider uses of causal concepts in contexts in which we have just statistical information.

Some of the axioms of MT are counter-intuitive. For example, the Principle of causal determinism seems to cause the above problem. Similarly, the

[4]Jakob 2006 p. 280.

Principle of causality seems counter-intuitive, consider for example an alternative world which contains rare cases of spontaneous occurrence of lung cancer. Contrary to the Principle of causality, in such alternative world we would still have been entitled to state (G).

Finally, the analysis of meaning of causal claims, based on MT article seems to be unsuccessful. Consider for example (T)-like sentence:

(I) Scarlett Johansson is the sort of person for whom contact with grass pollen tends to cause a skin rash.

According to Jacob's analysis the meaning of (I) is:

(i) Contact with grass pollen P is causally relevant for skin rash R, and

(ii) Scarlett Johansson instantiates a conjunction of event types X together with which contact with grass pollen is a minimally sufficient condition PX, which is part of a minimal theory of R.

Neither of the clauses imply occurrences of events which are the arguments of the causal relation and therefore (I) is a general causal claim. But (I) does not seem to convey nothing about anybody except Scarlett Johansson. Why it is a general causal claim? And why we should accept (i) as a part of its meaning? There is a possible answer, namely that the causal relevance in (i) is a very weak notion and therefore (i) does not adds much. This seems to be plausible if we consider how causal relevance is understood in MT. On the other hand, the answer starts to be very unintuitive if we remind ourselves that the meaning of (G) and (G)-like sentences are exhausted by (i) and (i)-like clauses. It seems obvious that we mean something stronger when we state (G).

The problem is clear when we consider (I) together with its wide[5] equivalent:

(J) Contact with grass pollen causes skin rash.

The meaning of (J) is:

(i) Contact with grass pollen P is causally relevant for skin rash R.

[5]It is wide in Hitchcock's terminology; in Jakob's one both (I) and (J) are general.

If we compare both meanings it is clear that according to Jacob's analysis (I) is stronger then (J). This means that we can infer (J) from (I) but not the other way around. This seems false. (I) is true, but (J) seems false. Therefore the theory seems to be in trouble.

We can ask here, what went wrong? The most obvious answer is that the meaning of general causal claims like (G) or (J) predicted by MT is much to weak. It is imposed by the Principle of persistent relevance and the Principle of causal relevance. Both principles can be satisfied by a single instance and therefore the truth conditions for a general causal claim are easy to satisfy. It is enough that one member in a given population (e.g., Scarlet Johansson in population of humans) instantiates a given causal tendency for the whole population to instantiate it. This is reflected in implausible predictions concerning meanings of causal claims. It seems that we mean something stronger when we use general causal claims.

The Jacob distinction is also unintuitive. A claim is a singular causal claim, according to the MT, if it refer to actual events standing in causal relation. If a clause is not a singular causal claim it is general one. Because of this, claims like (I) which refers to a tendency instantiated by a single instance are classified as a general causal claims. This seems to be at least misleading. (I)-like sentences do not describe anything general so why we should classify them as general causal claim?

Similarly, it seems that Hitchcock's distinctions are more fruitful. To see that let us go back to the aim of the Hichcock's article, namely an analysis of causal claims used as advices. Hitchcock's distinctions can be used to single out the type of causal claims which are the most naturally used in this role – the wide tendency causal claims. At the same time, the Jacob's distinction cannot do that or at last it cannot do that in a similarly elegant way. Plausibly, the subset of general causal claims would be most suitable to serve in this way. At the same time it is not clear if it can be described more precisely. Therefore at least in this respect, Hitchcock's distinctions perform better than Jacob's one.

4.4 Conclusion

In my paper I argued that MT is not a promising theory of causation. As we have seen, the analysis is both too restrictive (it excludes causation by omission

and probabilistic causation) and at the same time not restrictive enough (the predicted meaning of general causal claims is too weak). Moreover, it seems that there is no way in which we can solve these problems without changing the theory in a substantial way.

One interesting question we can ask is: Are the features which cause the implausible predictions of MT present in its predecessor Mackie's INUS theory? As we have seen the main problem with MT is that it is too lenient in the way it classifies factors as general causes. It is enough for a class to be judged as a cause of some other class that one event from the first class is part of minimally sufficient condition for an event from the second class. This seems to be way too lenient criterion and analogical leniency is already present in INUS. As we have seen according to the theory, a cause of an event is insufficient but necessary (non-redundant) part of unnecessary but sufficient condition or this event. To see the analogy we can go back to our example the case of a fire. In this case, the short-circuit is insufficient but necessary (non-redundant) part of unnecessary but sufficient condition or a cause of the fire. The prediction is plausible but let us consider what Mackie is saying about other conditions being part of the sufficient set (in our case presence of oxygen and flammable material etc.):

> "That is, the formula "$AB\bar{C}$ or $D\bar{E}F$ or $\bar{G}\bar{H}I$ or . . ." represents a necessary and sufficient condition for the fire, each of its disjuncts, such as '$AB\bar{C}$', represents a minimal sufficient condition, and each conjunct in each minimal sufficient condition, such as 'A' represents an INUS condition."(Mackie 1980 p. 246)

Surprisingly, Mackie is writing here that all other individual conditions which are parts of all sufficient but unnecessary conditions are INUS conditions and therefore causes. This is puzzling, there seems to be too many causes. First of all, the presence of oxygen and all other implicit conditions are all causes of the fire. Secondly the quote seems to suggests that conditions which are members of unrealized unnecessary sufficient sets are also causes. This is perhaps an uncharitable interpretation. In any case, the attribution of causality to factors other then short-circuit seems to be less intuitive. Moreover, as far as I can tell, the theory does not give us any way to explain why short-circuit seems to be more important as a cause of fire than any other INUS

condition. To put it otherwise, INUS theory is not able to distinguish the cause from the contributory causes. This problem is analogical to the problem with MT. According to INUS theory it is enough for a singular causal claim that a given event is a insufficient and necessary part of any of many sufficient but unnecessary condition, no matter how unlikely or exotic this condition is. Similarly, according to MT theory, it is enough that a one member in a given population (e.g., Scarlet Johansson in a population of humans) instantiates a given causal tendency for the whole population to instantiate it. The problem was already present in INUS theory but going from singular to general causal claims seems to make it worst.

As we have seen, contrary to the main claim of Jakob's article, it seems that Hitchcock's distinctions perform better then the new distinction defined on the basis of MT. The new distinction classifies some of the claims in a misleading way for example, it classifies (I) as a general causal claim.

Finally, we can ask if there is a part of Jakob's critique which remains valid? I think that there is. In the last part of his article Jakob points out that the second of two Hitchcock's distinctions, the one which distinguishes between wide and narrow causal claims, does not describe two semantically different kinds:

"The essential qua necessary difference between (S) and (W) is the mere number of instantiations which is one in (S) and thousands in (W). This difference has nothing to do with the nature of the causal relations expressed in (S) and (W)..."[6]

This seems true, the first distinction does all the semantic work. It seems plausible that many instances of relations analogical to the one which makes (S) true would make (W) true. Plausibly, the difference between wide and narrow causal claims is no more semantically significant than the difference between "wide" and "narrow" categorical claims like:

(L) David have lung cancer.

(M) There are thousands of cases of lung cancer.

On the other hand, this does not make the second distinction unimportant. As we have seen it has both epistemic and pragmatic importance.

[6]Jakob 2006 page 286.

Chapter 5

Causal Conditionals, Tendency Causal Claims and Statistical Relevance

5.1 Introduction and Theoretical Background

Indicative conditionals—that is, conditionals that do not involve the auxiliary verb "would" and do not imply falsity of the antecedent—are important linguistic structures. We use them to predict and explain events, to formulate instructions, and to describe causal relationships. For example, we can describe a causal tendency using explicit causal wording:

(1a) A lot of rain causes the ground to be waterlogged.

or by means of an indicative conditional

(1b) If it rains a lot, then the ground will be waterlogged.

This paper presents an empirical study of **causal conditionals**: that is, indicative conditionals that express a causal relationship. From now on, we will refer to them simply as "conditionals". This focus excludes, for example, Dutchman conditionals and purely inferential conditionals.

Studying the connection between conditionals and causal claims is interesting for many reasons. If there is a link, we could develop a unified semantics

for both types of expressions. Indeed, empirical evidence shows that indicative conditionals such as (1b) are often paraphrased by causal claims such as (1a) (e.g., Frosch and Byrne 2012). However, in order to formulate a plausible hypothesis between both types of sentences, we have to distinguish between the causal claims which are related to conditionals and those that are not.

A suitable distinction has been proposed by Christopher Hitchcock (2001, 219–220). According to the author, we can distinguish two kinds of causal claims, which have related but different meanings. The verb "to cause" used in actual causal claims is a success verb: it implies that the events described by its arguments took place. Because of this feature, actual causal claims are used for explaining occurrences of particular events, for example:

(2) James Dean's recklessness caused his accident.

There does not seem to be a systematic connection between the truth of an actual causal claim and the truth of the corresponding (indicative) conditional. The success aspect of an actual causal claim—that both cause and effect occurred—cannot be expressed adequately by means of an indicative conditional, implying that the truth conditions for both types of sentences should be different, too.

Therefore we focus on **tendency causal claims**, which describe the general tendency of a cause to bring about an effect, without implying the actual existence of the causal relata. An example of such a claim is (1a) or:

(3a) Pressing the red button causes the fire alarm to go off.

The corresponding conditional is (example taken from Declerck and Reed 2012):

(3b) If you press the red button, the fire alarm goes off.

True conditionals seem to correspond systematically to true tendency causal claims. For instance, if (1a) or (3a) are true, the corresponding conditionals (1b) and (3b) should also be true. Examining this relationship in detail will be a major goal of our paper. From now on, we use the term "causal claims" as a shortcut for tendency causal claims.

In the above cases, the link between the cause and effect was quite strong. Some tendency causal claims express a weaker relationship: the cause raises

the probability of the effect, but the effect may not be likely even in the presence of the cause. An example of such case is:

(4a) Smoking causes cancer.

We would classify this sentence as true, but the corresponding conditional:

(4b) If one smokes, one will get cancer.

seems false.

What is the difference between a true tendency causal claim, which supports the corresponding conditional, like (1a) or (3a), and those that do not, like (4a)? A plausible explanation is the difference in **strength of the causal link** between antecedent and consequent. Causal strength is typically interpreted as the degree to which the cause C brings about the effect E. While pressing the red button triggers the fire alarm for sure, smoking increases your risk of developing cancer—but not to a level where it is more likely than not. Intuitively, fire alarm is stronger determined by pressing the button than presence of lung cancer by smoking.

To express the concept of causal strength and causal relevance, probability enters the game. Probabilistic theories of causation classify C as a cause of E if and only if C raises the probability of E in all/some/the relevant contexts (Cartwright 1979; Eells 1991; Suppes 1970). The probability-raising intuition is also present in interventionist accounts of causation (Pearl 2000; Spirtes, Glymour, and Scheines 2000). All this suggests that **statistical relevance between cause and effect** could be a good predictor of whether causal claims are classified as true or false. In particular, a causal claim is more likely to be classified as true the stronger the statistical association is (for discussions of probabilistic measures of causal strength, see Cheng 1997; Eells 1991; Fitelson and Hitchcock 2011; Pearl 2001; Sprenger 2018).

The second possible explanation of the difference between causal claims that support the corresponding conditional and those that do not is the different role of **conditional probability**. Suppositional theories of conditionals (e.g., Adams 1975; Edgington 1995; Ramsey 1926) claim that "If C, then E" *expresses* the subjective degree of belief in E given C. For tendency causal claims, no comparable thesis has been advanced. Conditional probability is also frequently used for predicting judgments on causal conditionals. For example,

Ernest Adams (1965, 1975) explicates the acceptability of a conditional by the formula

$$Acc(C \rightarrow E) = p(E|C) \qquad \text{(Adams' Thesis)}$$

Similarly, Stalnaker (1968, 1975) develops a propositional semantics of conditionals where they have a definite truth value and their probability corresponds to that very same conditional probability

$$p(C \rightarrow E) = p(E|C) \qquad \text{(Stalnaker's Thesis)}$$

suggesting that the subjective degree of belief $p(E|C)$ could be an adequate predictor of how causal conditionals are classified (for well-known theoretical problems with Stalnaker's Thesis, see Lewis 1976). Both Adams' and Stalnaker's Thesis are widely supported by patterns observed in natural language reasoning (Adams 1975; Douven and Verbrugge 2013; Egré and Cozic 2011; Evans et al. 2007; Over 2016; Over et al. 2007). On the other hand, recent papers point out that the connection between conditionals and conditional probability crumbles when the antecedent is not (statistically, causally, or otherwise) relevant for the consequent (e.g., Douven and Verbrugge 2010; Skovgaard-Olsen, Singmann, and Klauer 2016). Alternative accounts such as the Evidential Support Theory (EST, Douven 2008, 2016) therefore demand not only that the conditional probability of E given C be high, but also that C raise the probability of E. In other words, C needs to provide evidential support for E (see also Crupi and Iacona 2019b; Krzyzanowska 2015; Krzyżanowska, Collins, and Hahn 2017).[1]

In this paper, we study the commonalities and differences between classifying causal and conditional claims as true or false from an empirical point of view, especially with respect to the role of probability and statistical relevance in predicting these classifications. More precisely, we investigate whether statistical relevance measures affect causal or conditional claims more strongly. There are many experiments devoted both to conditionals and to causal claims (e.g., Frosch and Byrne 2012; Sloman and Lagnado 2015, —see also Douven 2016 for a survey) but, as far as we know, there are no experiments devoted directly to testing the relation between both kinds of expressions. This paper

[1]The evidential support theory can be maintained either for truth conditions or for acceptability conditions; both varieties are to be taken seriously.

aims to fill this gap and to obtain a comprehensive picture of the factors that predict the classification of causal conditionals. In the next section we formulate the hypotheses that we test in our experiment.

5.2 The Hypotheses

The baseline idea of our paper is that causal and probabilistic factors predict the judgment of a causal conditional as true or false. While this claim is in agreement with most of the theoretical and empirical literature, it is too vague to be tested experimentally.

Our first hypothesis concerns the logical relationship between causal conditionals and the corresponding tendency causal claims. Two different kinds of relation are possible:

H1.a (Necessity) Causal conditionals are classified as true *only if* the corresponding tendency causal claim is classified as true.

H1.b (Sufficiency) Causal conditionals are classified as true *if* the corresponding tendency causal claim is classified as true.

H1.a seems prima facie plausible since absence of a causal relation has undermined the acceptability of conditionals in previous empirical research (e.g., Skovgaard-Olsen, Singmann, and Klauer 2016). On the other hand, in the light of the above examples (e.g., the pair (4a)/(4b)), we would expect that H1.b fails in empirical investigation. Some causal conditionals are expected to be classified as false although the corresponding tendency causal claim appears true (e.g., "smoking causes cancer"). We operationalize these hypotheses by demanding that of all causal conditionals evaluated as true, only a small percentage of the corresponding tendency causal claims are evaluated as false (H1.a). Similarly, for an overwhelming percentage of all tendency causal claims evaluated as true, the same claim in conditional form needs to be evaluated as true (H1.b). For the respective thresholds we consider a strict interpretation (5 and 95%) and a lenient interpretation (10 and 90%); as we will see later, this difference in interpretation does not affect the conclusions we draw.[2]

[2]It would also be interesting to test a quantitative version of H1.a and H1.b: the probability of classifying the conditional claim as true increases with the perceived strength of the causal tendency.

Similarly, we would like to test whether statistical relevance measures predict the evaluation of a causal claim as true or false. We focus on measures that depend on $p(E|C)$ and $p(E)$ since these quantities are the easiest to elicit and form a natural class of statistical relevance measures (i.e., they all conform to the Final Probability Incrementality condition, Crupi 2015; Sprenger and Hartmann 2019, ch. 1).[3] Among them, we consider the measures that are most frequently discussed and defended in the vast literature on probabilistic measures of evidential support (e.g., Crupi, Tentori, and González 2007; Fitelson 2001):

$$
\begin{aligned}
&d(C,E) = p(E|C) - p(E) \quad l(C,E) = \log \frac{p(E|C)}{1-p(E|C)} - \log \frac{p(E)}{1-p(E)} \\
&r(C,E) = \log \frac{p(E|C)}{p(E)} \qquad z(C,E) = \begin{cases} \frac{p(E|C)-p(E)}{1-p(E)} & \text{if } p(E|C) \geq p(E) \\ \frac{p(E|C)-p(E)}{p(E)} & \text{if } p(E|C) < p(E) \end{cases}
\end{aligned}
\tag{SR}
$$

The logarithmic transformation maps r and l to a scale that is adequate for predictor variables in a Generalized Linear Mixed Model (GLMM). All measures can be motivated intuitively: $d(C,E)$ and $r(C,E)$ are natural indicators of how much supposing C increases the probability of E on an additive and a multiplicative scale. The measure $l(C,E)$ corresponds to the ratio between posterior and prior odds on E when C is learned in the meantime (i.e., the Bayes factor). Finally, $z(C,E)$ can be conceptualized as a generalization of logical entailment between C and E to the case of uncertain reasoning. We can now formulate our hypothesis:

H2.a Statistical relevance measures in the class (SR) predict the classification of a tendency causal claim as true or false.

As a criterion for evaluating H2.a, we adopt the statistical significance of including statistical relevance as a predictor variable, plus a reasonable, non-negligible effect size. Effect size is measured by how much variance in the data can be explained by the predictor variables and expressed numerically by the squared correlation coefficient R^2. For an effect size to be meaningful, we

[3] A measure depending on $p(E|C)$ and $p(E|\neg C)$ like the popular measure $\Delta p = p(E|C) - p(E|\neg C)$ would require the participant to make two conflicting suppositions ("Suppose $C/\neg C$ is the case. How likely do you consider E?"). This is something we would like to avoid in this experiment, and therefore we focus on measures that depend on $p(E|C)$ and $p(E)$.

demand that it exceed the value $R^2 = 0.09$, which is conventionally identified with the lower bound of a medium effect (Cohen 1988).

Then, we ask which of the two Causal Bayesian Networks in Figure 5.1 is more adequate: the picture where probabilistic factors predict causal claims only via statistical relevance, or the picture where they add predictive value on top of statistical relevance. See Figure 5.1. In other words, the question is whether we can "compress" the two probabilities into a single variable for the purpose of predicting the classification of causal claims. This leads us to our next hypothesis:

H2.b Statistical relevance measures in the class (SR) predict the classification of a tendency causal claim as true or false almost as well as the statistical relevance taken together with conditional probability and probability of the consequent.

As a criterion for whether statistical relevance is an adequate proxy, we demand that the explained variance does not meaningfully increase over and above that what is already explained by the model that contains just statistical relevance as a predictor variable. Similar to H2.a, the difference should not exceed $R^2 = 0.09$.

