Learning, Conditionals, Causation

Mario Konrad Günther

Graduate School of
Systemic Neurosciences
LMU Munich

Learning, Conditionals, Causation

Mario Konrad Günther

Dissertation at the
Graduate School of Systemic Neurosciences
LMU Munich

submitted by
Mario Konrad Günther
born in Laupheim

Munich 2018

First Reviewer:	Prof. DDr. Hannes Leitgeb
Second Reviewer:	Prof. Dr. Stephan Hartmann

Date of Submission:	October 2, 2018
Date of Defense:	January 24, 2019

First Supervisor:	Prof. DDr. Hannes Leitgeb Chair of Logic and Philosophy of Language Co-Director of the Munich Center for Mathematical Philosophy LMU Munich
Second Supervisor:	Prof. Dr. Stephan Hartmann Chair of Philosophy of Science Co-Director of the Munich Center for Mathematical Philosophy LMU Munich
Third Supervisor:	Prof. Dr. Stephan Sellmaier Head of the Research Center for Neurophilosophy and Ethics of Neurosciences LMU Munich
Fourth Supervisor:	Prof. Dr. Holger Andreas Department of Economics, Philosophy, and Political Science University of British Columbia

First Reviewer:	Prof. DDr. Hannes Leitgeb
Second Reviewer:	Prof. Dr. Stephan Hartmann

Third Examiner:	Prof. Dr. Andreas Herz
Fourth Examiner:	Prof. Dr. Igor Douven

Für Maria und Konrad

Abstract

This dissertation is on conditionals and causation. In particular, we (i) propose a method of how an agent learns conditional information, and (ii) analyse causation in terms of a new type of conditional. Our starting point is Ramsey's (1929/1990) test: accept a conditional when you can infer its consequent upon supposing its antecedent. Inspired by this test, Stalnaker (1968) developed a semantics of conditionals. In Ch. 2, we define and apply our new method of learning conditional information. It says, roughly, that you learn conditional information by updating on the corresponding Stalnaker conditional. By generalising Lewis's (1976) updating rule to *Jeffrey imaging*, our learning method becomes applicable to both certain and uncertain conditional information. The method generates the correct predictions for all of Douven's (2012) benchmark examples and Van Fraassen's (1981) Judy Benjamin Problem. In Ch. 3, we prefix Ramsey's test by suspending judgment on antecedent and consequent. Unlike the Ramsey Test semantics by Stalnaker (1968) and Gärdenfors (1978), our strengthened semantics requires the antecedent to be inferentially relevant for the consequent. We exploit this asymmetric relation of relevance in a semantic analysis of the natural language conjunction 'because'. In Ch. 4, we devise an analysis of actual causation in terms of production, where production is understood along the lines of our strengthened Ramsey Test. Our analysis solves the problems of overdetermination, conjunctive scenarios, early and late preemption, switches, double prevention, and spurious causation – a set of problems that still challenges counterfactual accounts of actual causation in the tradition of Lewis (1973c). In Ch. 5, we translate our analysis of actual causation into Halpern and Pearl's (2005) framework of causal models. As a result, our analysis is considerably simplified on the cost of losing its reductiveness. The upshot is twofold: (i) Jeffrey imaging on Stalnaker conditionals emerges as an alternative to Bayesian accounts of learning conditional information; (ii) the analyses of causation in terms of our strengthened Ramsey Test conditional prove to be worthy rivals to contemporary counterfactual accounts of causation.

Précis

Philosophers have devised various theories of the relation between cause and effect. Few of the theories have occasionally been popular. Today, none is widely agreed upon. So far, any account of causation is plagued by counterexamples. In addition, it is hard to tell whether some account tallies best with our common sense of what causes what. Hence, it is safe to say that no philosophical account of causation has yet succeeded. This is astonishing – given how pervasive and familiar causation is. Causes help us to understand and explain what is going on around us. Causes help us to intervene in the course of events to bring about certain effects, or prevent others from occurring. As thinkers and agents, we are – quite naturally – interested in causal relations. We wonder why the coffee machine is not working, why our colleague seemed so gloomy today, whether a certain diet would support our health. Everyone values knowledge of what causes what. The more surprising that there is no unanimous theory of causation at our disposal.

The importance of causation is not restricted to anyone's everyday life, but extends to the special sciences. The goal of biology, neuroscience, medicine, economics, and history – to name just a few – is to discover the causes of their respective target phenomena. Within philosophy alone, the research on causation in metaphysics and epistemology impacts contemporary debates on mental causation, action theory, decision theory, learning theory, semantics, scientific explanation, and moral and legal responsibility. The ubiquity of causation awards the prospect of a successful account of causation with a broad interest.

