



Variation of information as a measure of one-to-one causal specificity

Pierrick Bourrat^{1,2} 

Received: 22 September 2017 / Accepted: 14 September 2018 / Published online: 23 November 2018
© Springer Nature B.V. 2018

Abstract

The interventionist account provides us with several notions permitting the qualification of causal relationships. In recent years, there has been a push toward formalizing these notions using information theory. In this paper, I discuss one of them, namely causal specificity. The notion of causal specificity is ambiguous as it can refer to at least two different concepts. After having presented these, I show that current attempts to formalize causal specificity in information theoretic terms have mostly focused on one of these two concepts. I then propose and apply a new information-theoretic measure which captures the other concept.

Keywords Causation · Interventionist account · Causal specificity · Information theory

1 Introduction

Within the interventionist account, causation between two variables C and E is characterized as follows:

Minimal criterion of causation: C is a cause of E , if there is at least one intervention on C that changes the value of E in a set of background conditions Z .¹

An intervention on a variable can be regarded as a particular kind of change which, *at a given point in time*, produces nothing but a change in the value of that variable

¹For a version of this criterion see Woodward (2013, Section 4). For a much more detailed version see Woodward (2003a, p. 59).

✉ Pierrick Bourrat
p.bourrat@gmail.com

¹ Department of Philosophy, Macquarie University, North Ryde, NSW 2109, Australia

² Department of Philosophy, Charles Perkins Centre, School of History and Philosophy of Science, The University of Sydney, Camperdown, NSW, 2006, Australia

(Woodward 2003a, p. 14).² The minimal criterion of causation tells us that if, following at least one intervention on C , some change is observed on the value of E when compared to the value of E when no intervention on C is performed, then C is a cause of E in Z . In the absence of a change in E , C is not a cause of E in Z .

Although the minimal criterion of causation can help us to make a distinction between causes and non-causes, it does not permit to make distinctions among causes (Woodward 2010). Intuitively however, not all causes are equal. When providing a causal explanation, some causes might be deemed more relevant, more insensitive to changes, or as having other desirable properties. This is the problem of causal selection (for more on this problem, especially in a biological context see Franklin-Hall 2015). For instance suppose one lights a match. Most of us will have the intuition that the action of striking the match is a more important cause for the match lighting than the presence of oxygen in the surroundings of the match. Yet, intuitions alone cannot serve as criteria for causal selection.

In order to move the causal-selection problem away from intuitions, a number of properties of causal relationships have been proposed to supplement the minimal criterion within the interventionist account.³ One of these properties, which will be the focus of this paper, is causal ‘specificity’ (Weber 2006, *in press*; Griffiths et al. 2015; Pocheville et al. 2017; Waters 2007). Specificity is a notion that is both pervasive and ambiguous in biological sciences (Woodward 2010, 301). One area in which the notion of causal specificity has been invoked is molecular biology. Within this context, authors have attempted to assess the role played by DNA when compared to other factors of the cell (causes) during transcription such as RNA polymerases. Transcription is the process of synthesizing an RNA molecule from a DNA sequence. Should one consider the sequence of DNA involved in the production of a messenger RNA (mRNA) as a more important cause of the sequence of the mRNA produced, than the polymerase involved in this process? One is tempted to respond in the affirmative because intuitively this sequence seems to have a higher specificity with respect to the sequence of the mRNA produced during transcription. Yet, without a precise measure of what is meant by ‘specificity’, this intuitive answer is insufficient and might be misleading. This is the context in which Griffiths et al. (2015) provided a measure of causal specificity. In this paper, building on previous philosophical work on this topic (see Weber 2006, *in press*, Woodward 2010, Griffiths et al. 2015, Pocheville et al. 2017), I extend this work and provide complementary measures of causal specificity. Following previous work, my analysis will be limited to nominal variables.

In Section 2, starting from the distinction made in Woodward (2010), I present in more depth what one can mean by ‘causal specificity’ with the help of causal diagrams to which I will refer in later sections. In Section 3, I show that the measure based on mutual information proposed by Korb et al. (2011), Griffiths et al. (2015), and Pocheville et al. (2017) of causal specificity corresponds to the range of causal

²For a more precise characterization of the notion of intervention see Woodward (2003a, pp. 98–99).

³For a number of these see Woodward (2003a, 2010, Griffiths et al. (2015), Pocheville et al. (2017).

influence notion of causal specificity,⁴ but not to one-to-one causal specificity. In Section 4 I propose a principled method for choosing the grain of description of causal relata. To measure one-to-one causal specificity, in Section 5, building on the work of Griffiths et al. (2015), I propose a different information-theoretic measure based on the notion of variation of information, which is a measure related to mutual information. I then show that this second measure, in and of itself, does not permit capturing the range of causal influence notion of causal specificity. Finally, in Section 6, I revisit a version of the example proposed by Griffiths et al. (2015) on transcription in light of my measure of causal specificity. I conclude by claiming that a more complete formal characterization of causal specificity in the philosophy of causation literature ought to take into account both the range of causal influence and the one-to-one causal specificity of a relation and thus must use both mutual information and variation of information to do so.