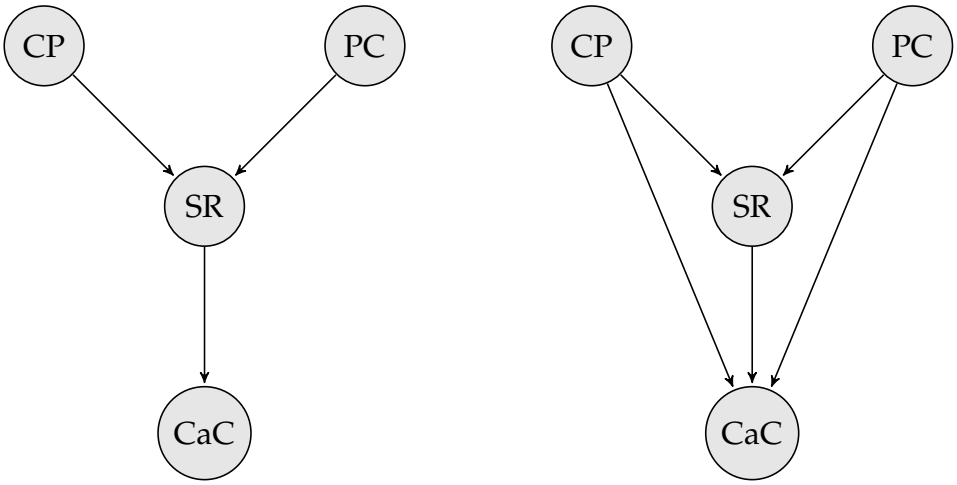

Figure 5.1: Two possible models for predicting the classification of causal claims, represented as Causal Bayesian Networks. CP = conditional probability, PC = probability of consequent, SR = statistical relevance, CaC = causal claim classification.

Third, while there is substantial debate about the empirical adequacy of Stalnaker's Thesis and Adams' Thesis (Douven 2016; Douven and Verbrugge

2013; Skovgaard-Olsen, Singmann, and Klauer 2016), even critics agree on a weak version of those theses: for relevant subclasses of causal conditionals, $p(E|C)$ is moderately correlated with the classification of a conditional as true. We should therefore expect that this probability predicts the classification of a causal conditional at least to some degree.

H3 The conditional probability $p(E|C)$ predicts the classification of causal conditionals of the form "if C, then E" as true or false.

H3 is evaluated on the same basis as the H2.a: adding the conditional probability as a predictor variable needs to be statistically significant, and the effect size as measured by the correlation coefficient must exceed $R^2 = .09$.

Finally, we come to our most interesting hypothesis which posits a specific difference between causal and conditional claims. The intuitive idea is that a conditional is classified as true if and only if two conditions are satisfied: (1) the causal claim is classified as true, in agreement with H1.a; (2) the probability of the consequent, given the antecedent, $p(E|C)$, is "high enough". As explained in the introduction, this hypothesis is supported by research in the tradition of Stalnaker (1968) and Adams (1975) where conditionals are evaluated according to the relevant conditional probability.

H4.a In the class of tendency causal claims classified as true, the conditional probability $p(E|C)$ predicts the evaluation of conditional claims as true.

H4.b In the class of tendency causal claims classified as true, the statistical relevance measures in (SR) predict the evaluation of conditional claims as true.

We expect a positive effect in the first hypothesis and a null effect in the second hypothesis: the effect of statistical relevance on the classification of a conditional is "absorbed" by the evaluation of the causal claim as true. In a Causal Bayesian Network, this would correspond to the claim that the truth of the causal claim screens off statistical relevance from the truth of the conditional claim. In this way, we explain the limitations that probabilistic theories (e.g., Adams' Thesis) experience when causal relevance between antecedent and consequent is suspended. The criteria evaluating for H4.a and H4.b correspond to those for H2.a and H3: statistical significance plus meaningful effect $R^2 > .09$.

Our overall picture amounts to modifying Evidential Support Theory as follows: where EST demands, on top of high conditional probability, statistical relevance between C and E for assessing a conditional as true, our account demands in addition the presence of a causal link between C and E. Of course, this need not be in tension with EST since we conjecture that statistical relevance is actually a good predictor of classifying a causal tendency claim as true.

5.3 Experimental Design and Methods

Participants

Participants were recruited via Amazon's Mechanical Turk (`www.mturk.com`). Mechanical Turk directed the participants to the experiment that was run on the Qualtrics platform (`www.qualtrics.com`). In return for their participation, subjects received a small monetary compensation. Seventy-four native English speakers participated in the experiment. Eighteen participants were excluded because they failed to give the correct response to at least one of the control questions. All participants indicated to have participated seriously. At the end of the experiment, participants were asked an open question about what they thought the experiment was about. None of the participants displayed clear knowledge of the purpose of the experiment. In total, 56 [39 female, mean age = 40.59 years, s.d. = 11.69 years] participants were included in the analysis.

Design

We used a within-subjects design were each participant evaluated 19 scenarios. These scenarios were presented in random order on the participants' computer screen. The participants were instructed to answer questions with the requirement that each question needed to be answered to be able to progress to the next item (i.e., forced-choice). The entire experiment was conducted in English.

Control questions were used as a check on participants' attention and participation. Randomly dispersed throughout the experiment, participants had to give a correct answer to several repeats of the *elimination questions* "For quality control, please select answer category five. If you do not select five, the survey will be terminated." In addition, subjects had to rate on a five-point

Likert scale how seriously they participated in the experiment (1= "completely unserious", 5 = "completely serious"). We included an open question on the purpose of the experiment, because knowledge of the experiment could influence the behaviour of participants. Our exclusion criteria were: failure to give the answer 5 on one of the elimination questions; rating their seriousness in participation as 3 or lower; or describing the experiment as being about conditionals and causation or something similar.

Material and Procedure

Participants had to evaluate tendency causal claims and the corresponding conditional claims in hypothetical scenarios, where a certain situation was described (e.g., the effect of prohibiting alcohol consumption on the crime rate). Through several pre-studies, 19 scenarios were selected on comprehensibility from a list of 60. When directed from Mechanical Turk to Qualtrics, participants first received instructions on the experiment. After the instructions they were presented with the scenarios in a randomized sequence. The 19 scenarios consisted of four questions each, eliciting probability judgments as well as dichotomous judgments on causal and conditional claims:

(Unconditional) Probability of Consequent This question elicits the probability of a certain development without making specific assumptions (e.g., "how likely is it that the crime rate will decline in the next five years?").

Conditional Probability of the Consequent This question elicits the probability of the same development under a specific assumption stated in the antecedent (e.g., "how likely is it that the crime rate will decline in the next five years if alcohol consumption is made illegal?").

Causal Claim This question asks the participants to evaluate the truth or falsity of the causal connection between antecedent and consequent (e.g., "Making alcohol consumption illegal will cause the crime rate to decline in the next five years").

Conditional Claim This question asks the participants to evaluate the truth or falsity of the corresponding indicative conditional (e.g., "If alcohol consumption is made illegal, then the crime rate will decline in the next five years.")

The first two questions had to be answered on a visual analog scale of probability percentages from 0% to 100%. The third and fourth question had to be answered with either "true" or "false". We reproduce the entire experimental material in Appendix A and B.

5.4 Results

Hypotheses H1.a and H1.b

Combined, the data consisted of 1064 entries; 56 participants responded to 19 scenarios. We evaluated H1.a and H1.b by simply checking the frequency statistics for the relevant categories (see Table 5.4). Of the 531 data points where a causal conditional was classified as true, 25 data points classified the corresponding tendency causal claim as false. This corresponds to a percentage of 4,71% and therefore confirms our hypothesis H1.a that perceived presence of a causal relationship is *necessary* for classifying a causal conditional as true, both for the strict 5% and the lenient 10% threshold. By contrast, Hypothesis H1.b that classifying a causal conditional as true is a *sufficient* condition for classifying the causal conditional as true was not borne out by the data: of 611 data points where the tendency causal claim was evaluated as true, only 506 evaluated the corresponding conditional as true. This percentage of 82,82% is clearly below the thresholds of 90% and 95% necessary to establish sufficiency.

Hypotheses H2.a and H2.b

A Generalized Linear Mixed Model (GLMM) was used to test hypothesis H2.a and H2.b. We used a logit link function, as the outcome variable for each hypothesis was binary (0 = False, 1 = True). We added participants and scenario number as crossed random effects, because of difference in content between

Contingency Table		Conditional Claim		Total
		True	False	
Causal Claim	True	506	105	611
	False	25	428	453
Total		531	533	1064

Table 5.1: Classification of causal and conditional claims as true and false.

the scenarios, and possible differences in their interpretation between participants. We used the R package *lme4* (Bates et al. 2015) to estimate the GLMM's regression coefficients, variance components, and the amount of variance in the outcome explained by the predictor(s).

All statistical relevance measures predict the classification of tendency causal claims as true or false, thus supporting hypothesis H2.a (see Table 5.2). For all statistical relevance measures, the analyses show a positive and meaningful association ($R^2 > 0.09$) with the proclivity of participants to assess the tendency causal claim as true. Specifically, the coefficients indicate the estimated increase in log odds (per unit of the relevance measure) of tendency causal claims being indicated as true versus false.[4] The random-effects for scenario and participant, though non-zero, appear to be minor. Specifically, their coefficients are only slightly larger than their standard errors. Based on the test statistics (z-values) and amount of explained variance (R^2), the association between $z(C,E)$ and classification of tendency causal claims was the strongest (i.e., largest coefficient with respect to its standard error). The weakest association was with the $r(C,E)$ measure.

To test hypothesis H2.b, the change in R^2 is assessed when *probability of the consequent* $p(E)$ and *conditional probability* $p(E|C)$ are added to the models of H2.a. Unfortunately, these predictors cannot be added to the $d(C,E)$ model, because they perfectly define, without transformation, this statistical relevance measure (see the SR equations). In other words, these predictors are incapable of explaining additional variance in peoples tendency to indicate causal claims as true over and above $d(C,E)$. For the other three statistical relevance measures, only $z(C,E)$ is an adequate proxy for the effects of *probability of the consequent* and *conditional probability*. Specifically, when the two probability predictors were added to the models the R^2s increased to about 0.47 for all statistical relevance measures, meaning that only $z(C,E)$ is a good proxy for the probability variables (i.e., R^2 difference is below the 0.09 threshold; $0.47 - 0.40 = 0.07$).

[4]Please note that the statistical relevance measures are not measured on the same scale. Thus the coefficients indicated in the tables cannot be meaningfully compared in the sense that the largest coefficient is the best predictor.

Type of Effect	Variable	Coefficient	Std. Error	z	p
Fixed effects	intercept	-0.15	0.31	-0.48	0.63
	$d(C,E)$	5.85	0.53	11.04	<0.0001
Random effects	Participant	1.14	1.07		
	Scenario	1.26	1.12		
Explained variance: $R^2 = 0.34$. Residual degrees of freedom = 1060.					
Fixed effects	intercept	0.26	0.35	0.75	0.46
	$r(C,E)$	1.16	0.14	8.36	<0.0001
Random effects	Participant	1.02	1.01		
	Scenario	1.76	1.33		
Explained variance: $R^2 = 0.22$. Residual degrees of freedom = 1060.					
Fixed effects	intercept	-0.04	0.32	-0.12	0.90
	$l(C,E)$	0.84	0.08	10.08	<0.0001
Random effects	Participant	1.10	1.05		
	Scenario	1.33	1.15		
Explained variance: $R^2 = 0.34$. Residual degrees of freedom = 1060.					
Fixed effects	intercept	-0.24	0.29	-0.84	0.40
	$z(C,E)$	3.74	0.30	12.42	<0.0001
Random effects	Participant	1.10	1.04		
	Scenario	0.99	0.99		
Explained variance: $R^2 = 0.40$. Residual degrees of freedom = 1060.					

Table 5.2: The Generalized Linear Mixed Model (GLMM) for the dependent variable *Causal Claim* as a function of *Statistical Relevance*, quantified by various measures.

Hypothesis H3

To test H3, a GLMM similar to the test of H2 was used. Again, we used a logit link function and added participants and scenario number as crossed random effects. The results show a strong and positive association between the conditional probability ascribed to the scenarios and the log odds of the corresponding causal conditional being considered as true. Specifically, with every percentage-point increase in conditional probability these log odds are estimated to increase by 0.07 (an increase of 7 over the full 100 percentage points). Most importantly, the model explains 32% ($R^2 = 0.32$) of the variance in the participants tendency to indicate causal conditionals as either true or false. In short, H3 is supported by the observed data.

Hypotheses H4.a and H4.b

To test hypotheses H4.a and H4.b, only those data points were used where the tendency causal claim was indicated as true (see Table 5.4). For analyses H4.a.

Type of Effect	Variable	Coefficient	Std. Error	z	p
Fixed effects	intercept	-4.61	0.46	-10.13	<0.0001
	Conditional Probability	0.076	0.006	13.06	<0.0001
Random effects	Participant	1.84	1.36		
	Scenario	0.64	0.80		
Explained variance: $R^2 = 0.34$. Residual degrees of freedom = 1060.					

Table 5.3: The Generalized Linear Mixed Model (GLMM) for the dependent variable *Causal Claim* as a function of *Conditional Probability*.

and H4.b. we again used GLMMs to estimate the coefficients and the variance explained by the predictors. Similar to the results for H3, results show a positive association between the conditional probability ascribed to the scenarios and the log odds of corresponding causal conditional being considered as true (see Table 5.4). Although the amount of variance explained is greatly reduced, from 32% to 12%, it is still considered meaningful ($R^2 > 0.09$) and thus supporting H4.a. None of the statistical relevance measures made a meaningful difference in explaining the variance in the participants' tendency to indicate the causal conditionals as true of false ($R^2 < 0.09$ in all cases, see Table 5.5), thus supporting our conjecture of a null effect and contradicting hypothesis H4.b.

Type of Effect	Variable	Coefficient	Std. Error	z	p
Fixed effects	intercept	0.89	0.62	-1.14	0.15
	Conditional Probability	0.05	0.008	5.79	<0.0001
Random effects	Participant	3.39	1.84		
	Scenario	0.45	0.67		
Explained variance: $R^2 = 0.13$. Residual degrees of freedom = 607.					

Table 5.4: The Generalized Linear Mixed Model (GLMM) for the dependent variable *Conditional Claim* as a function of *Conditional Probability* when *Causal Claim = "true"*.

Type of Effect	Variable	Coefficient	Std. Error	z	p
Fixed effects	intercept	1.89	0.38	4.99	<0.0001
	$d(C,E)$	2.04	0.62	3.28	0.001
Random effects	Participant	2.80	1.67		
	Scenario	0.64	0.80		
Explained variance: $R^2 = 0.04$. Residual degrees of freedom = 607.					
Fixed effects	intercept	2.15	0.37	5.82	<0.0001
	$r(C,E)$	0.36	0.16	2.17	0.030
Random effects	Participant	2.72	1.65		
	Scenario	0.65	0.81		
Explained variance: $R^2 = 0.02$. Residual degrees of freedom = 607.					
Fixed effects	intercept	1.89	0.38	4.99	<0.0001
	$l(C,E)$	0.37	0.09	4.00	0.0001
Random effects	Participant	2.94	1.71		
	Scenario	0.64	0.80		
Explained variance: $R^2 = 0.07$. Residual degrees of freedom = 607.					
Fixed effects	intercept	1.63	0.38	4.31	<0.0001
	$z(C,E)$	1.73	0.41	4.26	<0.0001
Random effects	Participant	2.70	1.64		
	Scenario	0.60	0.78		
Explained variance: $R^2 = 0.07$. Residual degrees of freedom = 607.					

Table 5.5: The Generalized Linear Mixed Model (GLMM) for the dependent variable *Conditional Claim* as a function of *Statistical Relevance* (quantified by various measures) when *Causal Claim* = "*true*".

5.5 Evaluation and Discussion

There is a natural mapping between tendency causal claims ("Smoking weed causes dizziness") and causal conditionals in the indicative mood ("if somebody smokes weed, he will feel dizzy"). In the presented study, we tested various hypotheses about the classification of such sentences as true or false, especially with respect to predicting these classifications as a function of statistical relevance. This is a highly relevant research question since the influence of probabilistic factors on causal claims and causal conditionals has been studied extensively, but in separate literatures. Unlike for counterfactuals, we do not yet have a semantic or pragmatic theory connecting both types of claims. We therefore conducted a study where participants classified a given causal claim and the corresponding conditional as true or false, and estimated in addition two probabilistic variables: the conditional probability of the consequent, given the antecedent, and the probability of the consequent simpliciter. Our specific interest was in finding whether these probabilistic features could reliably pre-

dict the classification of causal claims/causal conditionals as true or false.

Our informal discussion at the beginning of this paper has suggested that the conditions for classifying a causal conditional as true will be more demanding than the conditions for classifying the corresponding tendency causal claim as true. This claim, expressed in hypotheses H1.a and H1.b, has been supported convincingly by our experimental results. Then, we built on probabilistic theories of causation and causal strength in order to formulate our second pair of hypotheses: classification of a causal claim as true or false can be reliably predicted by measures of statistical relevance (H2.a). More precisely, statistical relevance measures can act as a good proxy for the full probabilistic data in a Generalized Linear Mixed Model (GLMM): predicting the classification of tendency causal claims by statistical relevance explains almost as much of the variance in the data as using all available probabilistic predictors (H2.b).

These hypotheses have been confirmed by our data, but H2.b has to be qualified since the outcome depends on the choice of the statistical relevance measure. Only the z-measure—a generalization of logical entailment to uncertain reasoning—meets our prespecified effect size threshold. This phenomenon is a classical case of *measure sensitivity*: the theoretical and empirical properties of statistical relevance, incremental confirmation or evidential support depend on the particular measure chosen (Fitelson 1999, 2001). We can only speculate why z fares better than l, and clearly better than d and r in our case. One plausible explanation would be that the z-measure normalizes statistical relevance, as measured by d, by *maximal* statistical relevance—that is, raising the probability of E to unity. This means that the same probability difference counts more if the conditional probability is high than when it is low. This feature would match with intuitions that *ceteris paribus*, causes are stronger if they are likely to bring about an effect (e.g., as expressed in regularity theories or necessitarian theories of causation). Apart from its contribution to studying causal and conditional judgments, this paper thus adds to an already existing literature that singles out z as a particularly apt measure of statistical relevance—both on the basis of axiomatic representation theorems and empirical performance (Crupi and Tentori 2013; Crupi, Tentori, and González 2007).

For the classification of the causal conditional, we have established that the conditional probability is indeed a reliable predictor, in line with a weak version of Adams' Thesis (H3). The effect size in the GLMM is very remarkable

($R^2 = 0.32$). This relationship remains intact when restricted to the set of causal claims classified as true (H4.a) although the effect size decreases to $R^2 = 0.12$, showing that the classification of the causal claim already predicts a good part of the classification of the causal conditional.