Our ordinary concept of causation exhibits two truisms. First, causes produce certain effects. Second, causes make a difference. That is, things would be different if the causes of some effects were absent. This difference-making idea underlies Lewis's (1973c) analysis of causation in terms of counterfactual conditionals. These conditionals are of the form 'if the cause had not been, the effect never had existed'. Lewis says that an event causes another if both events occur, and if the former had not occurred, so would not have the latter. Since Lewis's proposal counterfactual accounts have risen to some prominence in the contemporary debate on causation. Most of these accounts are meant to provide a theory of actual, singular, or token causation. The goal of such a theory is to figure out whether this particular event caused that particular event. In this dissertation, we put forth a novel analysis of actual causation in terms of a conditional that differs in kind from counterfactuals. Thereby, we achieve an analysis of causation in terms of production rather than counterfactual dependence. After all, causes bring about certain effects, they do not only make a difference.

Analyses of causation in terms of certain conditionals reveal that causal relations have a conditional structure. This observation, unfortunately, does not help much to analyse causation. For conditionals are as controversial as causation itself. And for a good reason: conditionals seem to be intimately tied to causation, even if we do not know exactly how. One indication for this conjecture is that conditionals, like causal relations, play a central role in reasoning and learning. However, there is no consensus emerging on what conditionals, and for that matter causal relations, mean. It is not even agreed upon whether conditionals have truth conditions, and thus express propositions at all (see Edgington (1995)). This makes it easy to predict that conditionals will continue to puzzle philosophers, epistemologists, logicians, and cognitive scientists alike.

In spite of the controversies surrounding conditionals, much research has originated from the same source, viz. Ramsey's (1929/1990) test for the acceptance of conditionals. The idea is that you accept a conditional 'if A then C' when you can infer its consequent C upon supposing the antecedent A. Inspired by this test procedure, Stalnaker (1968) has developed a semantics of conditionals by replacing (hypothetical) belief states by a (set of) possible worlds. Roughly, the Stalnaker conditional $A > C$ is true just in case C is true in the possible world that satisfies A and is otherwise most similar to the actual world. In a slogan, $A > C$ is true when the most similar A-world satisfies C. Adams (1975) has developed another semantics of conditionals. He has taken seriously Ramsey's phrase that the evaluation of a conditional requires to fix your degree of belief in the consequent given the antecedent. Accordingly, you accept 'if A then C' when your "degree of belief in C given A" is high. Yet another influential Ramsey Test semantics has been developed within Gärdenfors's (1988) theory of belief revision. You accept a conditional 'if A then C' if a minimal change of your beliefs to accommodate A makes you believe C as well. However, Gärdenfors (1986) showed that this version of the Ramsey test is inconsistent with his theory of belief revision under a mild assumption of non-triviality. It was Hansson (1992) who defended the Ramsey Test against Gärdenfors's inconsistency theorem. In part, Hansson saves the Ramsey Test by an alternative representation of belief states: he has used belief bases rather than belief sets to model the dynamics of belief. In contrast to belief sets, belief bases are in general not closed under logical consequence. Hence, belief bases are a less idealised model of a belief state allowing for a distinction between explicitly held beliefs and merely derived beliefs. Up to now, the Ramsey Test inspires work on conditionals and belief changes (see, e.g., Bradley (2007), Leitgeb (2010), Rott (2011), Rott (2017)).

Following Ramsey (1929/1990), Adams (1975) represents degrees of belief by a probability distribution P. He goes on to stipulate the probability of an indicative conditional as the conditional probability of the consequent given its antecedent, i. e. $P(\text{if } A \text{ then } C) := P(C \mid A)$. Lewis (1976) has shown that this stipulation does not hold for Stalnaker's Ramsey Test conditional $>$. Apart from trivial cases, the probability of a Stalnaker conditional does not equal the corresponding conditional probability, that is $P(A > C) \neq P(C \mid A)$. Lewis's result may be seen as a special case of Gärdenfors's inconsistency theorem, when belief states are modelled by probability functions and belief states change according to the rule $P'(C) = P(C \mid A)$. The setting of this special case describes the two core tenets of orthodox Bayesianism, where P is called the initial or prior degrees of belief and P' the final or posterior degrees of belief, and the rule changing the degrees of belief goes by the name of conditionalization. By conditionalization on A, a Bayesian agent learns a piece of evidence A with certainty. Jeffrey's (1965) generalisation of conditionalization allows a Bayesian agent to update her degrees of belief when the information A is learned with uncertainty. So far so good.