2 Dimensions of causal specificity

Woodward (2010) distinguishes two kinds of causal specificity, namely what I call the ‘range of causal influence’⁵ and what he calls the ‘one cause-one effect’ (p. 310) notion of specificity (henceforth ‘one-to-one specificity’). The range of causal influence of a variable C over another variable E concerns the number of interventions on C that produce changes in the value of E and the extent to which these changes are different from each other. The higher the number of interventions that produce changes in the value of E that are different from other changes produced by different interventions, the higher the range of causal influence. Woodward (2010) invokes this notion in the context of molecular biology, following claims made by, among others, Davidson (2001, pp. 1–2) and Waters (2007), that DNA is a more specific cause for the production of mRNA than other causes because the DNA sequences control with some precision the production (hence the specificity) of mRNA sequences, whereas the variation in many other cellular resources would typically lead either to the production or absence of production (hence the lack of specificity) of the same mRNA sequences.

One-to-one specificity captures the extent to which one cause is connected to a single effect, and the extent to which one effect is connected to a single cause. The focus of Woodward’s analysis with respect to one-to-one specificity is mostly the *variable* level. By this, I mean that the one-to-one correspondence is between the variables C and E . Yet, another possible focus of analysis, which will concern me here, is the *value of a variable* level. By this, I mean that the one-to-one correspondence occurs, not between the variables, but between *values* of the variable. Although these two views on one-to-one specificity seem *prima facie* conceptually distinct, they are connected. This is because two or more variables can be aggregated into a single one. For instance, a number n of variables X_1, X_2, \dots, X_n , assuming they are nominal and independent, can be re-described as a single ‘aggregate’ variable W . Suppose that

⁴Causal specificity qua range of influence is called ‘causal power’ by Korb et al. (2011).

⁵This corresponds to Woodward’s ‘INF’ notion of specificity Woodward (2010, p. 305).

each variable has v values with some given probability for each value. The number of values of W will simply be n^v .⁶ Although Woodward presents one-to-one specificity from the variable perspective, he discusses in passing the possibility of “lumping” or “splitting” a variable (Woodward 2010, p. 312), and applies the notion of one-to-one specificity between states or values of variables rather than the variables themselves (Woodward 2010, p. 313). In the remainder of the article, I will frame my analysis with respect to one-to-one specificity from the value perspective.

As shown by Woodward (2010, pp. 308–310), one-to-one specificity has been invoked in many different scientific contexts, including epidemiology (Weiss 2002; Hill 1965; Höfler 2005; Rothman and Greenland 2005), immunology (Langman 2000), the study of enzymes (Suzuki 2015), or behavioral genetics (Kendler 2005). Other contexts, not mentioned by Woodward, include the evolution of mutualism (e.g., fig/wasp coevolution, see Machado et al. 2005) and host/prey specificity for biological control (Brodeur 2012). We will see in Section 6 how it is also relevant in molecular biology.

Inspired by Woodward’s analysis, I propose that when one asks whether a relationship between two variables C and E which satisfies the minimal criterion of causation, is causally specific, the question can take at least three meanings which correspond to two concepts.

1. One-to-one Specificity

- (a) **Specificity of the cause for the effect** (“specificity of the cause”, for short)
To what extent does a value of C cause values of E which are different from values of E caused by other values of C ? Or in other words, for each value of E to what extent is this value determined by a single value of C ?
- (b) **Specificity of the effect for the cause** (“specificity of the effect”, for short)
To what extent is a value of E caused by values of C which are different from the values of C which cause other values of E . Or in other words, for each value of C to what extent does this value determine a single value of E ?

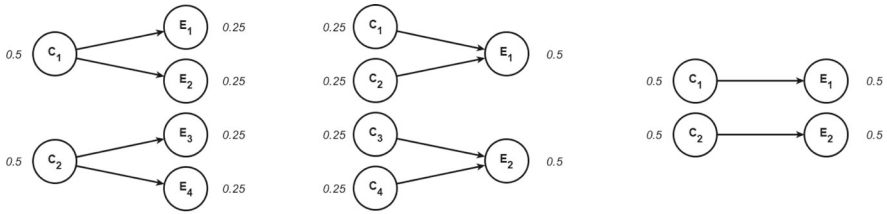
- 2. **Range of causal influence** What is the number of values of C that lead to different values of E , assuming that for each value of C there corresponds one and exactly one value of E ?

I regroup the ideas of specificity of the cause and of specificity of the effect under the general notion ‘one-to-one causal specificity’ identified by Woodward (2010), because they are the two complementary aspects of a bijective causal mapping.⁷

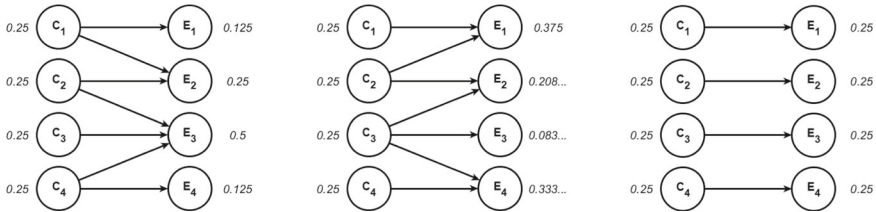
If the answer to question (1a) is that every value of E is caused by one and only one value of C , as is illustrated in the diagrams (a), (c), and (f) of Fig. 1, then we have a case of maximal specificity of the cause (see the caption of Fig. 1 to understand

⁶In cases where the assumption of independence is violated, the process of aggregation will involve a more complex procedure.