However, statistical relevance is not any more a relevant predictor if the corresponding causal claim is classified as true (H4.b): the relationship between the various measures and the target variable is still statistically significant, but this is to be expected for such a large data set and the effect size is too small to be of theoretical interest ($R^2 < 0.09$ for all statistical relevance measures).

If we extract a theory of how reasoners classify causal conditionals, our results suggest that causal conditionals are supported by sufficiently strong causal claims: (a) the corresponding tendency causal claim needs to be classified as true and (b) the odds for classifying the conditional as true co-vary with the conditional probability $p(E|C)$. Statistical relevance can act as a proxy for the first component (especially when the z-measure is used), but not for the classification of the conditional. Figure 5.2 gives a schematic representation of the relationship between the different variables in our model, adding a latent variable that expresses the part of the causal classification judgment that cannot be reduced to statistical relevance, or other probabilistic predictors.

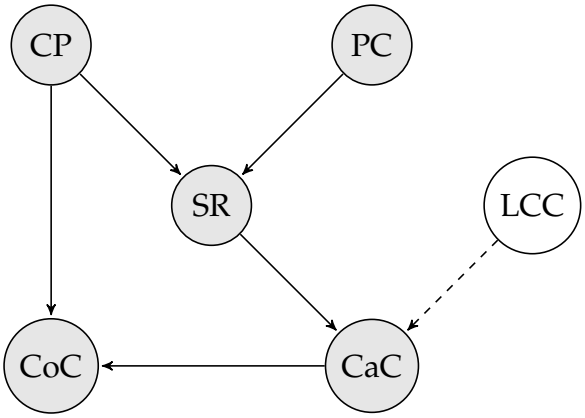

Figure 5.2: Predictors of the classification of a conditional claim (CoC) as true or false, represented via a Causal Bayesian Network. CP = conditional probability, PC = probability of consequent, SR = statistical relevance, CaC = causal claim classification, LCC = latent causal component.

All in all, this account resembles Igor Douven's Evidential Support Theory (Douven 2008, 2016; Douven and Verbrugge 2012) and more recent develop-

ments along these lines (Crupi and Iacona 2019b), and it provides empirical support for these research programs. Our account is different in that the explicit consideration of a causal component adds a "semantic" component which is absent in those proposals.

One of the limitations of our study is the exclusion of counterfactual conditionals, whose causal character has been studied extensively in the literature (Lewis 1973a, 1973b; Pearl 2000; Schulz 2017). We conjecture that analogical relations might hold between counterfactual conditionals and counterfactual causal claims. Consider, for example, the sentences "If Ben had gone to the party, he would have failed his exam" and "Attending the party would have caused him to fail his exam". The relationship between such pairs of sentences strikes us as a valuable object for further research (see also Schulz 2011). In addition, we might study whether the results change when we ask when we ask for the *assertability* of conditionals instead of truth (reasons for suspecting invariance are given by Douven and Krzyzanowska 2018), or when we replace a dichotomous choice by a continuous scale. Finally, we might elicit the probability of the effect given the negation of the cause (i.e., $p(E|\neg C)$) and see whether statistical relevance measures that depend on this quantity lead to different qualitative results. All in all, the interface of conditionals, causality and statistical relevance emerges as an important and fruitful area for future research.

Part II

Values, Objectivity and Replicability

Chapter 6

Values, Bias and Replicability

6.1 Introduction

The value-free ideal of science(VFI) is a view which claims that scientists should not use non-epistemic values when they are justifying their hypotheses. The view has enjoyed varying degrees of popularity. It was popular in the high days of neopositivism. Then it became controversial. It was both criticized (e.g., by Rudner 1953) and defended (e.g., by Jeffrey 1956). In the last fifty years it was generally regarded as obsolete, for example, by Douglas 2009, Elliott 2011 or Tsui 2016. Recently some researchers started, once again, to look at it more favorably, for example, Betz 2013 or John 2015. My paper fits into the last group. I will defend the VFI by showing that if we accept the use of non-epistemic values prohibited by it we are forced to accept, as legitimate scientific conduct, some of the disturbing phenomena of present-day science, for example, preference bias. Secondly, I will show that value-laden science contributes to the replication crisis.

In the second section, I will discuss the VFI in more detail. Then, in the third section, I will present two views which were proposed as alternatives to the VFI: Douglas' value-laden science and Ludwig's proposal concerning ontological choices. In the fourth section, I will present problems caused by the rejection of the VFI. I will conclude by showing how value-laden science contributes to the replication crisis in science and, following Betz 2013 and Levi 1960 argue that the VFI is realizable.

6.2 The Value-Free Ideal: motivation and controversy

Generally speaking, the value-free ideal is a view which claims that reasoning in science should not be influenced by non-epistemic values of the scientist. I will follow Steel 2010 in understanding epistemic values as values which promote the attainment of truths, all other values are non-epistemic. Examples of non-epistemic values are fairness or equality while an example of epistemic one is internal consistency. The initial formulation of VFI is too strong, some of the uses of these values in science seem to be unproblematic. For example, there is nothing wrong with scientists choosing topics of their studies on the basis of their interests. I will defend the VFI restricted to the context of justification. It claims that scientists should not use non-epistemic values when they are justifying their claims.[1]

It is important to remember that the VFI is an ideal. A few sometimes overlooked consequences follow from this. First of all, no descriptive claim follows from the VFI. It may be the case that no one complies with a given norm but it is still a valid norm. Secondly, it seems that even if it is not possible to satisfy an ideal, this does not make it invalid. It may still be the case that by approaching an unattainable ideal we will become better off. It seems to be true for other ideals which are widely believed to be valid. For example, impartiality is typically understood in the following way:

> "The word "impartial" connotes absence of bias, actual or perceived." (UNODC 2019 p. 28.)

Both, a jury and the justice system overall are required to be impartial. At the same time, in light of the psychological evidence showing the prevalence of the biases is seems that there was never and never will be a perfectly impartial jury or the perfectly impartial judicial system. At the same time, it not seems to make the ideal of impartiality any less valid or attractive and possible improvements in impartiality any less beneficial (see e.g., Gobert 1988 or Cammack 1994). It may be similar in the case of different ideas, for example, a perfectly true (perfectly empirically adequate) scientific theory may be

[1]Some authors claim that the use of non-epistemic values is harmful even in the context of discovery (e.g., Okruhlik 1994).

unattainable. It does not show that we should not aim at truth or impartiality. The impossibility in all three cases seems to be analogous in nature. Plausibly, we are not able to fulfill any of those ideals because of our limitations.

As we have already seen, the VFI is controversial. Arguments both for and against it were presented. Du Bois 1898 presented an argument for the VFI based on the role of science in a democratic society. At the same time, the VFI was supported by a growing consciousness of the gap between values and facts (e.g., Weber 1949). On the other hand, a number of arguments were presented against it. Following Betz 2013, we can divide them into two groups.[2] The semantic arguments show that the VFI is not a sound position. For example, Dupré 2007 argues that the distinction between epistemic and non-epistemic values on which the VFI is based is not well-defined. The methodological arguments show that the VFI is unrealizable. For instance, Longino 1990 argues that because every scientific theory is necessarily undetermined by the empirical evidence and the gap between both has to be filled with non-epistemic values. Similarly, Douglas 2009, following Rudner 1953, argues that, in the case of policy-relevant science, scientists have to use non-epistemic values while making inductive inferences. Finally, Ludwig 2015 argues that scientists have to use non-epistemic values to make ontological choices on which their theories are based.

The methodological arguments rely on the assumption that an ideal has to be realizable in order to be valid. As I pointed above, this assumption seems to be problematic, at the very least, should be explicitly defended. As far as I know, it was not. The second assumption is that the only way to fill the identified gaps in scientific practice, be it justifying the inductive reasoning, ontological choice or any other of the proposed roles, is to use non-epistemic values. This assumption was undermined for example, in Levi 1960, Betz 2013 or John 2015. If arguments present in those papers are sound and therefore the second assumption is false then VFI is realizable after all. I will discuss this in the last section.

Authors who criticize the VFI usually propose alternative ideals of scientific conduct. Not surprisingly, the core of all these proposals is a claim that scientists have to or even should use non-epistemic values when they are justifying

[2]For the discussion concerning relation between two types of arguments see: ChoGlueck 2018.

their hypotheses. In the next section, I will sketch two of such proposals.

6.3 Value-Laden Science

In this section I will present two counterproposals to the VFI.

Heather Douglas describes in the fifth chapter of her book "Science, Policy, and the Value-Free Ideal" the value-laden ideal of science. She starts by distinguishing two ways in which a scientist can use non-epistemic values. Firstly, she can use them in a direct way, when they serve her as reasons and evidence. Secondly, she can use them in an indirect way. According to Douglas, non-epistemic values can be used in a direct way just in the early phases of the scientific process. For example, scientists can decide on the basis of their values which problem they will work on. During justification, the part of the scientific process interesting for our purposes, scientists should not use values in a direct way, because values should not be treated as reasons. On the other hand, in opposition to the VFI, she claims that values should play an indirect role in the context of justification. What exactly is meant by the *indirect role*? Scientists have to be responsible for the theories they accept. Therefore, according to Douglas, in order to make the final judgment concerning the acceptance of a given hypothesis a scientist has to take into consideration the non-epistemic consequences of her possible mistake. The essential part of such considerations is played by non-epistemic values. They are necessary to assess how costly is a possible error and therefore what amount of evidence is necessary to accept the hypothesis in question. It is important for our purposes to add that, according to Douglas, making the final decision to accept a hypothesis is not the only part of the justification where scientists should use non-epistemic values:

> "Choices regarding which empirical claims to make arise at several points in scientific study. From selecting standards for statistical significance in one's methodology, to choices in the characterization of evidence during a study, to the interpretation of that evidence at the end of the study, to the decision of whether to accept or reject a theory based on the evidence, a scientist decides which empirical claims to make about the world."(Douglas 2009, 103)

As far as I understand, she claims that they should use non-epistemic values

to make all decisions which are not determined by the evidence.

Ludwig 2015 proposes another role for non-epistemic values. He claims that scientists have to use them when they make ontological choices. For example, choices concerning the meanings of theoretical terms used by their theories, such as species or intelligence. According to the author, such choices cannot be avoided or made without the use of non-epistemic values and therefore scientists have to use them. Again, it seems to be less of an argument against the VFI then against its realizability.

6.4 An argument for the Value-Free Ideal

Now, I will show that the rejection of the VFI legitimizes unacceptable scientific practices.

Consider scientists who test some hypothesis, for example, the null hypothesis that some new medical product has no negative side effects. Clearly, the result of the experiment will impact the society in which scientists live. If it is a (false) negative, the policy-makers are likely to ban the product. A lucrative industry and many job opportunities will not be created. On the other hand, if it is a (false) positive, the potentially harmful product will be used in clinical practice. Now, according to Douglas' picture, a scientist should consider the negative impact of both false-negative and false-positive results, and on the basis of that adjust the amount of evidence necessary for accepting the hypothesis. Let us assume that there are two scientists who conduct two experiments with the aim to test the hypothesis concerning the harmfulness of the product. Both of them choose an established method and, luckily, they get the results that perfectly well match the distribution of the property in question in the population. At the same time, the results are not conclusive, the revealed correlation is not very strong. The first of the scientists was raised in a small town with high unemployment and sees in the new product, mainly, a great opportunity to reduce unemployment in poorer regions of the country. Consequently, she sets a high threshold of evidence in order to lower the chance of false positives. She concludes that the new product is not dangerous. The second scientist has a history of serious diseases in her family and therefore she thinks that her priority should be to defend the population against possible harm caused by the new product. Therefore, she a low standard for rejecting

the null hypothesis and concludes, on the basis of the collected evidence, that the product is dangerous.

According to Douglas' picture, both scientists perform perfectly well, at the same time they draw different conclusions from the same data. Neither of them used non-epistemic values in a direct way and we have no reason to suppose that they did anything else wrong. At the same time, it seems that something went wrong. The final conclusions were not determined by data but by the personal histories of both scientists. The example shows that even an indirect influence of non-epistemic values can be a deciding factor. Nothing hangs on the fact that the scientists are using the non-epistemic values to fix the sufficient level of evidence rather than to make any other methodological choice.

Douglas does not specify to which degree one can lower or raise a sufficient level of evidence. She just points out that values should not weigh stronger than the evidence. If we can raise the required level of evidence arbitrarily high or low, it is hard to see how that can be a consequence of her view rather than an additional postulate.

The problem is not just theoretical. Time and again, scientists used their non-epistemic values in the way permitted by Douglas to deliver results that fit their non-epistemic interest. An example of that is the preference bias (see e.g., Stelfox et al. 1998) which consists of the fact that empirical studies are significantly more likely to support a result preferred by researchers or their employers. Biased studies are not frauds, scientists do not fabricate their results or use other clearly unacceptable strategies. At the same time, there seems to be something wrong with these experiments. To see that, let us consider an example from Wilholt 2008. Industry-funded studies that tested the toxicity of a chemical substance called Bisphenol A tend to use a special strain of rats that are less susceptible to the substance. In effect, none of the industry-funded experiments showed a carcinogenic effect of the substance in opposition to 90% of government-funded studies. The non-epistemic values were not used in a direct way. On the other hand, it seems that the scientists used their non-epistemic values, perhaps self-interest while making the methodological choices concerning which strain of rats to use. According to Douglas' story, the industry-funded scientists did nothing wrong. At the same time, such cases seem to be really disturbing and as noted by Wilholt 2008, they significantly

decrease our trust in science.

It is similar in the case of Ludwig's proposal. As we remember, he claims that scientists should make their ontological choices on the basis of their non-epistemic values. At the same time, the author rightly claims that the results of experiments depend on these choices. Therefore, the results of the experiments depend on non-epistemic values. The way the values influence the results (through ontological choices or through the adjustment of the required level of evidence) does not make much of a difference.

Once again, the problem is not just theoretical. An example from Oreskes and Conway 2010 shows how tampering with a notion of causality can misrepresent a scientific result. I will assume, for the purpose of the example, that a choice concerning the conceptualization of causality is an ontological choice. Causality is not explicitly mentioned by Ludwig among ontological notions but it seems to play a similar role in a scientific experiment:

> "When asked if a three-pack-a-day habit might be a contributory factor to the lung cancer of someone who'd smoked for twenty years, Cline again answered no, you "could not say [that] with certainty. . . . I can envision many scenarios where it [smoking] had nothing to do with it." When asked if he was paid for the research he did on behalf of the tobacco industry, he acknowledged that the tobacco industry had supplied $ 300,000 per year over ten years — $3 million but it wasn't "pay," it was a "gift." "(Oreskes and Conway 2010, 31)

The specialist in question, Martin I. Cline, seems to be using an implausibly demanding version of the regularity theory of causation to misleadingly present scientific results. The theory he implicitly endorses claims that a necessary condition for z to be a contributory factor of y, is that in all (possible?) cases in which z is present y is. That would explain his reluctance to admit that "three-pack-a-day habit" is a possible contributory cause of lung cancer. At the same time, the theory of causality is too demanding. The questioned claim seems to be true, in light of the scientifically informed common sense, as soon as we accept any less demanding theory of causality. The quote is taken from a transcript of a trial in which Cline served as an expert. It is hard to imagine a more socially responsible function for a scientist. Given that he

received money from a tobacco company, a strong case can be made that the misleading ontological choice was caused by one of his non-epistemic values, once again self-interest.

Another example of an ontological choice being problematic because of the influence of non-epistemic was discussed in Bishop 1990. Dorothy Bishop discusses studies investigating an association between handedness and developmental disorders. The field is filled with inconsistent results:

"...developmental dyslexia has been linked to mixed hand preference (Harris, 1957), strong left-handedness (Geschwipd & Behan, 1982), inconsistent right-handedness (Schachter, Ransil, & Geschwind, 1987) and both left- and strong right-handedness (Apnett & Kilsbaw, 1984). There are also many studies that report no association (see review by Bishop, 1983)." (Bishop 1990)

She shows through simulation that if a scientist is free to choose the way she conceptualizes the handedness after she already obtains the experimental data, she is almost guaranteed to get a positive result. According to the author, the fact that scientists making their choices motivated by a desire to find a significant result at least partially explains the inconsistent results.

In both cases, problems are caused by the non-epistemic values of a scientist influencing her methodological choices. At the same time, I do not want to claim that all uses of non-epistemic values are equally problematic. Plausibly, some of the decisions motivated by the non-epistemic values do not lead to misrepresentations as one described above. The problem for value-laden science is that it is not able to distinguish between the problematic cases and a legitimate scientific conduct. Any value-laden proposal needs to be augmented with some additional criterion in order to exclude those and similar problematic cases. To be sure, there are different proposals concerning distinguishing acceptable and problematic uses of non-epistemic values. Many of them, including proposals appealing to the lexical priority of evidence (see e.g., Anderson 2004 or Brown 2013) and a proposal based on already discussed distinction between epistemic and non-epistemic values (Steel 2010) were critically discussed in Hicks 2014. Hicks' arguments are convincing, all proposals he discusses seems to fall short of providing a successful demarcation. On the other hand, his counter-proposal is not fledged enough to provide a clear verdict for each case and therefore it successfulness is hard to evaluate. There may be more general reason why all the examined proposal concerning the de-

marcation of problematic and acceptable cases seems to be unsuccessful. If the only feature common to all of the problematic cases is that they involve non-epistemic values motivating the crucial methodological decisions, all additions to value-laden science which would be successful in excluding them would have to exclude all the uses of non-epistemic values and therefore transform it into a VFI-like proposal. Is it plausible it is that there are no other features common to all of the problematic phenomena? To answer this question we have to first introduce *researchers' degrees of freedom* (Simmons, Nelson, and Simonsohn 2011; Wicherts et al. 2016). During designing and conducting scientific experiment a scientist has to make a lot of decisions concerning the exact shape of the experiment. For example, she has to decide: How to collect the data? When to stop collecting the data? Should some observations be excluded? How to understand theoretical terms? Each of these issues can be approached in many equally correct and some incorrect ways and this constitutes researchers' degrees of freedom.

Researchers' degrees of freedom can be misused in many different ways. First of all, a decision may be biased, that is, it can be made with a specific result in mind as in the case described by Wilholt 2008. Biases come in many different flavors (see e.g., MacCoun 1998 for a detailed discussion). Some of them are conscious other unconscious, some of them are caused by the preferences of the scientist (as the one we discussed), others by their prior beliefs (confirmation bias).

Secondly, after the experiment has been performed, a scientist can consider which combination of possible choices will likely generate a positive result and report just these choices. This practices are called Questionable Research Practices (see e.g., Simmons, Nelson, and Simonsohn 2011). Other examples of questionable research practices are: using under-powered statistical designs and optional stopping, which I will discuss in the next section. As far as I know, these phenomena (questionable research practices and biases) are not considered to belong to one homogeneous group and were not analyzed as such. At the same time, in all of the mentioned cases, non-epistemic values plausibly motivate a crucial decision(s). While it is hard to show that there is no other feature common to all the discussed problems, I think it is fair to claim that until such a feature, is identified, value-laden science is unable to explain why the mentioned practices are problematic. This seems to be a substantial

disadvantage given how important these problems are.