As compared to the learning of factual information, it is less clear how the norms of Bayesian epistemology apply to the learning of indicative conditionals (see Douven (2015)). How should you change your beliefs when you learn 'if A then C'? Virtually all Bayesian accounts agree that learning a conditional imposes some constraint on the posterior degrees of belief for the consequent given the antecedent. Reminiscent of Adams's stipulation, to learn 'if A then C' is often assumed to imply that $P'(C \mid A)$ equals approximately 1 (see, e.g., Evans and Over (2004, Ch. 8) and Oaksford and Chater (2007, p. 118)). In case a Bayesian agent learns uncertain conditional information, the constraint on the posterior takes the form $P'(C \mid A) = a \leq 1$. Then Bayesians usually apply Jeffrey conditionalization. A more sophisticated alternative has been proposed by Van Fraassen (1980a) and Williams (1980). The proposal is to model the learning of conditional information by minimising the Kullback-Leibler (KL) divergence between prior and posterior degrees of belief. When you learn uncertain conditional information, and so the constraint takes the form $P'(C \mid A) = a$ for $0 < a < 1$, minimising the KL divergence may yield different results from Jeffrey conditionalization. Van Fraassen's approach leads thus beyond the confines of orthodox Bayesianism.

Shortly after, Van Fraassen (1981) and Van Fraassen et al. (1986) have challenged the KL divergence minimizer and the orthodox Bayesian alike. They do so by putting forth a scenario of learning uncertain conditional information, the Judy Benjamin Problem. To say the least, this learning scenario has since proven to be a severe challenge for Bayesians of all varieties. And not only Bayesians are troubled. Douven and Romeijn (2011) and Douven (2012) survey the extant accounts of learning conditional information and observe that there are no satisfactory accounts of learning conditional information. All accounts – ranging from simple conditionalization on the material implication, over possible worlds accounts, to advanced Bayesian perspectives – fail to provide the correct results for Douven's benchmark examples. He concludes that a general account of learning conditional information is yet to be formulated. We aim to remedy this situation.

In light of Lewis's (1976) triviality result, it is dubious whether conditional probabilities, or at least constraints on conditional posterior degrees of belief, are the right tool to model the learning of conditional information. Luckily, an alternative to Bayesian conditionalization is not far to seek. In the same paper, Lewis has found a probabilistic updating rule, which he named imaging. Imaging on A transfers the probability shares associated to $\neg A$-worlds to the respective most similar A-worlds. We may interpret imaging on A as another way to learn A with certainty. Notably, the probability of a Stalnaker conditional equals the probability of its consequent after imaging on the antecedent, that is $P(A > C) = P^A(C)$. By replacing the updating rule of conditionalization with imaging, Adams's stipulation becomes a theorem for Stalnaker's conditional. Unlike standard conditionalization, imaging can be applied when the antecedent has probability 0 and it can be applied to nested conditionals. For instance, the image on the Stalnaker conditional $A > C$, $P^{A>C}(E)$, is well-defined and equals the probability of the nested conditional $(A > C) > E$. Hence, imaging seems to be a promising candidate to provide a rather general account of learning conditional information.

In Chapter 2, we put forth a method of learning conditional information. Roughly, an agent learns 'if A then C' by (Jeffrey) imaging on the Stalnaker conditional $A > C$. Imaging on $A > C$ amounts to (i) learning that the most similar A-world satisfies C; and that (ii) the probability share of each $\neg(A > C)$-world is transferred to its most similar $(A > C)$-world. Jeffrey imaging is our generalisation of Lewis's imaging that mirrors Jeffrey's generalisation of Bayesian conditionalization. In contrast to Lewis's imaging, our generalisation does not transfer the whole probabilistic mass – but only a part thereof – to the respective most similar worlds. Thereby, Jeffrey imaging opens the door to apply our learning method to uncertain information. Unlike extant Bayesian accounts, our method generates the correct predictions for Douven's (2012) benchmark examples and Van Fraassen's (1981) Judy Benjamin Problem. Finally, we adapt our method of learning conditional information to a method of learning causal information. The combination of the two methods provides a unified framework which allows us to clearly distinguish between conditional and causal information.

In Chapter 3, we strengthen Ramsey's test. The idea is to prefix Ramsey's test by a suspension of judgment:

> First, *suspend judgment on the antecedent A and the consequent C*. Second, add A hypothetically to your stock of beliefs. Finally, test whether you can infer C.