⁷The adjective ‘bijective’ is another way to specify the mapping between two variables as being one-to-one.



(a) Specific cause for the effect and nonspecific effect for the cause. (b) Nonspecific cause for the effect and specific effect for the cause. (c) Specific cause for the effect and specific effect for the cause (bijective mapping over two values).



(d) Nonspecific cause for the effect and nonspecific effect for the cause. Range of causal influence equal to that of (c) and (e) (e) Nonspecific cause for the effect value and nonspecific effect for the cause. Range of causal influence equal to that of (c) and (d) (f) Specific cause for the effect and specific effect for the cause (bijective mapping over four values).

Fig. 1 Causal diagrams representing different causal relationships between C and E . In all the diagrams presented in this figure equiprobability for the values of C is supposed. Each arrow pointing from one value of C represents the probability of this value causing the value of E it points to. If only one arrow leaves a given value of C , the probability is 1 as in (c) and (f). If more than one arrow leaves a value of C , as in (a), (d) and (e), then we suppose that each arrow has the same probability conditioning on C . E , in such a case, is indeterministic because one value of C does not always cause one value of E

the diagrams). If the answer to this question is now that some or all values of E are caused by more than one value of C , as is illustrated in (b), (d) and (e), then we have a case in which the cause is to some extent non-specific to the effect. In other words, some values of E are multiply realized. For instance, let us start from the example of Griffiths et al. (2015, p. 539 ff.) and imagine that a sequence of DNA and an RNA polymerase, symbolized by the variables DNA and POL respectively, produce an mRNA symbolized by the variable RNA during the process of transcription. Suppose a case in which four different sequences of DNA produce four different sequences of mRNA, assuming the same polymerase is transcribing the DNA sequence. There is in this case a high causal specificity of the cause (DNA) for the effect (RNA). Suppose now that polymerases with different structures are nevertheless all able to

produce the same mRNA,⁸ assuming the same DNA sequence is used. In this case we have a low one-to-one specificity of the cause (*POL*) for the effect (*RNA*), or at least lower than that of *DNA* sequences for *RNA*.

Moving to question (1b) and applying the same reasoning, if the answer is that every value of *C* causes one and only one value of *E*, as is illustrated in the diagrams (b), (c), and (f) of Fig. 1, then we have a case of maximal specificity of the effect. If now the answer to question (2) is that some or all values of *C* cause more than one value of *E*, as is illustrated in the diagrams (a), (d) and (e), then we have a case in which the effect is to some extent non-specific to the cause or indeterministic.⁹ Take again the example of the DNA sequences and the polymerase producing an mRNA, but suppose now that alternative splicing occurs as proposed by Griffiths et al. (2015, p. 543 ff.), so that a given sequence of DNA does not lead to the production of a single mRNA sequence, but to a number of them. Alternative splicing, more particularly *cis*-splicing, is the process by which portions of a gene called exons, are stuck together with different combinations to produce different mRNA sequences, which will later be translated into different amino acid sequences (Griffiths and Stotz 2013, pp. 54–56). Splicing decreases the causal specificity of the effect (*RNA*) for the cause (*DNA*).

Moving to question (2), if we take once again the two diagrams (c) and (f) of Fig. 1, which both have the same one-to-one specificity (bijection in each case),¹⁰ we can nevertheless see that the range of values of *C* over which the mapping between *C* and *E* is bijective is two in diagram (c) and four in diagram (f). A measure of causal specificity understood as range of control or influence ought to account for this difference and give a higher range of causal influence for the relationship presented in (f) than in (c). This is the notion of specificity which has been the target of Weber (2006, *in press*), Waters (2007), Griffiths et al. (2015) within the context of comparing the causal influence of DNA sequences with that of other cellular factors for the production of mRNA sequences. To illustrate the range of influence notion of causal specificity, take again the case of the DNA sequences and the polymerase producing an mRNA. Suppose again that there is some alternative splicing so that a single type of DNA sequence can lead to the production of more than one type of mRNA sequence. Suppose that there are only two possible types of DNA sequence, each of which can lead to four mRNA sequences with equal probability via the effect of some splicing mechanisms represented by the variable *S*. Suppose that there are four possible distinct splicing mechanisms. In this case, the range of influence of *DNA* on *RNA* would be lower than that of *S*, since by intervening on *DNA* one could control

⁸This is a realistic scenario as RNA polymerases with different structures are found in different stocks of a single strain of *Escherichia coli* (Jishage and Ishihama 1997).

⁹Whether the indeterminism is ontological or epistemic will depend on the nature of the case, such as whether the relation between *C* depends on some quantum events or whether there is some unknown variation in some of the background variables not cited in the diagram.

¹⁰I use here two bijective diagrams because it is the easiest way to illustrate the notion of range of causal influence. That said, I could have used any other type of relationship such that when making the comparison any amount of one-to-one unspecificity is discounted for assessing the range of causal influence.

only two sets of values of *RNA* (like a switch), while intervening on *S* would give control of four sets of values of *RNA*.¹¹

In the next section, I show that the measure of mutual causal information proposed, among others, by Griffiths et al. (2015) permits accounting for the range of causal influence notion of causal specificity but does not capture one-to-one causal specificity. Griffiths et al. note the limits of mutual information considered as the sole measure of specificity (Griffiths et al. 2015, pp. 537–538). Nonetheless, their exclusive focus on this latter notion detracts from the completeness of their account. In Section 5, I develop some of Griffiths et al.'s points by proposing a new information-theoretic measure that provides a precise explication of the notion of one-to-one specificity.