6.5 Value-Laden Science and The Replication Crisis

In this section, I will show how value-laden science contributes to the replication crisis and point to, in my opinion, convincing response to the methodological arguments against the VFI.

Somebody may still think that there is nothing wrong with using values in the way proposed by Douglas 2009 or Ludwig 2015. She may think that there is nothing wrong with two scientists reaching different conclusions from the same observation because of their different sympathies or interests. I do not think that this response is plausible or, in other words, that the bullet is worth biting. To see that, let us consider the *Replication Crisis* (for discussion see: Open Science Collaboration 2015 or Romero 2017). The crisis consists in the fact that the results of many scientific experiments are not replicable. This means that an experiment with similar or identical design conducted by different scientists (or even the second time by the same researchers) delivers different results. The exact percentage of replicable studies is unknown. Some approximation is provided by Open Science Collaboration 2015. The authors attempted to replicate the results of one hundred experiments from papers published in prestigious psychological journals. Less than half of the attempts were successful.[3] This is a disappointing result. In light of it, it is hard to have any confidence in the truth of a psychological hypothesis based on a single experiment. The crisis is generally perceived as something very disturbing, as the name itself suggests. Many philosophers and methodologists explore different ways to deal with it (e.g., Ioannidis 2005).

At the same time, according to Douglas' story, this situation is expected and natural. If different scientists share different values, they make different methodological decisions. As a result of that, they sometimes, or maybe even often, draw different conclusions even from the same data. If it is the case, the fact that the scientists responsible for the original study share different values than the scientists responsible for its replication may be a reason for the failure

[3] "Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result..."(Open Science Collaboration 2015, 943)

of the replication attempt. To see that, let us go back to our example (page 5). If the first scientist tests the hypothesis concerning the harmfulness of x and the second one will try to replicate the result, given the inconclusiveness of the data, the replication will fail no matter how similar they will be. In this case, given the differences in the statistical design, the second experiment is an attempt to conceptually replicate the result. The conceptual replication, unlike the direct replication, does not impose using an identical statistical design in both experiments (see e.g., LeBel et al. 2018). Therefore, in the case of a direct replication, a scientist who attempts to directly replicate an experiment cannot use her non-epistemic values to adjust the sufficient level of evidence, she has to follow the design of the original study. Even in case of the direct replication, there are some decisions that scientists can use to tamper with the outcome of the experiment. For example, it is typically not required for replication to use the same population as the original study. As we have seen, these choices can be used to change the results of the experiment.

Once again, the examples from Wilholt 2008 and Bishop 1990 shows that this mechanism is present in science. In the case of industry biased studies, it is hard to expect that the biased study will be replicated by an unbiased experiment. Similarly, in the case of studies of handedness is not highly probable that the result of any of the experiments will be replicated by a high-quality replication given their conflicting results and likely use of the questionable research practices.

The connection between the use of the non-epistemic values and the replication crisis is acknowledged in methodological literature. For example, consider two factors named as the causes of questionable research practices by Simmons, Nelson, and Simonsohn 2011:

> "(a) ambiguity in how best to make these decisions and (b) the researcher's desire to find a statistically significant result."(Simmons, Nelson, and Simonsohn 2011 p. 1359.)

The first cause is the existence of researchers' degrees of freedom, the second is the fact that scientists value statistically significant results. Both clauses are connected by an implicit assumption that this non-epistemic value motivates the methodological choices. The use of questionable research practices greatly inflates a chance of a false-positive result and therefore lowers the replicabil-

ity of the given study. Similar factors are predictors of false-positive results according to Ioannidis 2005:

> "Corollary 4: The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true."(Ioannidis 2005 p. 0698.)

> "Corollary 5: The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true."(Ioannidis 2005 p. 0698.)

Interestingly, some of the questionable research practices can be a way of "deciding what amount of evidence is necessary" recommended by Douglas and other defenders of value-laden science. A perfect example is the *optional stopping*.[4] A scientist using this procedure, instead of deciding beforehand how many subjects she will use in her experiment, decides during the experiment. She either stops and concludes the experiment after some subjects were tested or adds additional subjects. Unsurprisingly, scientists seem to be inclined to conclude their experiments when the correlation they hoped for is present. Similarly, to other questionable research practices, the procedure increases the rate of false-positive errors.

Another decision a scientist has to make while designing her experiment is how statistically powerful it will be. Statistical power is the likelihood that an experiment will detect the effect in case it exists. If the effect size is small a large number of participants is required to get a moderately powerful experiment. Because of that scientists are sometimes inclined to conduct many cheaper, less powerful studies instead of one more powerful study with the hope to get a significant result in at least one of them by luck. A low statistical power of an experiment is also responsible for its low replicability (e.g., Vazire 2016 or Ioannidis 2005).

It is not clear what exactly proponents of value-laden science have in mind when they write about "deciding what amount of evidence is necessary". On the other hand, given that neither of the above procedures involves using non-epistemic values in a direct way, it seems that they can be precisely cases of manipulation they advise scientists to use. At the same time, both contribute to the replications crisis.

[4]See e.g., Heide and Grünwald 2017 or Montori et al. 2005.

As we have seen, according to Douglas both scientists in our example performed perfectly well. Similarly, she advises a scientist to use her non-epistemic values to adjust the sufficient level of evidence, which is, at the very least, consistent with advice to use optional stopping or under-powered experiments as soon as it fits her values. Given the effects these practices have on replicability, it seems that proponents of value-laden science have to see a replication crisis as a natural consequence of scientists using their values in a perfectly admissible way. In other words, according to them, the replication crisis is not a bug, but a feature of science. If scientists are allowed to use their non-epistemic values in the context of justification, they can systematically use them to tamper with the results of their studies and, at the same time, increase the rates of false-positive results and contribute to the replication crisis. Consider this together with the role played by replicability according to many scientist or methodologists, for instance:

> "Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence."(Open Science Collaboration 2015, 943)

In light of the crucial role of replicability, the very idea of value-laden science seems to be misguided.

Finally, I want to point to an answer to what is often seen as a concluding argument against the VFI, namely the methodological argument. As we have seen, various authors (e.g., Rudner 1953 or Douglas 2009) claimed that because of under-determination by data scientists have to use non-epistemic values to decide what amount of evidence is necessary. Others point to different decisions that require the use of non-epistemic values (e.g., Ludwig 2015). These authors conclude that because of those gaps, VFI is unrealizable and we should look for alternative solutions. Many found this and similar arguments convincing and therefore the VFI unattainable. Even if we grant that from the fact that an ideal is impossible to realize it follows that it is not a valid ideal, a highly doubtful inference, it seems that the challenge was answered satisfactorily. Betz 2013 describes a general strategy for hedging scientific results in cases where some of the assumptions or methods are not standard or well-supported. For example, if the experiment design we use is not standard,

we can put our results in form of a conditional statement with the assumption that the design does not distort the results in the antecedent and original non-hedged results in the consequent. One can use the strategy in cases of other methodological decisions which, according to the methodological arguments, require the use of non-epistemic values. For instance, both scientists from our example could present their results in the following form: "Given the used threshold of evidence, the substance x is harmful/harmless". The author presents uses of this strategy from actual scientific practice. Surprisingly, even some philosophers which stand against the VFI seem to favor a similar solution. As noted by Betz, Douglas 2009 proposes a similar strategy. Similarly, Rudner 1953 points out that it is essential for a scientist to makes his value decision explicit. One way to do so is to follow Betz's strategy.

In some cases, the strategy may seem to be impractical. Perhaps, it is so in the cases of the ontological choices. It seems to be impractical to list the results of all the ontological choices done on behalf of an experiment. Similarly, some assumptions may be just too standard or widely shared to be listed. The strategy described in Wilholt 2008 seems to be more appropriate in these cases. As we have already seen, the industry-funded scientists use a special strain of rats while examining the toxicity of Bisphenol A. These rats are less susceptible to the substance and therefore the choice makes the experiment far less likely to show its harmful effect. In analyzing the preference bias Witholt hastiness to blame, what seems to be the main cause of the problem, namely non-epistemic values of the scientist affecting their methodological choices. Instead, he points out that in all cases of preference bias scientists failed to comply with some important scientific conventions. For example, in order to prevent the use of insensitive strains of animals scientific community formed the following convention:

> "Because of clear species and strain differences in sensitivity, animal model selection should be based on responsiveness to endocrine active agents of concern (i.e. responsive to positive controls), not on convenience and familiarity" (National Toxicology Program, US Department of Health and Human Services. 2001,vii)

This does not seem to be a satisfying analysis. What with the instances of similar practices that occurred before the norm was put in place? They inspired the norm but cannot be explained as a case of biased science by the lack

of compliance with a norm that was not yet constituted. Wilholt claims that in such cases the norms may be implicit. This defends his analysis against the obvious counter-examples but, at the same time, makes it much more speculative. We can always postulate an implicit norm. A more satisfying analysis should take into consideration what behavior of scientists was meant to be prevented by these norms in the first place.

Given the described cases, it seems plausible that these and similar norms aim to prevent the uses of non-epistemic values which can change the outcome of a given experiment. This suggests the following conceptualization of the preference bias:

> A study is preference-biased iff during its course a decision was made which was caused by the non-epistemic values of an involved scientists and this decision changed the outcome of the study.

It seems to be a more straightforward and satisfying explanation of preference bias.

At the same time, the process during which biased studies inspire new scientific norms is interesting for our purpose. In light of all the above, it seems that by doing so the scientific community releases the individual scientists from some of their responsibilities. Some of the choices, for which they were responsible before, are no longer required. Instead, they just need to follow the new convention. This restricts the way in which scientists can tamper with their experiments. Science is full of such conventions. A notable example is the value of formal statistical significance set by convention to 0,05.[5] Recently some conventions were proposed to regulate the statistical power required for an experiment to make it appropriate for publication for example:

> "If the effect being examined is likely to be in the range of typical published effects in social/personality psychology, which is an

[5]Another example is: "The diversity of the methodologies used in the assessment of the efficacy of β-alanine highlights the importance of clear logical progression through the different aspects of supplementation to eventually be able to produce a clear concise set of criteria for its efficacy. Authors should also make every attempt to demonstrate: the purity of the supplement used; the double-blinding of the treatments; and the reliability of the exercise tests or measure employed."(Hobson et al. 2012) As we see, here the authors express a need for a conventionally fixed efficiency criterion. Then in the next sentence, they formulate few methodological recommendations e.g., one concerning a need for demonstrating the purity of the substance used in the experiment. With the support of the community of scientist, the advice can become a fully-fledged convention.

> r of .21 or d of .43 (Richard, Bond, & Stokes-Zoota, 2003), then researchers should aim for a sample size that will provide at least 80% power to detect such an effect or more..."(Vazire 2016)

The convention specifies the statistical power for which scientist should aim in order to get published. If we interpret the question about the necessary level of evidence posed by proponents of value-laden science in terms of the statistical power then a scientist does not have to decide herself.[6] Instead, she just has to conform to the convention, which is at the same time a precondition for being published.

It is easy to see a relation between this convention-forming mechanism and Betz's strategy. Conventions play a role similar to assumptions. Neither are meant to be tested and both impose a particular methodological decision. From the perspective of an individual scientist, it seems to make little difference if she assumes something or follows a convention to the same effect. If an assumption becomes popular and well established enough, it can be turned by an appropriate sanctioning body into a convention.

In addition to heading and conforming to scientific conventions, there are other ways to restrict the flexibility of a given experiment and therefore the need for use of non-epistemic values. As we have seen, Ioannidis 2005 lists such flexibility as one of the main factors which increase the probability that a result of a scientific experiment is false. In response to this problem, he recommends standardizing the conduct and reporting of research designs (see e.g., Schulz, Altman, and Moher 2010). In line with that, many other ways to reduce the methodological flexibility were proposed and some of them becomes standard tools for ensuring research quality for example, preregistration (see e.g., Nosek et al. 2018) which amounts to a requirement that the experimental design used by a scientist has to be specified and registered before the data are collected or multiverse analysis (see e.g., Steegen et al. 2016) which consist in concluding a statistical analysis for some/all methodological choices that can be made during data analysis. Both of these requirements restrict ways in which scientists can use their non-epistemic values just as an additional convention. In principle, it seems is possible that the further development of scientific methodology will lead to the construction of a system of conventions

[6] A similar proposal was presented in Levi 1960.

and other precautions which will eventually make value-laden choices unnecessary. If so, the VFI may be realizable after all.

These strategies do not only make a strong case for the realizability of the VFI; they also give us a clear idea of what a scientist can do to approach it. She has to follow the standard scientific conventions for a given problem when justifying her claim. Whenever the conventions do not specify which methodological decision should be made, she should make the decision herself and explicitly present it as an assumption of her result. As we have seen, even if the choice is motivated by her non-epistemic values, a hedged conclusion does not depend on them. This easy-to-follow procedure restricts biased decisions. Either the convention was followed and therefore the non-epistemic values of scientists were not involved or the final result was hedged and because of that, it does not depend on the values. Additionally, she should preregister her study and if possible use a multiverse analysis. If we combine this strategies with the above argument the VFI becomes, once again, a viable and perhaps even attractive proposal.

6.6 Conclusion

In my article, I defended the VFI by showing that if we reject it, we are forced to accept, as legitimate scientific conduct, some of the disturbing scientific phenomenons like preference bias or questionable research practices. I showed how two popular proposals of value-laden science (Douglas 2009 and Ludwig 2015) lead to this problem and presented some examples. I also showed that value-laden science contributes to the replication crisis. Finally, I presented strategies for making the VFI realizable. Following Betz 2013, a scientist can hedge her final result and therefore made it independent from her (value-laden) methodological choices. Secondly, as proposed by Levi 1960, a scientific community can instantiate a scientific convention that recommends a particular solution for a given methodological problem and therefore makes a corresponding (value-laden) choice unnecessary.

Chapter 7

Objectivity for the Research Worker

7.1 Introduction: a Story About a Scientist

Despite its undeniable success (e.g., electricity, space flights, etc.), science seems to be in a difficult position today. In the last few years, many problematic cases of scientific conduct were diagnosed, some of which involve outright fraud e.g., Stapel 2012 while others are more subtle e.g., supposed evidence for precognition; Bem 2011. These particular issues and the general lack of replicability of scientific findings e.g., Open Science Collaboration 2015 have contributed to what has become known as the Replication Crisis e.g., Harris 2017. In addition, the general public has become aware of these problem, which reduced the general trust in science e.g., Lilienfeld 2012.

Let us imagine a scientist, Dr. Jane Summers. Dr. Summers does research in cognitive psychology. One day, she reads about the low replicablity rate of the results of psychological studies Open Science Collaboration 2015. She becomes very concerned about the value of scientific results in general and her own research in particular. As a result, she is resolved to investigate to what extent her work is at risk of irreplicability and to ensure that her current and future work is as robust as possible against such a fate. She decides that, apart from ensuring the accuracy and precision of her measurements, the methods she employs should not be significantly influenced by her feelings, values, biases

and other idiosyncrasies. To her, this means that they are objective Hawkins and Nosek 2012; Ziman 1996.[1] Objectivity can be attributed, among others, to scientific measurement, the tool for development/improvement of scientific theories, and/or to true-to-nature explanations. It ensures that study outcomes are not systematically biased e.g., over estimation of drug efficacy, under estimation of risk; Goldacre 2014, positive research results are not false-positives to a larger proportion than is allowed by the statistical method; Simmons, Nelson, and Simonsohn 2011, and are independently reproducible by other scientists Altmejd et al. 2019; Bavel et al. 2016; Lindsay 2015; Simons 2014. Dr. Summers considers objectivity to be essential to science[2] and its absence to be a cause of the crisis that threatens the foundations of her research field. Thus, she considers the assessment and safeguarding of scientific objectivity as being of vital importance.

It is therefore somewhat puzzling to her that a proper explication of objectivity appears to be lacking in science. She is unable to find tools for the qualitative and/or quantitative assessment of objectivity. Methodological reforms are inspired by problematic cases, for instance, measurement results inconsistent with the laws of nature e.g., precognition; Bem 2011 or failures to reproduce established experimental results e.g., Open Science Collaboration 2015, rather than a clear understanding of objectivity. She realizes that science could greatly benefit from having a definition of 'objectivity' that can be explicated in a quantitative or qualitative assessment of scientific practice.

Dr. Summers has a hunch that philosophy might be of assistance in defining objectivity. After a short review of the philosophical literature, she does not manage to find a notion of objectivity which is ready for use in scientific practice. some of the proposals are mostly descriptive while others are difficult or impossible to test in practice (see Section 7.2). In effect, Dr. Summers becomes disheartened and contemplates quitting her quest for objectivity.

It is our opinion that we, philosophers, should not disappoint scientists like Dr. Summers in this respect and that philosophy can and should do better. We believe that the philosophical literature currently lacks a scientifically useful

[1] A similar description of objectivity can be found on Wikipedia: `https://en.wikipedia.org/wiki/Objectivity_(science)`.

[2] Interestingly, there are not many references on the importance of objectivity for science. Scientists we spoke to consider its relevance obvious and self-explanatory to such an extend that it does not warrant explicit explanation and justification.

conceptualization of objectivity and we intend to fill this gap. We will present a new way of thinking about objectivity of scientific practice. We understand scientific practice as pertaining to empirical research, which include all activities done by scientist essential for this endeavor. These include study design, data collection and measurement, data analysis, result reporting etc. [3]

Given that objectivity seems relevant for science and current philosophical definitions appear to be impracticable, it is our primary aim to come up with a notion of objectivity that can be used by the individual scientist. Therefore, we consider our new notion to be normative rather than descriptive. It should be noted that our notion of scientific objectivity will be neither complete nor static; it is open to future supplementation and revision. In the next section, we will briefly discuss philosophical views on objectivity and how they are of dubious practical use. In the third section, we will present a new version of a negative approach to scientific objectivity. We conclude with a discussion of the implications and limitations of our conceptualization and close with an outline of a usable and testable instrument for testing the objectivity of scientific practice.

7.2 Philosophy on Objectivity

In philosophy of science, scientific objectivity is a well discussed notion. Following Reiss and Sprenger (2017), we can list three main ways of conceptualizing it. Firstly, objectivity can be understood as a faithfulness to facts. Secondly, something can be understood as objective when it is free from value commitments. Thirdly, objectivity can be understood as being free from scientists' personal biases. Recently proposals which have gained much popularity are pluralist notions of objectivity e.g., Douglas 2004; Megill 1994; Wright 2018. Such notions encompass some or all of mentioned individual notions (e.g., the value-free objectivity, value neutrality objectivity, procedural objectivity etc.). Finally, there are negative conceptions of objectivity e.g., Daston and Galison 2010; Hacking 2015; Koskinen 2018 which claim that the objectivity consist of the absence of certain factors. In case of Daston and Galison 2010, these are factors of scientific subjectivity which are recognized by the scientific com-

[3]Note that for reasons clarified in Sections 7.3 and 7.4, we restrict our definition to research that works with non-qualitative (quantitative or countable) data.

munity as particularly troubling or important in a given time period. In the case of Koskinen 2018, these factors are epistemic risks which arise from the imperfections of epistemic agents.