The resulting strengthened Ramsey Test semantics of conditionals requires that the antecedent is inferentially relevant for its consequent, unlike the semantics of conditionals due to Stalnaker (1968) and Gärdenfors (1978). Using Hansson's (1999) framework of belief bases, the relevance exhibited by our strengthened semantics is often asymmetric. No wonder, then, that we can analyse the conjunction 'because' of natural language by our strengthened Ramsey Test semantics.

In Chapter 4, we analyse actual causation in terms of our newly developed strengthened Ramsey Test conditional. The strengthened conditional is meant to express a relation of production. C produces E just in case being agnostic on C and E, you can infer C by supposing E. Roughly, we propose that C is a cause of E if C produces E and $\neg C$ does not also produce E. Hence, we reduce causation to (beliefs about) events (or facts) and generalisations. Our analysis solves the problems of overdetermination, conjunctive scenarios, early and late preemption, switches, double prevention, and spurious causation – a set of problems that still challenges counterfactual accounts of actual causation in the tradition of Lewis (1973c).

In Chapter 5, we carry over our analysis of actual causation to Halpern and Pearl's (2005) framework of causal models. Thereby, our analysis simplifies considerably on the cost of losing its reductiveness. We compare the resulting analysis to the account of Halpern and Pearl (2005) and its modification due to Halpern (2015). Both accounts define actual causation in terms of contingent counterfactual dependence. Roughly, contingent counterfactual dependence says that even if E does not counterfactually depend on C in the actual situation, E counterfactually depends on C under certain contingencies. Their definitions of actual cause still struggle with any set of problems including both overdetermination and conjunctive scenarios, unlike our analysis in the framework of causal models.

In Chapter 6, we refine Halpern and Pearl's (2005) definition of actual causation to allow for disjunctive causes as discovered by Sartorio (2006). She argues, against the verdicts of Lewis (1986b) and Halpern and Pearl (2005), for the existence of disjunctive causes. The switching scenario she considers suggests that a disjunctive fact or event can be a cause without one of its disjuncts being a cause. We show that our refinement of Halpern and Pearl's (2005) definition can take such disjunctive causes into account.

In Chapter 7, we impose properties of causation, as assumed in cognitive neuroscience, upon Woodward's (2005) interventionist account of causation. Within the resulting framework, we investigate to what extent we are justified to derive causal relations between mental properties and properties of the brain, if certain methods are used in the neuroscientific studies. The upshot is that, for methods as diverse as Functional Magnetic Resonance Imaging and Transcranial Magnetic Stimulation, cognitive neuroscientists should dare to interpret their findings as establishing genuine causal relations.

In Appendix A, we apply our method of learning conditional information to Douven and Romeijn's (2011) Jeweller Example. Appendix B contains the proofs of Chapter 3. Appendix C provides the precise definitions for the belief changes underlying the analysis of causation in Chapter 4. Appendix D contains a proof showing that, under certain assumptions, a notion of counterfactual dependence is sufficient for causation according to our definition of Chapter 4. A more detailed summary of each chapter may be found in Section 1.5 of the Introduction.

In sum, we have moved from the Ramsey Test in two directions. First, we have proposed a method of learning conditional information. This method is based on Stalnaker's semantics of conditionals and Jeffrey imaging. As compared to extant Bayesian accounts, our framework does justice to the many facets of learning conditional information. We have gone in a second direction by strengthening the Ramsey Test. This has led us to a new conditional semantics amenable to an analysis of actual causation. In fact, we have analysed actual causation in terms of our strengthened Ramsey Test conditional twice over. The first analysis uses belief bases, while the second is embedded in a framework of causal models. In these guises, our strengthened Ramsey Test conditional gives rise to two analyses of actual causation that do not fall short of contemporary counterfactual accounts. The paths we have gone show the on-going fruitfulness of Ramsey's ideas.

Sources. Chapter 2 of this dissertation builds on the publications Günther (2018) and Günther (2017a). Chapters 3, 4, and 5 build on the publications Andreas and Günther (2018), Andreas and Günther (2019), and Andreas and Günther (2018), respectively. Chapter 6 builds on Günther (2017b).

Acknowledgements

First and foremost, I would like to thank Hannes Leitgeb. He gave me all the freedom I could wish for to pursue my research interests. At the same time, he provided excellent guidance whenever I faced a difficulty. His advice has been invaluable. He knows, in particular, what is yet to be improved upon. His proposals for clarification and relevant literature have always been dead-on. Besides the persistent support he gave me, his profound knowledge makes him an admirable supervisor. For the future, I hope he continues to be there when I need advice.