3 The limits of mutual causal information

The diagrams presented in Fig. 1 provide an intuition about the levels of one-to-one causal specificity and the range of causal influence a given causal relationship between two variables has. That said, they do not, on their own, provide a precise measurement of these two quantities. For instance, looking solely at Fig. 1, it would be hard to determine whether diagram (a) shows a different range of causal influence and/or one-to-one specificity between *C* and *E* than diagrams (d) and (e).

Korb et al. (2011) and Griffiths et al. (2015) came independently to the same measure of causal specificity between two variables, defined as the amount of mutual information transferred from the causal variable to the effect variable. In this section I show that the measure of causal specificity proposed by these authors corresponds solely to the notion of range of causal influence and must be complemented by an additional measure to account for one-to-one specificity.

I follow the treatment of Griffiths et al. (2015, p. 538) who define causal specificity in information-theoretic terms as:

SPEC: the specificity of a causal variable is obtained by measuring how much mutual information interventions on that causal variable carry about the effect variable.

The mutual information between two variables corresponds to the the amount of uncertainty lost on a variable given knowledge of the other, or again the amount of redundant information between two variables. Mutual information is closely related to the notion of Shannon entropy used in information theory (Shannon 1948). The entropy of a variable corresponds to the amount of uncertainty about that variable. Uncertainty is often measured in *bits*, but other units can be used. When measured in *bits*, the entropy of a variable represents the expected minimal number of questions with a yes/no answer one has to ask to know the value of the variable with certainty. For instance, if the variable is ‘outcome of tossing a fair coin’, its entropy is 1 *bit*.

¹¹The sets of values of *RNA* controlled by *DNA* have a cardinality of 4 while those of *RNA* have a cardinality of 2.

To know for sure that the toss effectively resulted in ‘tails’ or ‘heads’, one just needs to ask in expectation at least one question (such as ‘is it tails?’ or ‘is it heads?’). If the variable is now the ‘outcome of tossing two fair coins successively’, the entropy is two *bits* because the expected minimal number of yes/no questions that must be asked is two, one for the first coin and another for the second coin (I assume here that the tosses are independent).

Formally, the mutual information between the variable E and the variable C can be calculated as

$$I(E; C) = H(E) - H(E|C) \quad (1)$$

where $I(E; C)$ is the mutual information between E and C , $H(E)$ is the entropy of E , and $H(E|C)$ is the entropy of E conditional on C , that is once we know the value of C .¹² Following Griffiths et al. (2015), by using the $\text{do}(\cdot)$ operator (symbolized with ‘ $\hat{\cdot}$ ’ in equations) proposed by Pearl (2009), which represents intervening on a variable, one can provide a causal version of mutual information from \hat{C} to E ($I(E; \hat{C})$), as follows:

$$I(E; \hat{C}) = H(E) - H(E|\hat{C}) \quad (2)$$

Here, \hat{C} represents the variable C to which the value has been set by an intervention. I will refer to this measure as ‘mutual causal information.’

The motivation underlying the use of the $\text{do}(\cdot)$ operator is that it permits to separate cases in which any mutual information between E and C is the result of a third (confounding) variable from cases in which the mutual information between the two variables does not depend on the value of a third variable. That is, if the mutual information between E and C is calculated when the values of C occur naturally, a parent variable of both C and E might be the reason (at least partly) why the mutual information between E and C is positive. By setting the values of C by interventions with the $\text{do}(\cdot)$ operator, any putative causal arrow heading to C is broken, so that the mutual information calculated is causal as opposed to statistical. Griffiths et al. (2015, Appendix B) and Pearl et al. (2016, pp. 53–55) provide more detailed explanations with examples for the motivation underlying the use of the $\text{do}(\cdot)$ operator.

The causal version of mutual information matches the range of causal influence notion of causal specificity. This can easily be seen if one notices that the mutual causal information from \hat{C} to E in diagram (c) of Fig. 1 is 1 *bit* against 2 *bits* in diagram (f) of the same figure. This result is expected since the range of values in the latter diagram, assuming equiprobability for the values of C , is twice that of the former, and there is maximal one-to-one specificity (bijective mapping) in both diagrams. Equiprobability for the values of C corresponds to a case of maximum uncertainty about C , that is of maximum entropy. Interpreted causally, this implies that no value of C is privileged to determine the *possible* range of influence of C on E . As argued by Griffiths et al. (2015), this matches with Woodward’s analysis of specificity qua range of influence. That said, they recognize that depending on the context it might be more appropriate to use other distributions for C . In particular, they discuss the propositions made in different biological contexts by Waters (2007),

¹²For a short formal introduction to the notion of entropy, conditional entropy and mutual information and how to calculate them see Griffiths et al. (2015, Appendix). For a longer introduction see Stone (2015).

who argues that the actual distribution for C is the relevant one in the context of classical genetics, that of Weber (in press), who argues that one should use the “biologically normal” distribution, and that of Griffiths and Stotz (2013, 198-199) who argue that in some contexts, such as behavioral developmental or systems biology, researchers are interested in *potential* causes, as opposed to actual ones, so that the uniform distribution (equiprobable values) is the most relevant one. For my purpose, I will restrict my analysis to the case of maximum entropy (equiprobable values) since it is the one that matches the best with Woodward’s initial treatment of causal specificity.