Despite all this interest, it seems that a conceptualization of scientific objectivity which can be used by a scientist has not yet been proposed. Firstly, most of the proposals are aimed at being descriptively correct. This, in light of the conflicting intuitions and conceptual confusion surrounding objectivity, is a very useful endeavor but also distinct from the task of formalizing a notion that is normatively useful. In other words, it is not clear if these notions can fulfil the normative task of guiding scientific practice. Some authors are explicit about the descriptive nature of their proposals. For example, the aim of Heather Douglas (2004) famous article seems to be descriptive:

> *In this paper, I will lay out a complex mapping of the senses of objectivity. This mapping will make two contributions to current discussions. First, it will dissect objectivity along operationally distinct modes.[...] Second, the mapping will allow me to cogently argue that the different meanings of objectivity I explore here are not logically reducible to one core meaning.* Douglas 2004, p. 454-455

Similarly, Inkeri Koskinen (2018) is explicit about it descriptive nature of her proposal:

> *In this article I defend a risk account of scientific objectivity. The account is meant to be a largely descriptive or even a semantic one; my aim is to draw together ideas presented in recent discussions, and to clarify what we philosophers of science do when we identify distinct, applicable senses of objectivity or call something objective.* Koskinen 2018, p. 1

For other proposals, it is clear that they are descriptive due to their methodological approach. In the case of the Datson and Galison (2010) for instance, their historical methodology makes it a clear that it is a descriptive proposal.

Secondly, some of the normative conceptualization of objectivity are not suitable for use by scientists. Such a notions needs to be testable. Otherwise, how is a scientist to asses if given scientific practice is objective or not? An example, *value-free objectivity* seems to fail in this respect. Value-free objectivity is based on a more general *value-free ideal*. The value-free ideal claims that

scientists should not use their non-epistemic values like 'equality' of 'fairness' when they justify their claims e.g., Betz 2013. This conception of objectivity claims that a scientific justification is objective as long as it is not influenced by non-epistemic values. There might be reasons to believe that this notion is normatively compelling e.g., Betz 2013; Sober 2007, though it is easy to see the difficulty of its implementation. In general, we do not have access to scientists' intentions, thus we cannot judged what motivated their decisions and actions. The same can be said for value-neutral objectivity which also connects the objectivity to the values motivating the choices of scientists.

Detailed discussion of practical usability of all the proposed notions of objectivity is beyond the scope of our paper. However, we hope that this cursory sketch provides an idea of the problems with putting these notions into practice and motivate the value a new conceptualization of objectivity.

7.3 To see it from the other side: problems in science and the via-negativa approach to objectivity

There is no generally accepted positive definition of 'health' in health care and the medical sciences.[4] Fortunately, this does not prevent doctors from healing ailments and researchers from developing new drugs and technologies. A positive definition of health is unnecessary, when the instances that reduce or endanger health can be defined and addressed. In brief, health is what remains when the particular infirmities are removed.[5] Health care and medical science appears to be successful, even in the face of changing definitions, diagnostics, and disagreements about ailments. We believe that this via-negativa approach can also be applied to the concept of scientific objectivity.

[4]The World Health Organization defines health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity." Organization et al. 1950, p.100. Which is essentially a negative definition augmented with an well-being requirement. The definition is considered controversial and with little added benefit over the original negative definition e.g., Jadad and O'Grady 2008. Adding 'well-being' just kicks the can down the road.

[5]In this paper, the negative definition of 'health' is used as an analogy to clarify our approach. Our conceptualization has no stake in this definition or its controversies.

Our negative approach resembles to some extent other negative proposals in philosophy e.g., Daston and Galison 2010; Koskinen 2018. Just like these approaches, we conceptualize objectivity as the absence of certain factors, however, our aim and justification are different. Specifically, the purpose of our notion is to be relevant to and practicable by scientists. The identification and justification for objectivity-reducing factors comes from empirical research itself. We postulate that susceptibility to factors that have been empirically or methodologically identified as potential causes of, for instance, systematic bias or low replication rates e.g., Simmons, Nelson, and Simonsohn 2011 constitutes non-objectivity of scientific practice.

During scientific practice, objectivity can be compromised in various ways. Researchers make certain decisions when they design their study and collect, process, and analyze their data. The possibility of choosing between two or more options in these instances can be called *researchers' degrees of freedom* Simmons, Nelson, and Simonsohn 2011; Wicherts et al. 2016. The misuse of which can result in biased and/or irreproducible outcomes. The ways in which scientists can misuse this freedom can be grouped into two categories. Firstly, a scientist can make a priori decisions concerning the research design and data collection, which can preclude certain outcomes or make them more/less likely (i.e. introduce systematic bias in a certain direction). Secondly, a scientist has to make decisions on how to process and analyze the data, which allows her to try all possible combinations of decisions until a positive/desired result is found.

In this section, we first focus on problematic practices taking place before or during conducting the study (see Subsection 7.3). Then, we discuss that which can be problematic after the study was conducted and the data is analyzed (see Subsection 7.3). The section is concluded with a tentative conceptualization of objectivity resilience to such problematic practices.

Problems before and during Research: Design, Data collection, and Measurements

During the early stages of a scientific experiment (e.g., designing the observational study or experiment, sampling, measurement, etc.), a scientist has to make several decisions, which could influence the final result. In some cases, a scientist might make such choices with the aim of obtaining a specific result

in mind. Such decisions introduce systematic bias e.g., Fanelli, Costas, and Ioannidis 2017.[6]

In science, biased research seems to results from the influence of beliefs or prejudices of the scientist on her methodological decisions. For example, scientists make methodological choices that increase the likelihood of getting results that align with the preferences of those that provide the research funds this is known as 'funding bias' Jones and Sugden 2001; Nelson 2014. Similarly, a scientist can adjust the design of her experiment or observational study, conciously or subconciously, in order to increase the probability that the results will support her prior beliefs.

Typically, we can pinpoint a single decision or small number of decisions responsible for the biased outcomes(s). Usually, it takes place in the early phases of the scientific process. For example, the biased studies presented in Wilholt 2008 involve scientists choosing a specific strain of experimental animals, which made the experiments significantly less likely to show the toxicity of the tested substance (in line with the preference of the funding institution). Next to sample selection, bias can be introduced in many of the other decision that a scientist has to make when designing and conducting research, specifically:

1. Which measurement (outcome measure) to use?
2. Which kind of independent variable (experimental manipulation) to use?
3. Which sample to select and how?
4. Setting of the experiment or observational study (when and where)?
5. How and to what extent do researcher and research subject interact?
6. How to perform the measurement (e.g., blinded or unblinded)?

Recognition of features that can introduce bias is reflected in proposals concerning how to counter it. For example, Wilholt 2008 proposed establishing conventions which regulate the way scientist should conduct their studies as a remedy to funding bias. In the case of choosing insensitive animals, he proposed to adopt the following convention:

[6]For clarification, bias as discussed in this paper is different from bias in psychology e.g., MacCoun 1998 where it is used to classify cognitive heuristics (e.g., confirmation bias, bandwagon effect, anchoring, etc.). These heuristics might indirectly influence results, but these distant causes are irrelevant to our approach.

> "Because of clear species and strain differences in sensitivity, animal model selection should be based on responsiveness to endocrine active agents of concern (i.e. responsive to positive controls), not on convenience and familiarity" National Toxicology Program, US Department of Health and Human Services. 2001, p.vii

Different conventions are and can be implemented in order to impose methodological restrictions on scientists. Some of them force scientists to measure the direct outcome of interest instead of a proxy, use standardized tests or measurements, use random sampling from the population, use random allocations of participants to conditions, use equal group treatment, use blind or double blind design (experimental studies), and/or use data collectors that are blind to the research (observational studies). All of the mentioned conventions restrict the range of biasing decisions a scientist can make.

Problems After Experiments or Observations: Data Management, Analysis Specification, and Result Reporting

After a researcher has run the experiment and the data has been collected, several decisions have to be made. For instance, the data needs to be processed (e.g., removing outliers, combining variables, binning variable values, etc.), the statistical model needs to be specified (e.g., linear model, multilevel model, structural equation model, etc.), and finally the dependent and predictor variables need to be select that are to be included in the model. The assumption is that for each step only one (and the most appropriate) of the possible options is selected. However, the general rate of false-positive results[7] is increased when, instead of taking a single option for each step, several possible combinations of options are explored and only the combinations that culminate in positive results are reported e.g., John, Loewenstein, and Prelec 2012; Simmons, Nelson, and Simonsohn 2011; Szucs 2016; Wicherts et al. 2016. These behind-the-scenes

[7]No matter how precise an instrument is and no matter how stringent the evidence requirements are, there is always a non-zero probability that the result of a study does not reflect reality (e.g., positive outcome of a HIV test when the person is actually HIV negative). Statistical methods of analysis come with certain rules and assumptions, which must be followed in order for this probability to have a known maximum. In other words, if a study is performed according to its rules, none of the assumptions are violated, and it is repeated a large number of times, the proportion of false-positive results (i.e., an effect is observed while actually no effect exists) is at most equal to this probability, which is or can be known.

practices that covertly influence results go by the name of *questionable research practices*. The causes of these practices may be the scientists' (sub)conscious beliefs or preferences, the ambiguity or ignorance about how the methods works and what the statistics are/mean, or the desire to find/see associations and structure in what is being studied. Concretely, at least the following decisions need to be made by a researchers when dealing with quantitative data and performing statistical analyses this incomplete list is adapted from: Bakker, Dijk, and Wicherts 2012; Kass et al. 2016; Nelson, Simmons, and Simonsohn 2018; Simmons, Nelson, and Simonsohn 2011; Wicherts et al. 2016:

1. How to handle incomplete or missing data?
2. How to pre-processes data (e.g., cleaning, normalizing, etc.)?
3. How to process data, deal with violations of statistical assumptions (e.g., normality, homoscedasticity, etc.)?
4. How to deal with outliers?
5. Which measured construct to select as primary outcome?
6. Which variable to select as dependent variable out of several that measure the same construct?
7. How to score, bin, recode the chosen dependent variable?
8. Which variables to select as predictors out of the set of measured variables?
9. How to recode or restructure these predictors (e.g., combining variables, combining levels of a variable, etc.)?
10. If and which variables to additionally include as covariates, mediators, or moderators?
11. Which statistical model to use?
12. Which estimation method and computation of standard errors to use?
13. If and which correction for multiple testing to use?

14. Which inference criteria to use (e.g., p-values and alpha level, Bayes factor, etc.)?

Note that, if such decisions needs to be made and how many option the scientists has to choose from depends on how the study was designed and the structure and size of the data that was collected.

Currently, there are already some potential strategies for restricting uses of questionable research practices (i.e., ad hoc decision making in order to get positive results, also known as p-hacking). For instance, a) preregistration of the study from design to analysis e.g., Chambers 2013; Nosek et al. 2018; Wicherts et al. 2016; b) data and analysis blinding e.g., MacCoun and Perlmutter 2015; and c) run several/all of the (theoretically) possible tests in a *multiverse analysis* Steegen et al. 2016.[8] It should be noted that such strategies are not mutually exclusive and that combinations are possible, because they all restrict researcher's degrees of freedom without introducing new ones. For instance, not all decisions can be made in advance, precluding their preregistration. In such a case, some of these can be caught by data blinding, because the scientist might not know what the data will look like in advance, though has an analysis plan that can be communicated to the independent data analysis. In addition, the multiverse analysis can be employed for those elements of the research that have an exploratory nature that do not allow for data blinding and handing the analysis to someone else.

To Sum up: a conceptualization of objectivity

Our negative version of conceptualizing objectivity ties it to scientific problems that result from the decisions and actions of individual scientists. These problems are notoriously hard to detect. For instance, a report of a study during which questionable research practices were used can be indistinguishable from the report of a study during which they were not used. If objectivity is just absence of these problems then our notion is not testable. On the other hand, we can easily tell if precautions against such problems (e.g., preregistration) are present and thus how resilient a given practice is. Therefore, we state that scientific practice becomes more objective when it becomes more resilient to

[8] In-depth discussion of these strategies is beyond the scope of this paper. Function, benefits, and limitations of these strategies can be found in the cited papers.

actions and decisions that can influence its outcome; concretely, when:

a) the study design and data collection becomes more resilient to the scientists' influence on the data;

b) and the data processing and analysis become more resilient to ad hoc decision making and selective reporting of positive results.

In the limit, a practice is objective when it is impervious to biasing influences and precludes ad hoc decisions and actions.

Our approach has two clear advantages: it is empirically verifiable and it does not require perfect/objective and universally agreed factors that reduce objectivity. Our notion, in opposition to traditional conceptualization (e.g., value-free objectivity), ties objectivity to features of scientific practice, the existence of which can be empirically tested (e.g., was the study preregistered or not). These features can be collected in a form of a checklist (see the Appendix C for a first setup), which can serve as a ready-to-use tool for assessing objectivity of scientific reports (e.g., manuscripts, published papers, grant applications, etc.).[9] In addition, objectivity according to this conceptualization can be verified by assessing the extent of systematic bias and inflated false-positive rates in a body of literature. The presence of objectivity promoting features like preregistration decreases the chance of a given study being a false positive. Therefore we can indirectly test the objectivity of a study for instance, by testing the consistency of results between original studies and their replications when they are preregistered in comparison to those that are not.

The second advantage follows from the first. We do not claim that the list of objectivity reducing factors on which our conceptualization is based is exhaustive. Moreover, some factors might be considered controversial as objectivity reducing or it may not be objective how factors are included, while other are not. This is not problematic for our proposal, because a) the identification and inclusion of factors is based on robust empirical results and methodological considerations; and b) their impact on the quality of the study, as explained in the previous paragraph, can be empirically verified.

[9]Similar checklist have been developed and are in wide use as tools for assessing methodological quality of studies e.g., Downs and Black 1998; Sindhu, Carpenter, and Seers 1997 when, for instance, appraised for inclusion into a systematic review e.g., Haidich 2010.

7.4 Discussion: Conclusions, Implementations, Limitations and suggestions for further research

In this paper, we have offered a novel and practicable theory of scientific objectivity. We have argued that many, if not all, philosophical attempts at defining objectivity are at least not practicable as they currently stand and likely to be impossible to test in scientific practice (see Section 7.2). In our approach, we have used findings from empirical research and methodological considerations to identify features of scientific practice considered to be problematic. We postulate that resilience to these features constitute objectivity. Given this list of features, scientific practice approaches objectivity when it becomes less vulnerable to actions and decisions by scientists that can influence its outcome. In this section, we discuss the limitations and implications of our conceptualization. In the Appendix C, we present a draft for a tool that can be used to assess the objectivity of scientific endeavours (e.g., published papers, submitted manuscripts, proposed research in grant applications, etc.). In addition, we suggest investigations into our tool to test and improve its validity and reliability. We close this paper with a detailed illustration of how this tool could be usefully implemented.

Limitations

Incompleteness. Plausibly, in our paper we do not reach a complete list of ways in which scientific practice can be compromised. Therefore, it is most likely that we did not reach a complete definition of objectivity, though rather a list of currently identified necessary conditions. However, our approach does provide a framework for learning from empirical research and methodological developments when, where, and how particular factors compromise scientific objectivity. Even with this limitation, we believe that our conceptualization is an improvement over previous attempts of conceptualizing objectivity and can still be used in a fruitful way (see Section 7.3).

Ritualization. Some might argue that restricting researchers in the proposed way will actually reduce objectivity. For instance, the (faulty) use of the Null Hypothesis Significance Testing procedure (NHST) has been equated with a restrictive ritual; a practice that discourages informed reasoning and prescribes

certain actions and decisions. The NHST ritual has been considered to be the main cause of the inflated number of false-positive results in science Gigerenzer 2004; Ioannidis 2005; Stark and Saltelli 2018, which is the opposite of what an objective method should achieve. However, the NHST ritual only appears to restrict researchers and provides just the illusion of objectivity. In particular, apart from inference criterion (i.e., an observed statistic lower than a conventional threshold), this ritual does not restrict (mis)use of degrees of freedom (mentioned in Section 7.3) at any point during the research process. Specifically, and in contrast to recommendations of our proposal, ad hoc decision-making in data management, analysis, and result reporting are not prohibited in the NHST ritual. It might even be considered that this partial formalization enshrines a false sense of objectivity that is actually harmful to the quality of scientific results e.g., Gigerenzer 2004; Simmons, Nelson, and Simonsohn 2011. In other words, if the ritual had been restrictive in ruling out questionable research practices, it would actually promote objectivity. Our conceptualization does recommend these additional restrictions. Also, in contrast to the conservative nature of a ritual, our conceptualization is (meant to be) adaptive; developed in accordance with novel discoveries concerning problematic scientific practices and methodological changes in science.

Restricted. Our conceptualization is restricted to practice of quantifiable or countable research, which precludes qualitative research and non-empirical practices. Qualitative research is currently omitted from our definition, because, to our knowledge, empirical research and methodological considerations on the particulars of systematic bias and false-positive rate inflation in the use of qualitative methods are currently absent in the academic literature. Non-empirical practices, like simulation studies, and theoretical reasoning are lacking, because they solely or heavily rely on the discretionary choices of the researcher, precluding the possibility of objectivity as we conceptualize it. However, the empirical claims and hypotheses that follow from such practices can be empirically tested/verified and these tests can be evaluated against our conceptualization of objectivity.

Objective research does not guarantee true results nor validity or reliability. Even if the work of a scientist did not suffer from anything that could jeopardize the research's objectivity, it is still possible that the results are not true (i.e., do not reflect or represent reality). It could be as innocent as a false-positive or

it might be that the measurement instrument is not adequate for investigating the phenomenon at hand. Either way, we should be clear that objectivity of a practice can not be equated with scientific truth generation. Similarly, even when scientific practice is (as close to) objective (as possible), maybe of a low reliability (i.e., noisy measurement) or lacks validity (i.e., does not measure what it is supposed to measure). According to our notion, validity and reliability, though necessary to guarantee the quality of results, are not necessary for objectivity of the scientific practice that produced the results.

Exploratory research and serendipitous discoveries. Many (if not most of the) famous scientific breakthroughs have been serendipitous discoveries. These discoveries were most likely the product of exploratory research that were neither done by unbiased scientist nor completely free from practices that would now be labeled as 'questionable'. It should be noted that we do not object to these practices and even see them as a vital part of science. However, when it comes to verifying these findings and integrating them in the rest of science, we firmly believe that these discoveries should be tested with a practice that is as objective as possible.