I would like to express my gratitude to Stephan Hartmann. It was his class back in the Masters program of the Munich Center for Mathematical Philosophy (MCMP) that made me familiar with the problem of learning conditional information from a Bayesian point of view. Many discussions later, I finally outlined my own non-Bayesian method of learning conditional information. This has led directly to my first and second publication. Moreover, he introduced me to a great network of academics and provided many platforms to give talks. He gave me, for instance, the responsibility to organise the ECAP9 Symposium on “Current Trends in Neurophilosophy”, and even to hold his lecture on Bayesian networks for the MCMP Master students. It was more than a pleasure to have him as a supervisor.

I am grateful to Stephan Sellmaier. He was always ready to provide advice concerning both, academia and life. Throughout my PhD program, I felt strong support from his side and the environment he provided. In particular at the beginning, he accelerated my personal and academic development considerably. He always had an ear for all the PhD students in his Research Center for Neurophilosophy and Ethics of Neurosciences, which is part of the Neurophilosophy division of the Graduate School of Systemic Neurosciences (GSN). There, he created a space where we could try out new ideas in front of critical but friendly colleagues. I profited from this environment a lot.

My appreciation is due to Holger Andreas. I learned a lot from the frequent and ample discussions we had. We are in the middle of developing a research project on the Ramsey Test and causation. The first fruits of our collaboration can be found in chapters 3, 4, and 5. Over the years, we exchanged numerous drafts commenting on each other’s contributions. I am grateful for his thoroughness and patience with me. It is no overstatement that the dissertation in this form would not have been possible without him.

I am indebted to Igor Douven. He proposed a – broadly speaking – Bayesian answer to the question what we learn when we receive conditional information. Without any formal commitment, he commented extensively on the earliest draft I produced on said question. Despite my non-Bayesian tendencies in approaching the question, he and Stephan Hartmann valued my proposal and encouraged me to submit it. Please excuse me for being an enfant terrible as regards Bayesianism.

Special thanks go to Andreas Herz, director of the Bernstein Center for Computational Neuroscience (BCCN) in Munich. He, the BCCN, the GSN, and the MCMP strongly supported a conference I organised together with Holger and Kay Thurley on “Causation, Explanation, Conditionals”. Andreas was eager to establish an interdisciplinary exchange between neuroscientists and philosophers, for which I am very grateful.

I owe a great many thanks to Hans Rott. He asked crucial questions at a talk I gave on learning conditional information and took plenty of time to carefully read my drafts on the Ramsey Test. His comments pointed to gaps, followed by proposals to overcome these and to work out interesting comparisons with extant literature. As a result, I could always improve upon the respective manuscript. His comments helped me, in particular, to rebut objections a reader might have before these were actually levelled. It is noteworthy that some core ideas of this dissertation originate more or less directly from Hans's work on belief revision and the Ramsey Test. Together with Holger, I attempted to carry Hans's ideas further and push them into the arena of causation. Hans contributed significantly and in ingenious ways at each important step during my PhD studies – and all of that without any formal commitment.

I would like to explicitly thank both, the Graduate School of Systemic Neurosciences and the Munich Center for Mathematical Philosophy, in particular the institutional heads Benedikt Grothe, and Hannes and Stephan. All of them are working hard to create a small heaven for PhD students. The GSN made it possible that I could organise the above mentioned conference, for which I feel honoured and am more than grateful. As regards financial support, academic events, soft-skill courses, management, and social activities, the GSN sets a bar so high that no wishes remain unfulfilled. Combined with the likewise rich offer of the MCMP, the opportunities were so numerous that I could not even take up half of the highlights I would have had liked to. It is hard to imagine a better and more vibrant environment for personal and academic development. I am happy to be part of the connection between the GSN and the MCMP.

I would like to thank Olivier Roy, Julian Nida-Rümelin and Fiorella Battaglia for continuous help and support on various occasions.

Finally, I would like to express my gratitude to Atoosa Kasirzadeh, Cameron Beebe, and Gary Mullen for reading my all too early drafts and for an enjoyable time. The same applies to my GSN fellows, especially the Neurophilosophy group, and my former MCMP classmates. I apologise to the many people that helped me over the last couple of years, but which I forgot to mention. I dedicate this work to my parents, Maria and Konrad Günther.

Contents