When the causal relationship is non-bijective, mutual causal information can be seen as the range of values over which the minimal criterion of causation would have been satisfied, had the mapping between \widehat{C} and E with this mutual causal information been bijective (Pocheville et al. 2017). This means that, considering a constant number of possible values for C and E , interventions on C that do not make a difference to E (multiple realization of E) and interventions of the same class (that is intervention changing the value of C from one value to the other) that lead to different possible values of E (indeterminism of E), will not contribute to reducing the mutual causal information from \widehat{C} to E when the same number of interventions are possible with a bijective mapping. That is why if we now calculate the mutual causal information from \widehat{C} to E in the non-bijective diagrams (a), (b), (d) and (e) of Fig. 1, in spite of the fact that some of the diagrams have a different number of values for \widehat{C} and/or E , we nonetheless find 1 *bit* in each case.

Thus, in spite of differences in causal structure and number of values for both \widehat{C} and E , a measure of mutual causal information, if interpreted as causal specificity without clearly distinguishing range of influence from one-to-one specificity, tells us that there is no difference in causal specificity between the bijective case of diagram (c) and any of the other four non-bijective diagrams. Yet, considering the distinction made in Section 2 between range of causal influence and one-to-one specificity, differences in one-to-one specificity between some of the diagrams in Fig. 1 with a mutual causal information from \widehat{C} to E of 1 *bit* are not captured by this measure.

These differences in one-to-one specificity are far from trivial. To see why, let us take a slightly more concrete example. Suppose that we study two strains of bacteria and that in one strain (S_1), we identify two possible alleles (mutations) (m_1 and m_2) on a gene that lead bacteria S_1 to resist a given antibiotic (a) while two other alleles on the same gene (m_3 and m_4) lead this bacteria not to resist this antibiotic ($\neg a$). In the second strain (S_2), we suppose that the same gene (or more accurately, the homologous gene) can only exhibit two alleles, namely m_1^* which provides a resistance to the antibiotic (a), and m_2^* which provides no resistance to the antibiotic ($\neg a$). These two cases have effectively and respectively the same structure as the diagrams presented in (b) and (c), replacing C by M and M^* (for ‘mutation’), and E by A (for ‘antibiotic resistance’). Hence the mutual causal information in both cases, from \widehat{M} to A for S_1 and \widehat{M}^* to A for S_2 , assuming equiprobability of the causal variables, is 1 *bit*.

Yet, at least in some biological contexts, one could not treat the relationship $M \rightarrow A$ for S_1 as having the same specificity as the relationship $M^* \rightarrow A$ for S_2 . A simple measure of the ratio of testing interventions on M , that is interventions that

make a difference on the effect variable (using Woodward and Hitchcock's 2003b terminology), namely A in our example, shows this. This measure is $\frac{2}{3}$ (two interventions out of three for each value of C) in the case of S_1 , against 1 (one intervention out of one for each value of M^*) in the case of S_2 . If one wanted to physically intervene in the same way in the two systems in order to change the resistance of bacteria, one would not have the same success rate in the two systems. The success rate would be 66.67 % with S_1 while it would be 100 % with S_2 .¹³ A measure of mutual causal information between the causal relata does not permit capturing this difference in one-to-one specificity. In Section 5, I provide an information-theoretic measure that captures this difference. But before doing so I need to provide a principled reasoning for deciding what the appropriate grain of description for the causal relata should be.

So far, I assumed that the way one describes the variables C and E is unproblematic. Yet, without any principled reasoning, there exists an infinite number of ways one can potentially describe them, depending on the fineness of the description. Furthermore, in some cases, several values of C and/or E can be aggregated into one single value without changing the range of influence of the causal relationship. The case of the two strains of bacteria presented above is one of them. This fact seems to imply that one-to-one specificity is, at least to some extent, an arbitrary distinction. In the next section I show that there is a non-arbitrary way to ground the choice for a grain of description.

4 Range of influence and grains of description

Pocheville et al. (2017) propose that the appropriate method to choose the grain of description for a causal explanation is to aggregate the values of the causal relata in a way that minimizes the entropy of C and E but maximizes the mutual causal information from \hat{C} to E .¹⁴ Doing so, they claim, fixes objectively a 'grain' at which to describe the causal relationship. This is equivalent to considering that the relevant grain to describe a causal relationship is one that, for a given case, maximizes the ratio of possible interventions that make a difference on the effect variable (i.e. maximizes the ratio of testing interventions) on the total number of possible interventions.

Although following strictly this approach is on the right track, it must be developed to take one crucial aspect of the scientific practice into account, namely that interventions on variables (by means of experimentations, for instance) are never made in isolation from each other. Rather they are made within a paradigm that fixes the type of intervention performed on biological or other systems and, at least in some cases, with the aim to compare the effects of the same class of interventions being performed on different but homologous systems. And indeed, Pocheville et al. (2017)

¹³The example of the two strains of bacteria only involves a difference in specificity of the cause. Other examples involving differences in specificity of the effect or both differences in specificity of the cause and of the effect could be constructed (see Section 6).