Implications and Applications

The primary implication of our approach to objectivity is that has application in science, in contrast to traditional theories of objectivity. For instance, our notion can be used to assess and address currently salient problems in science i.e., the replication crisis: Harris 2017 and evaluate suggested solutions to problematic scientific practices. Concretely, our conceptualization of objectivity can be captured in an tool that can be tested and calibrated (see Section 7.3).

Increasing objectivity of scientific methods is a necessary step in remedying problems, such as the replication crisis e.g., Harris 2017. The replication crisis is constituted by the fact that results from many scientific experiments are not reproduced in replication studies for a discussion see: Open Science Collaboration 2015; Romero 2016. Concretely, that experiments with similar or identical designs conducted by different scientists (or by the same researchers for the second time) delivered widely different results. The exact percentage of replicablility is unknown, though some indication might be gleaned from large scale replication projects e.g., Open Science Collaboration 2015. In the case of the Open Science Collaboration (2015), hundreds of scientists collaborated to

attempt replication of one-hundred experiments published in prestigious psychological journals. Less than half of the attempts were successful;[10] clearly a disappointing result.

Replicability can be compromised by many factors. One of them is the misuse of degrees of freedom e.g. Simmons, Nelson, and Simonsohn 2011; Wicherts et al. 2016. Specifically, biased studies are more likely to deliver results which fit the particular interest of the scientist (see Section 7.3) or general interest in positive results (see Section 7.3), which therefore will likely disagree with the results of unbiased experiments; decreasing the overall replicability. Now, if the objectivity of scientific practice (i.e., resistance against bias and questionable research practices) is increased, then replicability on any metric will increase. In light of that, increasing objectivity seems to be a necessary steps toward solving the replication crisis and its effectiveness will be clearly observable in the published record of research.

In addition, our notion gives clear indications of which suggested solution to problematic scientific practices will most likely be successful. Some of these restrict scientists directly (e.g., preregistration requirement, random sampling, randomization, etc.), while others make it harder to exploit degrees of freedom (e.g., blind analysis). Because of that, they improve the objectivity to a certain extent. On the other hand, for some of the proposals it is not clear if they are capable of improving objectivity. The *Reformist Package* is and example of such a proposal. It requires that the first author of a paper on a scientific experiment states all potential conflicts of interest. This amounts to explicitly listing all sources of funding which supported his/her work and claiming full responsibility. The Reformist Package has some proponents in scientific literature e.g., Stelfox et al. 1998 and some of the most important scientific journals (e.g., Lancet, Journal of the American Medical Association, etc.) adopted it in their publishing policy. However, according to our conceptualization, it is not clear at all if the proposal improves the objectivity. The Package is forcing scientists to reveal potential causes of systematic bias in the form of financial ties, but it does not safeguard the experiment against actions that can introduce this bias. Our conceptualization predicts that the Reformist Package is ineffective in dealing with the influence funding agencies have, via their researchers, on

[10] A clear and formal definitions of replication is still absent and several benchmarks were used in this paper. On none of them did the replication rate exceed 50%.

the results. This is corroborated by the dissatisfaction concerning its ineffectiveness is common in current literature e.g., Schafer 2004, and is supported by the results of empirical research e.g., Cain, Loewenstein, and Moore 2005.

Finally, our conceptualization of objectivity is compatible with, and follows the spirit of meany traditional theories of objectivity. Our notion is based on the intuition that objectivity is essentially about minimizing the influences that the individual traits of a scientist have on her research (results). This intuition inspired many other conceptualizations of objectivity, for instance, value-free objectivity and procedural objectivity (see Section 7.2). Specifically, the value-free conception of objectivity claims that a scientific justification is objective as long as it is not influenced by non-epistemic values. However, in contrast to our conception, the value-free objectivity is hard to assess in practice and therefore use, because there is no reliable way to know what the motivation was for a given methodological choice. Furthermore, our notion is consistent with all descriptive theories, because we do not claim anything about how the concept is used and understood by scientists or natural language users. Besides, some of these descriptive conceptualizations seem to be based on the above mentioned intuition as well. For example, the epistemic risk account of objectivity of Koskinen 2018, seems to be similar in spirit to our proposal. It claims that objectivity consists in averting epistemic risks arising from imperfections of epistemic agents. Adhering to the recommendations of our proposal averts some of such risks, for example, the risk of delivering a biased result due to study design choices (see Section 7.3). In other words, her description of how objectivity is understood fits our recommendation concerning in which sense science should be objective. Regulatory objectivity, described in Cambrosio et al. 2006, is another example of a descriptive conceptualization based on the same intuition. It is built on the historical analysis of objectivity from Daston and Galison 2010. Regulatory objectivity consists of conventions which aim to ensure research quality, specifically:

> *Regulatory objectivity*, that is based on the systematic recourse to the collective production of evidence. Unlike forms of objectivity that emerged in earlier eras, regulatory objectivity consistently results in the production of conventions, sometimes tacit and unintentional but most often arrived at through concerted programs of action. Cambrosio et al. 2006, p.1

Recent developments are interpreted as the emergence of a new type of objectivity. Implementing and developing such conventions fit our recommendations for the prevention of methodological choices that can bias results or inflate false-positive rates. Again, there is coherence between our normative proposal and the descriptive theory which describes how scientists understand the objectivity.

Conclusions

Let us once again imagine our scientists, Dr. Jane Summers. Dr. Summers is starting a new experiment (e.g., the effects of caffeine on attention, short-term memory, and long-term memory in psychologically healthy adults) but this time she has a grasp on the notion of objectivity and will use (some of) the precautions against objectivity reducing practices. Specifically, when she designs the study, she ensures that for all intents and purposes the participants selection is random from the population of interest (e.g., males and females, age 21 and up that do not suffer from psychological disorders) and that the non-response rate is not biased (e.g., equal non-response in age and gender), that the measurement instruments come with published validation (i.e., standardized test for attention and memory), the participants' allocation to conditions (e.g., coffee with a high dose of caffeine or decaffeinated coffee) is random, and the experiment is double-blinded (i.e., both participant and experimenter are unaware of experiment condition and purpose). Dr. Summers preregisters the study design and the analysis (e.g., structural equation model) of the main effect of interest (e.g., caffeine positively affects long-term memory, mediated by attention and short-term memory). She will have her data blinded and processed by an independent researcher. In addition, she reserves a room for a multiverse analysis. In Dr. Summers' case, not much is known about the complex relation between dependent and independent variables and its mediation or moderation by participant characteristics (e.g., sex, age, daily caffeine consumption, etc.). Thus, apart from the main model suggested by theory and previous research, she wishes to explore other theoretically possible options. Specifically, she performs and reports the results of the analyses of all theoretically possible models and summarizes their results in one or more multiverse analyses. By taking these steps, Dr. Summers restricts many ways in which her study can be biased and thereby improves the objectivity of her

work.

To summarize, in this paper we have argued for a practicable notion of scientific objectivity. We showed that some of the most popular and representative approaches in philosophy cannot be put into practice in their current form. We presented our empirically informed version of via negativa approach to objectivity and conceptualization of objectivity as methodological resilience. Finally, we showed that and how this new conceptualization can plausibly be used by scientists. In the present form, our theory is far from being perfect or complete. At the same time, like science itself, it has the potential to be adjusted and developed to move ever closer to adequacy and completeness.

Chapter 8

Conclusion

In this section, I will conclude my thesis by discussing the obtained results and possible directions for future studies.

8.1 Conditionals and Causal Claims

The first part of my thesis was devoted to the study of indicative conditionals. In the first chapter I critically discussed leading proposals dealing with probability and acceptability of conditionals: The Equation, AT and QAT. I attempted to show that newly reached consensus concerning the validity of those proposals may be premature. Firstly, I show that the AT and QAT lack empirical support and the evidence for The Equation should be reevaluated in light of the results of experiments which include the irrelevant and negatively relevant conditionals. Secondly, I showed that theoretical arguments do not favor any of the theses. The triviality proofs show that it is at least costly to accept any of them. Similarly providing a semantic basis for the theses is problematic. I also pointed to the more attractive alternatives to the theses that were proposed or are at least possible. In light of that, I conclude that we should re-think the role of the theses in future studies of conditionals.

In the second chapter, I showed that the Ramsey Test is in tension with the results of some of the experiments devoted to the acceptability of conditionals. I also analyzed counterarguments to the claim that the presence of positive probabilistic relevance (of antecedent for consequent) is a necessary condition for the acceptability of a conditional and show that neither of them

is convincing.

The third chapter defends the distinction into actual and tendency causal claim which was used in the forth one. I showed that the Minimal Theory of Causation on which the alternative causal distinction is based is problematic. For example, it relates the contents of general and singular causal claims in a unintuitive way. The new distinction is equally problematic for example, it categorizes a claim that describes individual tendency as a general causal claim.

In the fourth chapter, I (together with my coauthors Jan Sprenger and Noah van Dongen) used Hitchcock's distinction between actual and tendency causal claims to formulate several hypotheses concerning the relation between the indicative conditionals and tendency causal claims. The hypotheses were tested and the results show that the causal indicative conditional is evaluated as true if the corresponding causal claim is evaluated as true and the assigned conditional probability of the consequent given antecedent is high.

8.2 Objectivity, Replication and Values

In the second part of my thesis, I discussed the classical problems of philosophy of science, scientific objectivity, and values in science. In my discussion of these issues I related them to recent developments in scientific methodology revolving around the Replication Crisis.

In the fifth chapter, I tried to show that the rejection of VFI forces one to legitimize some of the Questionable Research Practices which are considered to be one of the causes of Replication Crisis. I also showed that the crisis itself seems to be a natural consequence of, at least some versions of, the value-laden ideal of science. Finally, I pointed toward some ways of making the VFI more plausible by defending it against the methodological arguments which show that it is not realizable. Firstly, argued that the unrealizability of ideal does not make it invalid. Secondly, I pointed to some ways in which one can argue for the realizability of VFI.

In the sixth chapter, I (together with my co-author Noah van Dongen) proposed a novel definition of scientific objectivity. Our idea is, roughly, that the scientific practice is objective if and only if it demonstrably does not involve problematic scientific practices like Questionable Research Practices or biased

methodological decisions. Consequently, objectivity consist of methodological rigidity. We also showed which concrete methodological improvements promote objectivity, for example, preregistration or forming new scientific conventions. Finally, we presented a sketch of an instrument for testing scientific objectivity of an experiment, for example, presented in a scientific article.

8.3 Direction for Future Work

Presented results suggest many possible future projects. The first chapter is mainly negative, it argues against the present orthodox view concerning probability and acceptability of indicative conditionals. It suggests at least two projects. Firstly, one may try to defend the received view, such defense would have to include an explanation of results of empirical studies that go against the theses and in the case of The Equation a plausible way out of trivialization. On the other hand, one can and some did present the alternative proposal concerning the probability or acceptability of conditionals. Each of such new proposals opens interesting avenues of research. First of all, such proposals can be empirically tested and, as we have seen, such experiments were presented. Another way to approach the new proposals would be to explore their theoretical standing. One may, for example, explore if a new proposal is susceptible to a trivialization-like argument or if it can be supported by any of the proposed semantics for conditionals. In the case of the second chapter dedicated to RT, the obvious future project would be to test the predictions of RT and RT+ empirically in a study which includes irrelevant and negatively relevant conditionals and compare their performance.

The connection between causal claims and indicative conditionals supported by the result of the experiment presented in the fourth chapter suggests some directions for future work. For example, given that we now know how two types of expressions are related we can use solutions developed in theories of causality as an inspiration for solutions for some of the controversial issues concerning conditionals, for instance, the probability of conditionals.

The argument against value-laden science presented in the fifth section has some obvious limitations. For example, there are some versions of the view which I did not explicitly address. Given the general consensus concerning the implausibility of the VFI it seems that such explicit discussion of those

proposals may be useful. Similarly, a systematic historical study showing how scientific methodology evolves toward the realization of VFI may be useful for making the points made in the chapter more persuasive.

In the sixth section we present a sketch of a survey for assessing the objectivity of a given experiment. In the future, we plan to test its inter-ratter reliability which means to test if different subject gives different or similar scores to the same experiment. Secondly, we want to test if a high score predicts higher replicability rates of other beneficial methodological features.

Appendix A

Instructions for the Experiment

"**Thank you for participating in our study.**

As a Mechanical Turk worker, you should at least have an approval rating of 95% if you want to be reimbursed for your participation.

Please read the questions carefully and answer each with a probability score (between **0%** and **100%**) or truth value ('**True**' or '**False**')

If you have any questions or comments, please leave them at the end of the survey.

You start the survey by pressing the double arrow at the bottom-right corner."

Appendix B

List of Scenarios and Questions

1a. John is a middle-aged man, how likely is it that he will be healthy?

1b. John is a middle-aged man, how likely is it that he will be healthy if he exercises daily?

1c. Daily exercising causes John to be healthy.

1d. If John exercises daily, then he will be healthy.

2a. How likely is it that a random person is skinny?

2b. How likely is it that a random person is skinny if his/her daily food intake is 4 apples and 3 cucumber sandwiches?

2c. Eating only 4 apples and 3 cucumber sandwiches a day causes people to become skinny.

2d. If people only eat 4 apples and 3 cucumber sandwiches a day, then they will become skinny.

3a. How likely is it that a random person will catch the flu?

3b. How likely is it that a random person will catch the flu if two-thirds of his/her co-workers already have it?

3c. Having people around oneself with the flu causes one to catch the flu.

3d. If people around oneself have the flu, then one will catch the flu.

4a. How likely is it that a random person has more than 10 friends?

4b. How likely is it that a random person has more than 10 friends if he/she uses MDMA at parties?

4c. Using MDMA at parties causes people to have more than 10 friends.

4d. If people use MDMA at parties, then they will have more than 10 friends.

5a. How likely is it that the crime rate will decline in the next 5 years?

5b. How likely is it that the crime rate will decline in the next 5 years if drugs (xtc, cocaine, weed) are legalized?

5c. Legalizing drugs (xtc, cocaine, and weed) causes the crime rate to decline over the next 5 years.

5d. If drugs (xtc, cocaine, and weed) are legalized, then the crime rate will decline over the next 5 years.

6a. How likely is it that the crime rate will decline in the next 5 years?

6b. How likely is it that the crime rate will decline in the next 5 years if alcohol consumption is made illegal?

6c. Making alcohol consumption illegal causes the crime rate to decline over the next 5 years.

6d. If alcohol consumption is made illegal, then the crime rate will decline over the next 5 years.

7a. How likely is it that the national birth rate (babies born per capita per year) will decline in the next 10 years?

7b. How likely is it that the national birth rate (babies born per capita per year) will decline in the next 10 years if contraception is mandatory until the age of 21?

7c. Making contraception mandatory until age 21 causes the national birth rate to decline.

7d. If contraception is made mandatory, then the national birth rate will decline.

8a. How likely is it that the national birth rate (babies born per capita per year) will decline in the next 10 years?

8b. How likely is it that the national birth rate (babies born per capita per year) will decline in the next 10 years if all men start wearing white socks in sandals?

8c. All men wearing white socks in sandals causes the national birth rate to decline.

8d. If all men start wearing white socks in sandals, then the national birth rate will decline.

9a. How likely is it that the national birth rate (babies born per capita per year) will decline in the next 10 years?

9b. How likely is it that the national birth rate (babies born per capita per year) will decline in the next 10 years if education is mandatory until age 21?

9c. Making education mandatory until age 21 causes the national birth rate to decline.

9d. If education is mandatory until age 21, then the national birth rate will decline.

10a. How likely is it that a random non-fiction book will be an international best seller (3000 copies sold in the first week)?

10b. How likely is it that a random non-fiction novel will be a best seller (3000 copies sold in the first week) if the author is Stephen Hawking?

10c. Having Stephen Hawking as the author causes a book to be a best seller.

10d. If Stephen Hawking is the author, then a book will be a best seller.

11a. How likely is it that a random non-fiction book will be an international best seller (3000 copies sold in the first week)?

11b. How likely is it that a random non-fiction novel will be a best seller (3000 copies sold in the first week) if the author is Sarah Palin?

11c. Having Sarah Palin as the author causes a book to be a best seller.

11d. If Sarah Palin is the author, then a book will be a best seller.

12a. John is a man in the late sixties, he eats fast food every day and does not exercise. How likely it is that he will develop a cancer?

12b. John is a man in the late sixties, he eats fast food every day and does not exercise. How likely it is that he will develop a cancer if he smokes a pack of cigarettes a day since he was a teenager?

12c. John's smoking will cause him to develop a cancer.

12d. If John smokes, he will develop a cancer.

13a. How likely it is that a random person will develop a cancer?

13b. How likely it is that a random person, exposed to a high dose of gamma radiation, will develop a cancer?

13c. Exposure to a high dose of gamma radiation causes cancer.

13d. If one is exposed to a high dose of gamma radiation, he will develop a cancer.

14a. How likely it is that a random child will be tall?

14b. How likely it is that a random child will be tall, given that she or he is drinking a lot of milk every day?

14c. Drinking a lot of milk everyday causes one to be tall.

14d. If a child drinks a lot of milk every day, he or she will be tall.

15a. How likely it is that a random child will be tall?

15b. How likely it is that a random child will be tall, if she or he is exercising every day?

15c. Exercising causes one to be tall.

15d. If one exercises every day, he or she will be tall.

16a. How likely it is that a random person will become a diabetic?

16b. How likely it is that a random person who eats three apples a day will become a diabetic?

16c. Eating three apples a day cause diabetics.

16d. If one eats three apples a day, he will become a diabetic.

17a. How likely it is that a random person will have all natural teeth at the age of sixty?

17b. How likely it is that a random person will have all natural teeth at the age of sixty, if he or she brushes one's teeth after every meal?

17c. Brushing one's teeth after every meal will cause one to have all natural teeth at the age of sixty.

17d. If one brushes one's teeth after every meal, she or he will have all natural teeth at the age of sixty.

18a. How likely it is that a random person will have all natural teeth at the age of sixty?

18b. How likely it is that a random person will have all natural teeth at the age of sixty, if he or she brushes one's teeth after every meal and visits a dentist once a month?

18c. Brushing one's teeth after every meal and visiting a dentist once a month will cause one to have all natural teeth at the age of sixty.

18d. If one brushes one's teeth after every meal and visits a dentist once a month, she or he will have all natural teeth at the age of sixty.

19a. What is the probability that the approval of the government will decrease in the next few months?

19b. What is the probability that the approval of the government will decrease in the next few months if the majority party proposes legislation that bans alcohol?

19c. A legislation proposal which bans alcohol, made by the majority party, will cause a decrease in approval of the government in the next few months.

19d. If the majority party proposes legislation that bans alcohol, the approval of the government will decrease in the next few months.

Appendix C

A draft of a tool for assessment of objectivity

Given the negative nature of our notion, our checklist consists of questions assessing how susceptible the study is to suffer from the mentioned problematic practices. Concretely, a checklist consists of yes-no questions that indicate the presence or absence of features that prevent problematic practices.

Questions concerning the study being bias-resilient:[1]

1. Was (were) the outcome measure(s) directly related to the phenomenon of interest as stated in the research aim or research question? (e.g., 'death rate' to 'death by cardiac arrest')

2. Was (were) the intervention(s) clearly related to phenomenon of interest as stated in the research aim or research question? (e.g., 'cardiac arrest reducing medication' to 'death by cardiac arrest')

3. Was sampling procedure random?

4. Was the sampling procedure capable of producing a sample representative of the population? (i.e., do inclusion/exclusion criteria allow all member of the population)

[1]These questions pertain only to experimental research. However, the questions can be adapted for observational research.

5. When the subjects are volunteers, was the (non-)response rate similar across participant characteristics? (e.g., equal between men and women)
6. Was the allocation of the subjects to the experiment conditions random?
7. Were both the experimenter and the subjects blind to the experiment condition?
8. Was the drop-out rate of subjects similar across the experiment conditions?
9. If any answer to these question was 'no', were proper steps taken to ameliorate the potential bias that could have resulted from it?

Questions concerning the study being resilient against bias and false-positive rate inflation due to questionable research practices:

1. Was the study preregistered?
2. If so, was the following specified in the preregistration:
   (a) Management of missing and incomplete data.
   (b) Pre-processing of data (e.g., how to clean and normalize).
   (c) Data processing and dealing with violation of statistical assumptions.
   (d) Management of outliers.
   (e) Statistical analysis/model.
   (f) Dependent variable(s) of the model.
   (g) Predictors/covariates of the model.
   (h) Estimation method and computation of standard errors.
   (i) Inference criteria.
3. If preregistered, did the final report conform to this preregistration? Specifically, did the final report conform to the preregistration on:
   (a) Management of missing and incomplete data.
   (b) Pre-processing of data (e.g., how to clean and normalize).

(c) Data processing and dealing with violation of statistical assumptions.

(d) Management of outliers.

(e) Statistical analysis/model.

(f) Dependent variable(s) of the model.

(g) Predictors/covariates of the model.

(h) Estimation method and computation of standard errors.

(i) Inference criteria.

4. If the final report did not completely conform to the preregistration, was the particular deviation handled by blinding data or blinded analysis?

5. If the final report did not completely conform to the preregistration, was the particular deviation handled by using a multiverse analysis and report all (theoretically) possible ways of handling this case?

6. If the study was not preregistered, was blinded data management and analysis used?

7. If blinded data management and analysis was used, was it used on the following:

   (a) Management of missing and incomplete data.

   (b) Pre-processing of data (e.g., how to clean and normalize).

   (c) Data processing and dealing with violation of statistical assumptions.

   (d) Management of outliers.

   (e) Statistical analysis/model.

   (f) Dependent variable(s) of the model.

   (g) Predictors/covariates of the model.

   (h) Estimation method and computation of standard errors.

   (i) Inference criteria.

8. If the study was not preregistered and the data management and analysis was not blinded, was a type of multiverse analysis performed?

9. If a type of multiverse analysis was performed, did it incorporate the following elements:

   (a) Management of missing and incomplete data.

   (b) Pre-processing of data (e.g., how to clean and normalize).

   (c) Data processing and dealing with violation of statistical assumptions.

   (d) Management of outliers.

   (e) Statistical analysis/model.

   (f) Dependent variable(s) of the model.

   (g) Predictors/covariates of the model.

   (h) Estimation method and computation of standard errors.

   (i) Inference criteria.

It needs to be noted that such a tool needs to be further developed, tested, and calibrated. To be useful, this objectivity checklist should of course be (to a large extent) reliable and valid. In the first case, inter-rater reliability and intra-rater reliability should be assessed. I.e., have different subjects use the tool and measure the agreement between their results Cohen 1960; Fleiss 1971, and have subjects use the tool on two or more different occasions on the same material and measure the similarity of results between these occasions Gwet 2008. Close similarity between scores indicate that users will in general give similar scores when using the checklist. If there are questions in the checklist that score low on inter-rater and/or intra-rater reliability, then they should be rephrased or dropped. The validity of the checklist can be assessed by empirically verifying if research that scores high on the checklist are less prone to produce problematic results (e.g., have a higher replication rate) in comparison to research that score low on the checklist (i.e., criterion validity). For now, we do not have a scoring system of the checklist. The easiest scoring system would be to use the proportion of 'yes' answers of the total number of questions that are relevant for the report that is evaluated. However, this scoring system should be further developed and tested. Also, scientific practice could be simulated to assess to what extent bias and false-positive rate inflation could still be introduced with varying levels of objectivity according to the checklist. Results

from such simulation studies could also be used to calibrate the scoring system. Finally, scientists could be enlisted to perform mock research to attempt to bias outcomes and/or produce false-positive results with varying levels of objectivity safeguards in play. These mock studies might identify weaknesses and gaps in the tool (and objectivity conceptualization), which could be used when calibration the tool and supplementing/removing elements.

Bibliography

Adams, Ernest W. 1965. "The Logic of Conditionals." *Inquiry* (Dordrecht) 8:166–197.

— . 1975. *The Logic of Conditionals: An Application of Probability to Deductive Logic*. D. Reidel Pub. Co.

— . 1998. *A Primer of Probability Logic*. Stanford: Csli Publications.

Altmejd, Adam, et al. 2019. "Predicting the Replicability of Social Science Lab Experiments."

Anderson, Elizabeth. 2004. "Uses of Value Judgments in Science: A General Argument, with Lessons From a Case Study of Feminist Research on Divorce." *Hypatia* 19 (1): 1–24. doi:10.2979/HYP.2004.19.1.1.

Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–554. doi:10.1177/1745691612459060.

Baratgin, Jean, et al. 2018. "The psychology of uncertainty and three-valued truth tables." *Frontiers in psychology.* 9 (): 1479. http://dro.dur.ac.uk/26304/.

Bates, Douglas, et al. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. doi:10.18637/jss.v067.i01.

Bavel, Jay J van, et al. 2016. "Contextual sensitivity in scientific reproducibility." *Proceedings of the National Academy of Sciences* 113 (23): 6454–6459.

Belnap, Nuel D. 1970. "Conditional Assertion and Restricted Quantification." *Noûs* 4 (1): 1–12. doi:10.2307/2214285.

Bem, Daryl J. 2011. "Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect." *Journal of personality and social psychology* 100 (3): 407.

Bennett, Jonathan. 2003. *A Philosophical Guide to Conditionals*. Oxford University Press.

Betz, Gregor. 2013. "In Defence of the Value Free Ideal." *European Journal for Philosophy of Science* 3 (2): 207–220.

Bishop, D. V. M. 1990. "How to increase your chances of obtaining a significant association between handedness and disorder." PMID: 1701771, *Journal of Clinical and Experimental Neuropsychology* 12 (5): 812–816. doi:10.1080/01688639008401022. eprint: https://doi.org/10.1080/01688639008401022. https://doi.org/10.1080/01688639008401022.

Bradley, Richard. 2000. "A Preservation Condition for Conditionals." *Analysis* 60 (3): 219–222.

Brown, Matthew. 2013. "The source and status of values for socially responsible science." *Philosophical Studies* 163 (). doi:10.1007/s11098-012-0070-x.

Byrne, Ruth, and P.N. Johnson-Laird. 2009. "'If' and the problems of conditional reasoning." *Trends in cognitive sciences* 13 (): 282–7. doi:10.1016/j.tics.2009.04.003.

Cain, Daylian M., George Loewenstein, and Don A. Moore. 2005. "The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest." *The Journal of Legal Studies* 34 (1): 1–25.

Cambrosio, Alberto, et al. 2006. "Regulatory objectivity and the generation and management of evidence in medicine." *Social Science & Medicine* 63 (1): 189–199. ISSN: 0277-9536. doi:https://doi.org/10.1016/j.socscimed.2005.12.007.

Cammack, Mark. 1994. "In Search of the Post-Positivist Jury." *Indiana Law Journal* 70 ().

Cantwell, John. 2008. "The Logic of Conditional Negation." *Notre Dame Journal of Formal Logic* 49 (3): 245–260. doi:10.1215/00294527-2008-010.

Carlstrom, Ian F., and Christopher S. Hill. 1978. "Book Review:The Logic of Conditionals Ernest W. Adams." *Philosophy of Science* 45 (1): 155–.

Cartwright, Nancy. 1979. "Causal Laws and Effective Strategies." *11* 13, no. 4 (): 419–437.

Chambers, Christopher D. 2013. "Registered reports: a new publishing initiative at Cortex." *Cortex* 49 (3): 609–610.

Cheng, Patricia W. 1997. "From Covariation to Causation: A Causal Power Theory." *Psychological Review* 104:367–405.

ChoGlueck, Christopher. 2018. "The Error Is in the Gap: Synthesizing Accounts for Societal Values in Science." *Philosophy of Science* 85 (4): 704–725. doi:10.1086/699191.

Cohen, Jacob. 1960. "A coefficient of agreement for nominal scales." *Educational and psychological measurement* 20 (1): 37–46.

— . 1988. *Statistical Power Analysis for the Behavioral Sciences*. Newark, N.J.: Lawrence & Erlbaum.

Cooper, William S. 1968. "The Propositional Logic of Ordinary Discourse." *Inquiry: An Interdisciplinary Journal of Philosophy* 11 (1-4): 295–320. doi:10.1080/00201746808601531.

Crupi, Vincenzo. 2015. "Confirmation." In *The Stanford Encyclopedia of Philosophy*, ed. by Ed Zalta. Retrieved from https://plato.stanford.edu/archives/sum2018/entries/confirmation/. https://plato.stanford.edu/entries/confirmation/.

Crupi, Vincenzo, and Andrea Iacona. 2019a. *The evidential conditional*. http://philsci-archive.pitt.edu/16479/.

— . 2019b. "Three Ways of Being Non-Material." Unpublished manuscript.

Crupi, Vincenzo, and Katya Tentori. 2013. "Confirmation as Partial Entailment: A Representation Theorem in Inductive Logic." *Journal of Applied Logic* 11:364–372.

Crupi, Vincenzo, Katya Tentori, and Michel González. 2007. "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues." *Philosophy of Science* 74:229–252. ISSN: 0031-8248.

Cruz, Nicole, et al. 2016. "Centering and the Meaning of Conditionals."

Daston, Lorraine, and Peter Galison. 2010. *Objectivity*. New York: Zone Books.

Declerck, Renaat, and Susan Reed. 2012. *Conditionals. A Comprehensive Empirical Analysis*. Mouton de Gruyter.

Douglas, Heather. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138 (3): 453–473.

— . 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Douven, Igor. 2008. "The Evidential Support Theory of Conditionals." *Synthese* 164 (1): 19–44.

— . 2015. *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches*. Cambridge University Press.

— . 2016. "Experimental Approaches to the Study of Conditionals." In *Companion to Experimental Philosophy*.

— . 2017. "How to account for the oddness of missing-link conditionals." *Synthese* 194, no. 5 (): 1541–1554. ISSN: 1573-0964. doi:10.1007/s11229-015-0756-7. https://doi.org/10.1007/s11229-015-0756-7.

Douven, Igor, and Karolina Krzyzanowska. 2018. "The Semantics-Pragmatics Interface." In *Further Advances in Pragmatics and Philosophy*, ed. by Alessandro Capone, vol. 2. Springer.

Douven, Igor, and Sara Verbrugge. 2010. "The Adams Family." *Cognition* 117 (3): 302–318.

— . 2012. "Indicatives, concessives, and evidential support." *Thinking and reasoning* 18 (4): 480–499.

— . 2013. "The Probabilities of Conditionals Revisited." *Cognitive Science* 37 (4): 711–730.

Douven, Igor, et al. 2019. "Conditionals and inferential connections: toward a new semantics." *Thinking and Reasoning* 0 (0): 1–41. doi:10.1080/13546783.2019.1619623. eprint: https://doi.org/10.1080/13546783.2019.1619623. https://doi.org/10.1080/13546783.2019.1619623.

Downs, S.H., and N Black. 1998. "The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions." *Journal of epidemiology and community health* 52 (): 377–84. doi:10.1136/jech.52.6.377.

Du Bois, W. E. Burghardt. 1898. "The Study of the Negro Problems." *The Annals of the American Academy of Political and Social Science* 11:1–23. ISSN: 00027162. http://www.jstor.org/stable/1009474.

Dubois, Didier, and Henri Prade. 1990. "The logical view of conditioning and its application to possibility and evidence theories." *International Journal of Approximate Reasoning* 4 (1): 23–46. ISSN: 0888-613X. doi:https://doi.org/10.1016/0888-613X(90)90007-O. http://www.sciencedirect.com/science/article/pii/0888613X90900070.

Dupré, John. 2007. "Fact and Value," 27–41. doi:10.1093/acprof:oso/9780195308969.003.0003.

Edgington, Dorothy. 1995. "On Conditionals." *Mind* 104 (414): 235–329.

Eells, Ellery. 1991. *Probabilistic Causality*. Cambridge University Press.

Egré, Paul, and Mikaël Cozic. 2011. "If-clauses and probability operators." *Topoi* 30 (1): 17.

Egré, Paul, Lorenzo Rossi, and Jan Sprenger. 2019. "De Finettian Logics of Indicative Conditionals."

Elliott, Kevin. 2011. "Is a Little Pollution Good for You?: Incorporating Societal Values in Environmental Research" (): 1–264.

Eva, Benjamin, Stephan Hartmann, and Soroush Rafiee Rad. 2019. "Learning from Conditionals." Fzz025, *Mind* (). ISSN: 0026-4423. doi:10.1093/mind/fzz025. eprint: http://oup.prod.sis.lan/mind/advance-article-pdf/doi/10.1093/mind/fzz025/28843008/fzz025.pdf. https://doi.org/10.1093/mind/fzz025.

Evans, Jonathan, Simon Handley, and David Over. 2003. "Conditionals and conditional probability." *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Evans, Jonathan, and David Over. 2004. *If*. Oxford University Press.

Evans, Jonathan, et al. 2007. "Thinking about conditionals: A study of individual differences." *Memory & Cognition* 35:1772–1784.

Fanelli, Daniele, Rodrigo Costas, and John P. A. Ioannidis. 2017. "Meta-assessment of bias in science." *Proceedings of the National Academy of Sciences* 114 (14): 3714–3719. ISSN: 0027-8424. doi:10.1073/pnas.1618569114. eprint: https://www.pnas.org/content/114/14/3714.full.pdf. https://www.pnas.org/content/114/14/3714.

Finetti, Bruno de. 1936. "La logique de la probabilité." *Actes du congrés international de philosophie scientifique*.

— . 1980. *Probabilità [Probability]*. Encyclopedia.

Fitelson, Branden. 1999. "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity." *Philosophy of Science* 66:S362–S378.

— . 2001. "Studies in Bayesian Confirmation Theory." PhD thesis, University of Wisconsin–Madison.

— . 2015. "The Strongest Possible Lewisian Triviality Result." *Thought: A Journal of Philosophy* 4 (2): 69–74. doi:10.1002/tht3.159.

Fitelson, Branden, and Christopher Hitchcock. 2011. "Probabilistic Measures of Causal Strength." In *Causality in the Sciences*, ed. by Phyllis McKay Illari, Federica Russo, and Jon Williamson, 600–627. Oxford: Oxford University Press.

Fleiss, Joseph L. 1971. "Measuring nominal scale agreement among many raters." *Psychological bulletin* 76 (5): 378.

Foley, Richard. 2009. "Beliefs, Degrees of Belief, and the Lockean Thesis." In *Degrees of Belief*, ed. by Franz Huber and Christoph Schmidt-Petri, 37–47. Springer.

Frosch, Caren A., and Ruth M.J. Byrne. 2012. "Causal conditionals and counterfactuals." *Acta Psychologica* 141 (1): 54–66. ISSN: 0001-6918. doi:http://dx.doi.org/10.1016/j.actpsy.2012.07.001. http://www.sciencedirect.com/science/article/pii/S0001691812001126.

Fugard, Andy, Niki Pfeifer, and Bastian Mayerhofer. 2011. "Probabilistic theories of reasoning need pragmatics too: Modulating relevance in uncertain conditionals." *Journal of Pragmatics* 43 (): 2034–2042. doi:10.1016/j.pragma.2010.12.009.

Fugard, Andy, et al. 2011. "How people interpret conditionals: Shifts toward the conditional event." *Journal of Experimental Psychology Learning Memory and Cognition* 37 (): 635–48.

Fuhrmann, André, and Isaac Levi. 1994. "Undercutting and the Ramsey Test for Conditionals." *Synthese* 101 (2): 157–169.

Gärdenfors, Peter. 1986. "Belief Revisions and the Ramsey Test for Conditionals." *Philosophical Review* 95 (1): 81–93.

Gibbard, Allan. 1981. "Two Recent Theories of Conditionals." In *Ifs*, ed. by William Harper, Robert C. Stalnaker, and Glenn Pearce, 211–247. Reidel.

Gigerenzer, Gerd. 2004. "Mindless statistics." *The Journal of Socio-Economics* 33 (5): 587–606.

Gobert, James J. 1988. "In Search of the Impartial Jury." *J. Crim. L. & Criminology*.

Goldacre, Ben. 2014. *Bad pharma: how drug companies mislead doctors and harm patients*. Macmillan.

Goodman, I., Hung Nguyen, and Elbert Walker. 1991. *Conditional Inference and Logic for Intelligent Systems: A Theory of Measure-Free Conditioning*. ISBN: 978-0-444-88685-9.

Graßhoff, Gerd, and Michael May. 2001. "Causal Regularities." In *Current Issues in Causation*, ed. by Wolfgang Spohn, Marion Ledwig, and Michael Esfeld, 85. Mentis.

Gwet, Kilem. 2008. "Intrarater Reliability." *Methods and Applications of Statistics in Clinical Trials* 2:473–485. doi:10.1002/9780471462422.eoct631.

Hacking, Ian. 2015. "Let Not Talk About Objectivity." In *Objectivity in Science*, ed. by Jonathan Y. Tsou, Alan Richardson, and Flavia Padovani. Springer Verlag.

Hadjichristidis, Constantinos, et al. 2001. "On the Evaluation of *If p then q* Conditionals." In *Proceedings ofthe 23rd Annual Meeting ofthe Cognitive Science Society*.

Haidich, AB. 2010. "Meta-analysis in medical research." *Hippokratia* 14 (Suppl 1): 29–37.

Hájek, Alan. 1989. "Probabilities of Conditionals — Revisited." *Journal of Philosophical Logic* 18 (4): 423–428.

— . 1994. "Triviality on the Cheap?" In *Probability and Conditionals: Belief Revision and Rational Decision*, ed. by Ellery Eells, Brian Skyrms, and Ernest W. Adams, 113–40. Cambridge University Press.