¹⁴Strictly speaking, Pocheville et al. only consider a case in which the entropy of \hat{C} can be reduced but do not discuss cases in which the entropy of E can be reduced. I assume here that they would argue for the same approach in such cases.

are clearly aware of this when they write in passing, referring to a theoretical example: “[w]hilst [our method] is the optimal way to discretise the variable[s ...] for this single experiment [...], it is not optimal for a wider experimental program!”. I extend here their reasoning.

To see why the method of maximizing mutual causal information while reducing the entropy of causal relata is not optimal in a context wider than a single experiment or case, let us take again our example with the two strains of bacteria and antibiotic resistance. In this example, M in S_1 and M^* in S_2 both have the same range of influence, namely 1 *bit*. The entropy of the causal relata in S_1 can be reduced without changing the mutual causal information, by aggregating the two alleles leading to the resistance to the antibiotic (m_1 and m_2) into one single value and aggregating the two alleles leading to the non-resistance to the antibiotic (m_3 and m_4) into a second single value. In doing so, the diagram (a) for S_1 would effectively become like the diagram (c) in Fig. 1.

Yet, contextualizing this case in a larger experimental program in which multiple examples of antibiotic resistance in bacteria are considered, another conclusion about which grain of description one ought to use to effectively compare the role of mutations in bacterial resistance in different strains is reached. In fact, one challenge for the strict ‘maximal mutual causal information for a minimal entropy’ method for this case, is that an intervention in the case of S_1 and of S_2 does not represent the same physical process: intervening on M in the case of S_1 would involve changing indeterministically the DNA sequence from either of two values to either of two other different values (each with a 0.5 probability). In contrast, intervening on M^* in the case of S_2 would involve changing the DNA sequence deterministically from one value to another one.¹⁵ The problem here is that to be able to compare explanations belonging to the same class, one ought to perform the same class of interventions. If this principle is not respected, the comparison of the effect of mutation on bacterial resistance will be made between different explananda and the ‘right’ grain of description might end up being very different in very similar cases. Furthermore, when scientists physically intervene on a class of biological systems using a particular technique, they fix the class of physical interventions performed and consequently put some constraints on the type of questions, using this particular molecular technique, one can answer. A strict ‘maximal mutual causal information for a minimal entropy’ approach does not permit taking these points into consideration.

These points acknowledged, choosing the right grain of description to be able to compare explanations of the same class, corresponds to choosing the grain of interventions on C that maximizes the mutual causal information from \hat{C} and E while minimizing the entropy on both C and E , and *considering the class of phenomena studied*.¹⁶

¹⁵Recall that we assume that other physical values were impossible in the case of M^* .

¹⁶It should be noted that this approach is prey to a form of the reference class problem (see Hájek (2007) for more on this problem), namely that there will be an infinite number of classes of phenomena a particular case belongs to that might lead to contradicting views about the right grain of description to choose. I will not attempt here to provide a solution to this problem which is a general one in many contexts.

Take again our cases with the two strains of bacteria which will provide a case in point. These two situations belong to a class of phenomena in which mutations lead to change in the phenotype observed in this species. In such a class of phenomena, considering a large number of strains, the class of interventions amounting to one and only one physical mutation (for example a point mutation on a gene) carries on average more mutual information on a phenotype than the class of interventions consisting physically in two possible mutations, that is, an indeterministic intervention leading to two indeterministic values of M . This is because in many cases of this class of phenomena, indeterministic interventions consisting in two physically possible mutations will lead to indeterministic outcomes on the effect variable that would have carried more mutual causal information had the intervention been physically deterministic. This is true even though locally, that is for some particular cases such as the case with S_1 , one type of intervention consisting physically in an indeterministic intervention will make no difference to the mutual causal information between the causal relata, when compared to a deterministic intervention in a case belonging to the same class. If this reasoning is correct, the grain of description for C corresponding to one physical possible mutation (rather than two with equiprobability) will be the correct one to use with both strains and more generally for this class of phenomena. This will amount to considering four values for C with S_1 and only two with S_2 .

With a principled way to decide the level of one-to-one specificity for a given causal explanation, I propose in the next section an information-theoretic measure of this dimension of causal specificity.

5 Variation of information as a measure of one-to-one causal specificity

We saw in Section 3 that the notion of mutual causal information, if it permits accounting for the range of causal influence notion of causal specificity, does not permit distinguishing causal relationships with differences in one-to-one specificity. To measure one-to-one specificity, using information theory, I propose to use a quantity different from mutual information, called ‘variation of information.’¹⁷ Variation of information between two variables E and C ($VI(E; C)$) is defined as the sum of two conditional entropies, namely the conditional entropies of C knowing E ($H(E|C)$) and of C knowing E ($H(C|E)$), so that

$$VI(E; C) = H(E|C) + H(C|E) \quad (3)$$

Variation of information is a measure related to mutual information in the following way. We have

$$VI(E; C) = H(E, C) - I(E; C) = H(E) + H(C) - 2I(E; C) \quad (4)$$

¹⁷For more on this concept see Meilă (2003).