— . 2012. "The Fall of "Adams' Thesis"?" *Journal of Logic, Language and Information* 21 (2): 145–161.

Hájek, Alan, and N. Hall. 1994. "The Hypothesis of the Conditional Construal of Conditional Probability." In *Probability and Conditionals: Belief Revision and Rational Decision*, ed. by Ellery Eells, Brian Skyrms, and Ernest W. Adams, 75. Cambridge University Press.

Harris, R. F. 2017. *Rigor mortis: how sloppy science creates worthless cures, crushes hope, and wastes billions*. New York: Basic Books.

Hawkins, Carlee Beth, and Brian A Nosek. 2012. "Motivated independence? Implicit party identity predicts political judgments among self-proclaimed independents." *Personality and Social Psychology Bulletin* 38 (11): 1437–1452.

Head, Megan L., et al. 2015. "The Extent and Consequences of P-Hacking in Science." *PLOS Biology* 13, no. 3 (): 1–15. doi:10.1371/journal.pbio.1002106. https://doi.org/10.1371/journal.pbio.1002106.

Heide, Rianne de, and Peter Grünwald. 2017. "Why optional stopping is a problem for Bayesians" ().

Hicks, Daniel J. 2014. "A New Direction for Science and Values." *Synthese* 191 (14): 3271–95. doi:10.1007/s11229-014-0447-9.

Hirvonen, Sanna, Natalia Karczewska, and Michał Sikorski. 2019. "On Hybrid Expressivism about Aesthetic Judgments." *Grazer Philosophische Studien* (Leiden, The Netherlands) 96 (4). https://brill.com/view/journals/gps/96/4/article-p541_541.xml.

Hitchcock, Christopher. 2001. "Causal Generalizations and Good Advice." *The Monist* 84 (2): 218–241.

Hobson, R. M., et al. 2012. "Effects of β-alanine supplementation on exercise performance: a meta-analysis." *Amino Acids* 43, no. 1 (): 25–37. ISSN: 1438-2199. doi:10.1007/s00726-011-1200-z. https://doi.org/10.1007/s00726-011-1200-z.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2, no. 8 (): 55–69.

Jackson, Frank. 1987. *Conditionals*. Blackwell.

Jadad, Alejandro R, and Laura O'Grady. 2008. "How should health be defined?" *BMJ: British Medical Journal (Online)* 337.

Jakob, Christian. 2006. "Hitchcock's (2001) Treatment of Singular and General Causation." *Minds and Machines* 16 (3): 277–287.

Jeffrey, Richard C. 1956. "Valuation and Acceptance of Scientific Hypotheses." *Philosophy of Science* 23 (3): 237–246.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–532. ISSN: 14679280. doi:10.1177/0956797611430953. arXiv: arXiv:1011.1669v3.

John, Stephen. 2015. "Inductive Risk and the Contexts of Communication." *Synthese* 192 (1): 79–96. doi:10.1007/s11229-014-0554-7.

Johnson-Laird, P, and Ruth Byrne. 2002. "Conditionals: A Theory of Meaning, Pragmatics, and Inference." *Psychological review* 109 (): 646–78. doi:10.1037//0033-295X.109.4.646.

Jones, Martin, and Robert Sugden. 2001. "Positive Confirmation Bias in the Acquisition of Information." *Theory and Decision* 50 (1): 59–99.

Kass, Robert E., et al. 2016. "Ten Simple Rules for Effective Statistical Practice." *PLoS Computational Biology* 12 (6): e1004961. doi:10.1371/journal.pcbi.1004961.

Kolbel, Max. 1997. "Expressivism and the syntactic uniformity of declarative sentences." *Crítica; revista hispanoamericana de filosofía* 29 (): 3–51.

— . 2000. "Edgington on Compounds of Conditionals." *Mind* 109 (433): 97–108. doi:10.1093/mind/109.433.97.

Koskinen, Inkeri. 2018. *Defending a Risk Account of Scientific Objectivity*. http://philsci-archive.pitt.edu/14890/.

Kratzer, Angelika. 1989. "An Investigation of the Lumps of Thought." *Linguistics and Philosophy* 12 (): 607–653. doi:10.1007/BF00627775.

Krzyzanowska, Karolina. 2015. *Between "If" and "Then". Towards an empirically informed philosophy of conditionals.* PhD thesis, University of Groningen.

Krzyżanowska, Karolina, Peter J. Collins, and Ulrike Hahn. 2017. "Between a Conditional's Antecedent and its Consequent: Discourse Coherence Vs. Probabilistic Relevance." *Cognition* 164:199–205. doi:10.1016/j.cognition.2017.03.009.

Krzyżanowska, Karolina, Sylvia Wenmackers, and Igor Douven. 2014. "Rethinking Gibbard's Riverboat Argument." *Studia Logica* 102 (4): 771–792.

LeBel, Etienne, et al. 2018. "A unified framework to quantify the credibility of scientific findings." 3 (): 389–402.

Levi, Isaac. 1960. "Must the Scientist Make Value Judgments?" *Journal of Philosophy* 57 (11): 345–357.

Lewis, David. 1973a. "Causation." *Journal of Philosophy* 70 (17): 556–567.

— . 1973b. *Counterfactuals*. Oxford: Blackwell.

— . 1976. "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review* 85 (3): 297.

Lilienfeld, Scott O. 2012. "Public skepticism of psychology: why many people perceive the study of human behavior as unscientific." *American Psychologist* 67 (2): 111.

Lindsay, D Stephen. 2015. *Replication in psychological science*.

Lindström, Sten. 1996. "The Ramsey Test and the Indexicality of Conditionals: A Proposed Resolution of Gärdenfors' Paradox." In *Logic, Action and Information*, ed. by André Fuhrmann and Hans Rott. de Gruyter.

Longino, Helen. 2004. "How Values Can Be Good for Science." In *Science, Values, and Objectivity*, ed. by Peter K. Machamer and Gereon Wolters, 127–142. University of Pittsburgh Press.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Ludwig, David. 2015. "Ontological Choices and the Value-Free Ideal." *Erkenntnis*, no. 6: 1–20.

Lycan, William G. 2003. "Real Conditionals." *Philosophical Quarterly* 53 (210): 134–137.

MacCoun, Robert. 1998. "Biases in the Interpretation and Use of Research Results." *Annual review of psychology* 49:259–87.

MacCoun, Robert, and Saul Perlmutter. 2015. "Blind analysis: hide results to seek the truth." *Nature News* 526 (7572): 187.

Mackie, J. L. 1980. *The Cement of the Universe: A Study of Causation*. Clarendon Press.

Marini, Margaret Mooney, and Burton Singer. 1988. "Causality in the Social Sciences." *Sociological Methodology* 18:347–409. ISSN: 00811750, 14679531. http://www.jstor.org/stable/271053.

McGee, Vann. 1989. "Conditional Probabilities and Compounds of Conditionals." *Philosophical Review* 98 (4): 485–541.

Megill, Allan. 1994. "Four Senses of Objectivity." In *Rethinking Objectivity*.

Milne, Peter. 2003. "The Simplest Lewis-Style Triviality Proof Yet?" *Analysis* 63 (). doi:10.1093/analys/63.4.300.

Montori, Victor M., et al. 2005. "Randomized Trials Stopped Early for BenefitA Systematic Review." *JAMA* 294, no. 17 (): 2203–2209. ISSN: 0098-7484. doi:10.1001/jama.294.17.2203. eprint: https://jamanetwork.com/journals/jama/articlepdf/201802/jrv50019.pdf. https://dx.doi.org/10.1001/jama.294.17.2203.

Murphy, Kevin R., and Herman Aguinis. 2019. "HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results?" *Journal of Business and Psychology* 34, no. 1 (): 1–17. ISSN: 1573-353X. doi:10.1007/s10869-017-9524-7. https://doi.org/10.1007/s10869-017-9524-7.

National Toxicology Program, US Department of Health and Human Services. 2001. *National Toxicology Program's report of the endocrine disruptors low-dose*. http://ntp-server.niehs.nih.gov/ntp/htdocs/liason/LowDosePeerFinalRpt.pdf.

Nelson, Julie A. 2014. "The Power of Stereotyping and Confirmation Bias to Overwhelm Accurate Assessment: The Case of Economics, Gender, and Risk Aversion." *Journal of Economic Methodology* 21 (3): 211–231.

Nelson, Leif D, Joseph Simmons, and Uri Simonsohn. 2018. "Psychology's Renaissance." *Annual Review of Psycholgy* 69:1–24. doi:https://doi.org/10.1146/annurev-psych-122216-011836.

Nosek, Brian A, et al. 2018. "The preregistration revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115 (11): 2600–2606. doi:10.1073/pnas.1708274114.

Oberauer, Klaus, and Oliver Wilhelm. 2003. "The meaning (s) of conditionals: Conditional probabilities, mental models, and personal utilities." *Journal of Experimental Psychology: Learning Memory and Cognition* 29:680–693.

Okruhlik, Kathleen. 1994. "Gender and the Biological Sciences." *Canadian Journal of Philosophy* 24 (sup1): 21–42.

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." *Science* 349 (6251): aac4716.

Oreskes, Naomi, and Erik M. Conway. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues From Tobacco Smoke to Global Warming*. Bloomsbury Press.

Organization, World Health, et al. 1950. *The preamble of the constitution of the World Health Organization*.

Over, David. 2016. "Causation and the Probability of Causal Conditionals." In *Oxford Handbook of Causal Reasoning*, ed. by Michael Waldmann. Oxford: Oxford University Press.

Over, David, and Jean Baratgin. 2016. "The "defective" truth table: its past, present, and future." In *The Thinking Mind: A Festschrift for Ken Manktelow*.

Over, David, and Nicole Cruz. 2018. "Probabilistic accounts of conditional reasoning."

Over, David, et al. 2007. "The probability of causal conditionals." *Cognitive psychology* 54 (): 62–97. doi:10.1016/j.cogpsych.2006.05.002.

Pearl, Judea. 2000. *Causality*. Cambridge: Cambridge University Press.

— . 2001. "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ed. by Jack Breese and Daphne Koller, 411–420. ISBN: 1558608001.

— . 2009. *Causality*. Cambridge University Press.

Pierre, Barrouillet, and Caroline Gauffroy. 2015. "Probability in reasoning: A developmental test on conditionals." *Cognition* 137 (). doi:10.1016/j.cognition.2014.12.002. http://gen.lib.rus.ec/scimag/index.php?s=10.1016/j.cognition.2014.12.002.

Proust, Joëlle. 2012. "The Norms of Acceptance." *Philosophical Issues* 22 (1): 316–333·

Ramsey, Frank P. 1926. "Truth and Probability." In *Philosophical Papers*, ed. by D H Mellor, 52–94. Cambridge: Cambridge University Press.

Ramsey, Frank Plumpton. 1990. "General propositions and causality." In *Philosophical papers*, ed. by David Hugh Mellor. Cambridge University Press.

Reiss, Julian, and Jan Sprenger. 2017. "Scientific Objectivity." In *The Stanford Encyclopedia of Philosophy*, Winter 2017, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Romero, Felipe. 2016. "Can the Behavioral Sciences Self-Correct? A Social Epistemic Study." *Studies in History and Philosophy of Science Part A* 60:55–69.

— . 2017. "Novelty Versus Replicability: Virtues and Vices in the Reward System of Science." *Philosophy of Science* 84 (5): 1031–1043.

Rott, Hans. 1986. "Ifs, Though, and Because." *Erkenntnis* 25 (3): 345–370. doi:10.1007/BF00175348.

Rudner, Richard. 1953. "The Scientist Qua Scientist Makes Value Judgments." *Philosophy of Science* 20 (1): 1–6.

Schafer, A. 2004. "Biomedical conflicts of interest: a defence of the sequestration thesis—learning from the cases of Nancy Olivieri and David Healy." *Journal of Medical Ethics* 30 (1): 8–24. ISSN: 0306-6800. doi:10.1136/jme.2003.005702. eprint: https://jme.bmj.com/content/30/1/8.full.pdf. https://jme.bmj.com/content/30/1/8.

Schulz, Katrin. 2011. ""If You'd Wiggled A, Then B Would've Changed"." *Synthese* 179 (2): 239–251.

Schulz, Kenneth F, Douglas G Altman, and David Moher. 2010. "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials." *BMJ* 340 (). doi:10.1136/bmj.c332.

Schulz, Moritz. 2017. *Counterfactuals and Probability*. Oxford: Oxford University Press.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological science* 22 (11): 1359–1366.

Simons, Daniel J. 2014. "The value of direct replication." *Perspectives on Psychological Science* 9 (1): 76–80.

Sindhu, Fahera, Lucy Carpenter, and Kate Seers. 1997. "Dovelopment of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique." *Journal of Advanced Nursing* 25 (6): 1262–1268. doi:10.1046/j.1365-2648.1997.19970251262.x.

Skovgaard-Olsen, Niels, Henrik Singmann, and Karl Christoph Klauer. 2016. "The relevance effect and conditionals." 1, *Cognition* 150:26–36. ISSN: 0010-0277.

— . 2017. "Relevance and Reason Relations." *Cognitive Science* 41 (S5): 1202–1215. doi:10.1111/cogs.12462. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12462. https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12462.

Skovgaard-Olsen, Niels, et al. 2017. "Relevance differently affects the truth, acceptability, and probability evaluations of 'and', 'but', 'therefore', and 'if then'." *Thinking and Reasoning* 23 (). doi:10.1080/13546783.2017.1374306.

Skovgaard-Olsen, Niels, et al. 2019. "Cancellation, Negation, and Rejection." *Cognitive Psychology* 108:42–71.

Sloman, Steven A., and David Lagnado. 2015. "Causality in Thought." *Annual Review of Psychology* 66:223–247.

Sober, Elliott. 2007. "Evidence and Value Freedom."

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. 2nd. New York: Springer.

Spohn, Wolfgang. 2012. *The Laws of Belief: Ranking Theory and its Philosophical Applications*. Oxford University Press.

Sprenger, Jan. 2015. "Conditional Degree of Belief."

— . 2018. "Foundations for a Probabilistic Theory of Causal Strength." *Philosophical Review* 127:371–398.

Sprenger, Jan, and Stephan Hartmann. 2019. *Bayesian Philosophy of Science*. oxford university press.

Stalnaker, Robert. 1968. "A Theory of Conditionals." In *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*, ed. by Nicholas Rescher, 98–112. Oxford: Blackwell.

— . 1975. "Indicative Conditionals." *Philosophia* 5 (3): 269–286.

— . 1976. "Stalnaker to Van Fraassen." In *Foundations of probability theory, statistical inference, and statistical theories of science*, ed. by W. Harper C. Hooker.

Stapel, Diederik. 2012. *Ontsporing*. Prometheus Amsterdam.

Stark, Philip B, and Andrea Saltelli. 2018. "Cargo-cult statistics and scientific crisis." *Significance* 15 (4): 40–43.

Steegen, Sara, et al. 2016. "Increasing transparency through a multiverse analysis." *Perspectives on Psychological Science* 11 (5): 702–712.

Steel, Daniel. 2010. "Epistemic Values and the Argument From Inductive Risk." *Philosophy of Science* 77 (1): 14–34. doi:10.1086/650206.

Stelfox, Henry Thomas, et al. 1998. "Conflict of Interest in the Debate over Calcium-Channel Antagonists." PMID: 9420342, *New England Journal of Medicine* 338 (2): 101–106. doi:10.1056/NEJM199801083380206. eprint: https://doi.org/10.1056/NEJM199801083380206. https://doi.org/10.1056/NEJM199801083380206.

Stroebe, Wolfgang, and Fritz Strack. 2014. "The Alleged Crisis and the Illusion of Exact Replication." PMID: 26173241, *Perspectives on Psychological Science* 9 (1): 59–71. doi:10.1177/1745691613514450. eprint: https://doi.org/10.1177/1745691613514450. https://doi.org/10.1177/1745691613514450.

Suppes, Patrick. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Szucs, Denes. 2016. "A Tutorial on Hunting Statistical Significance by ChasingN." *Frontiers in psychology* 7:1444. doi:10.3389/fpsyg.2016.01444.

Talbott, William. 2016. "Bayesian Epistemology." In *The Stanford Encyclopedia of Philosophy*, Winter 2016, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Tsui, Anne. 2016. "Reflections on the so-called value-free ideal: A call for responsible science in the business schools." *Cross Cultural & Strategic Management* 23 (1): 4–28. doi:10.1108/CCSM-08-2015-0101. eprint: https://doi.org/10.1108/CCSM-08-2015-0101. https://doi.org/10.1108/CCSM-08-2015-0101.

UNODC. 2019. *Commentary on the Bangalore Principles of Judicial Conduct*. United Nations Office on Drugs / Crime. eprint: https://www.unodc.org/res/ji/import/international_standards/commentary_on_the_bangalore_principles_of_judicial_conduct/bangalore_principles_english.pdf.

van Fraassen, Bas. 1976. "Probabilities of Conditionals." In *Foundations of probability theory, statistical inference, and statistical theories of science*, ed. by W. Harper C. Hooker.

Vazire, Simine. 2016. "Editorial." *Social Psychological and Personality Science* 7 (1): 3–7. doi:10.1177/1948550615603955. eprint: https://doi.org/10.1177/1948550615603955. https://doi.org/10.1177/1948550615603955.

Vidal, Mathieu, and Jean Baratgin. 2017. "A psychological study of unconnected conditionals." *Journal of Cognitive Psychology* 29 (6): 769–781. doi:10.1080/20445911.2017.1305388. eprint: https://doi.org/10.1080/20445911.2017.1305388. https://doi.org/10.1080/20445911.2017.1305388.

Warr, Jason. 2016. *An Introduction to Criminological Theory and the Problem of Causation*. doi:10.1007/978-3-319-47446-5.

Weber, Max. 1949. "'Objectivity' in Social Science and Social Policy." In *The Methodology of the Social Sciences*.

Wicherts, Jelte M, et al. 2016. "Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking." *Frontiers in Psychology* 7:1832.

Wijnbergen-Huitink, Janneke van, Shira Elqayam, and David Over. 2015. "The Probability of Iterated Conditionals." *Cognitive science* 39 4:788–803.

Wilholt, Torsten. 2008. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science Part A* 40 (1): 92–101.

Woodward, James. 2005. "Making Things Happen." *Philosophical Review* 114 (4): 545–547. doi:10.1215/00318108-114-4-545.

Wright, Jack. 2018. "Rescuing Objectivity: A Contextualist Proposal." *Philosophy of the Social Sciences* 48 (4): 385–406. doi:10.1177/0048393118767089. eprint: https://doi.org/10.1177/0048393118767089. https://doi.org/10.1177/0048393118767089.

Ziman, John. 1996. "Is science losing its objectivity?" *Nature* 382 (): 751–754. doi:10.1038/382751a0.