where $H(E, C)$ is the joint entropy of E and C .¹⁸ Using the $\text{do}(\cdot)$ operator, like in the case of mutual information, we can define a causal version of variation of information as

$$VI(E; \widehat{C}) = H(E|\widehat{C}) + H(\widehat{C}|E) \quad (5)$$

When the variation of causal information from \widehat{C} to E is zero, not only does this imply that all the entropy present in C is transferred to E , but also that it does so without any loss. In other words, it tells us that the mapping between C and E is bijective, which means that to every one value of C there corresponds exactly one value of E . When the variation of causal information is non-zero this implies either that part of the entropy of C is not transferred to E which means that some values of E are multiply realized and/or that the entropy of E is higher than the entropy of C in which case the relationship $C \rightarrow E$ is to some extent indeterministic. These two possibilities correspond well to the respective notions of the cause for the effect being to some extent unspecific when $H(\widehat{C}|E) > 0$ and of the effect for the cause being to some extent unspecific when $H(E|\widehat{C}) > 0$. We saw in Section 2 that these concepts are the two complementary ones that permit assessing the extent to which a relationship is one-to-one specific.

Thus, because of these different properties, variation of causal information is a good candidate to measure one-to-one causal specificity. More precisely, since a case of maximal one-to-one causal specificity from \widehat{C} to E corresponds to a nil variation of causal information, variation of causal information measures the one-to-one causal *unspecificity* from \widehat{C} to E . Taking our case with the two strains of bacteria presented in the previous section, we find that the variation of causal information from \widehat{M} to A (S_1) is positive (1 *bit*) while it is nil from \widehat{M}^* to A (S_2). This implies a higher (and maximal) one-to-one specificity for S_2 when compared to S_1 . Note importantly that the two measures obtained can be compared because one of them is 0 (maximally one-to-one specific). More generally, when the measures to be compared are non-nil, each variation of information measure must be normalized over the joint entropy between the cause and the effect (see Section 6 for an example). This is because variation of information is an absolute measure that does take into account the possible differences in entropy between the causal variables. Thus, everything else being equal, a causal variable with a higher entropy will have a higher variation of information with the effect variable, than a causal variable with a lower entropy. Normalizing the variation of causal information over the joint entropy permits to overcome this problem.

We now have at hand two different measures of causal specificity. One, namely mutual causal information, captures the range of causal influence dimension of causal specificity, but does not permit accounting for the one-to-one dimension of causal specificity. The other, namely variation of causal information, captures the one-to-one dimension of causal specificity. Whether variation of causal information also captures the range of causal influence dimension of causal specificity is what I tackle next.

Like with mutual causal information not capturing one-to-one specificity, it is quite straightforward to show that variation of causal information does not capture

¹⁸For more on the notion of joint entropy see Cover and Thomas (2006, Chap. 2).

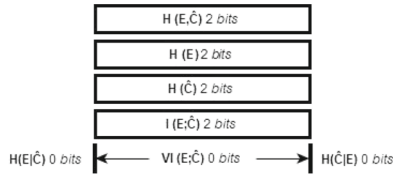
the range of causal influence of a causal relationship. To see this, suffice to take the diagrams (c) and (f) of Fig. 1 and calculate the variation of information in each case. Since the range of causal influence in diagram (f) is higher than that in diagram (c), if the variation of causal information is different in each case, this will indicate that this measure does not solely capture one-to-one specificity, but also potentially the range of causal influence. Yet, the variation of causal information in both cases is nil, which interpreted in one-to-one causal specificity terms implies that the relation is bijective in both cases.

Range of causal influence and one-to-one specificity are nevertheless related in the following way. Assuming a causal relationship with a constant joint entropy, which could involve an equal number of equiprobable values for both \widehat{C} and E (constant $H(\widehat{C})$ and $H(E)$), but not necessarily, the further away the causal mapping between C and E is from a bijective mapping with equiprobable values for C , both the lower the mutual causal information and the higher the variation of causal information from \widehat{C} to E will be. This can easily be verified with Eq. 4 and is illustrated with the graphical representations in Fig. 2. These are representations of the relationships between the different causal-information-theoretic measures discussed so far. We can see in Fig. 2, that the relationships (b) and (d) have the same joint entropy (4 bits). In (b) the mutual causal information is higher than in (d) (2 bits against 1 bit), which also implies a lower variation of causal information (2 bits against 4 bits). That said, knowing nothing about the mapping between \widehat{C} and E and their respective entropies, a given mutual causal information measure will tell us nothing about whether the relationship is bijective, and a given variation of causal information measure can tell us nothing about the range of causal influence of this relationship. Here again this falls out of Eq. 4 and this can be illustrated with Fig. 2 in which we can see that in spite of an equal mutual causal information (2 bits) from \widehat{C} to E in (a), (b), and (c), in each case the variation of information is different (0, 2 and 4 bits respectively). Similarly, in spite of an equal variation of causal information (4 bits) in (c) and (d), in both cases the mutual information is different (2 and 1 bits respectively).

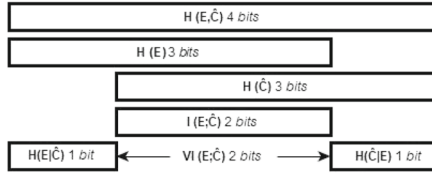
Considering the two dimensions of causal specificity, each of which is captured by a different causal information-theoretic measure, I propose that a high (low) causal specificity of C on E amounts to both a low (high) variation of causal information coupled with a high (low) mutual causal information from \widehat{C} to E . On the one hand, a low variation of causal information from \widehat{C} to E implies that the mapping between C and E is close to a bijective one (and bijective when the variation of information is 0). However, it does not permit telling, in and of itself, whether the number of values of C having an influence on the values of E is high. On the other hand, a high mutual causal information from \widehat{C} to E implies that a high number of values of C have an influence on the values of E . However, it does not permit telling whether the mapping between C and E is a bijective one.

6 Revisiting Griffith's et al. example

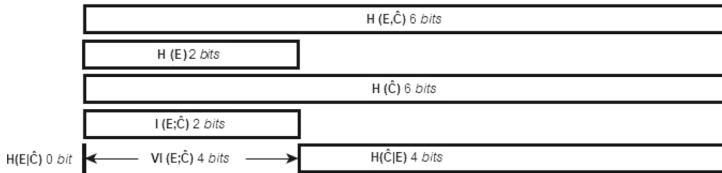
In Section 2, I presented the example of a DNA sequence and an RNA polymerase producing an mRNA, from Griffiths et al., which motivated the development of their



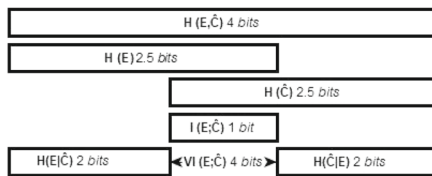
(a) Bijective causal relationship with 2 bits of mutual causal information and a nil variation of causal information.



(b) Non-bijective causal relationship with 2 bits of mutual causal information (same range of causal influence as in (a)) and the same amount of specificity of the cause and specificity of the effect (1 bit) (in total 2 bits of variation of causal information).



(c) Non-bijective causal relationship in which the one-to-one unspecificity is entirely due to unspecificity of the cause for the effect is multiply realized (4 bits of variation of causal information), and with 2 bits of mutual causal information (same range of causal influence as in (a))



(d) Non-bijective causal relationship with 1 bits of mutual causal information, and with the same amount of specificity of the cause and specificity of the effect (2 bits) which corresponds in total to a variation of causal information of 4 bits as in (c).

Fig. 2 Illustrations of the relationship between causal joint entropy $H(E, \hat{C})$, entropy of the cause $H(\hat{C})$, entropy of the effect $H(\hat{E})$, conditional entropies $H(\hat{C}|E)$ and $H(E|\hat{C})$, mutual causal information $I(\hat{C}; E)$ and variation of causal information $VI(\hat{C}; E)$

measure of range of influence. In their most basic example, they suppose a case in which the variable *DNA* has four possible values ($dna_1, dna_2, dna_3, dna_4$) with probability 0.25 for each. Each one of the four values of *DNA* leads at the same time

to a different value of RNA each resulting in the production of a different mRNA with probability 0.125 (rna_1 for dna_1 , rna_2 for dna_2 , and so forth), and to the value \emptyset with the same probability (see Fig. 3a). Second, they assume that the variable POL has two values, one of which leads nondeterministically (with probability 0.125) to the four different values of RNA producing an mRNA, while the second value leads to the value \emptyset with probability 0.5 (see Fig. 3b). With this scenario, they show that the range of influence of \widehat{DNA} on RNA on the one hand, and of \widehat{POL} on RNA on the other hand, is the same since both \widehat{DNA} and \widehat{POL} carry 1 bit of information to RNA . They conclude that the two relationships “are (given our working assumptions) equally causally specific”.

Yet, such is not the case if by ‘causally specific’ one refers to one-to-one specificity. In fact, doing the calculations for both relationships one finds that $VI(\widehat{DNA}; RNA) = 2 \text{ bits}$, while $VI(\widehat{POL}; RNA) = 1 \text{ bit}$. Because the relationships $DNA \rightarrow RNA$ and $POL \rightarrow RNA$ have different joint entropies, the absolute values cannot be directly compared. To be compared, each relationship must be normalized over the joint entropy of the two variables involved in this relationship. Once this is done, we obtain $NVI(\widehat{DNA}; RNA) = \frac{2}{3}$ (NVI for ‘normalized variation of information’), while $NVI(\widehat{POL}; RNA) = \frac{1}{2}$. This result indicates that \widehat{DNA} is less one-to-one specific than \widehat{POL} . By increasing the number of possible \widehat{DNA} sequences, the variation of information of \widehat{DNA} becomes lower than that of \widehat{POL} . With eight equiprobable sequences of DNA, we get: $NVI(\widehat{DNA}; RNA) = 0.625$ and $NVI(\widehat{POL}; RNA) = 0.6$, which leads to the conclusion that \widehat{POL} is slightly more one-to-one specific than \widehat{DNA} . With 16 equiprobable sequences of DNA, we get: $NVI(\widehat{DNA}; RNA) = 0.6$ and $NVI(\widehat{POL}; RNA) = \frac{2}{3}$, which now leads to

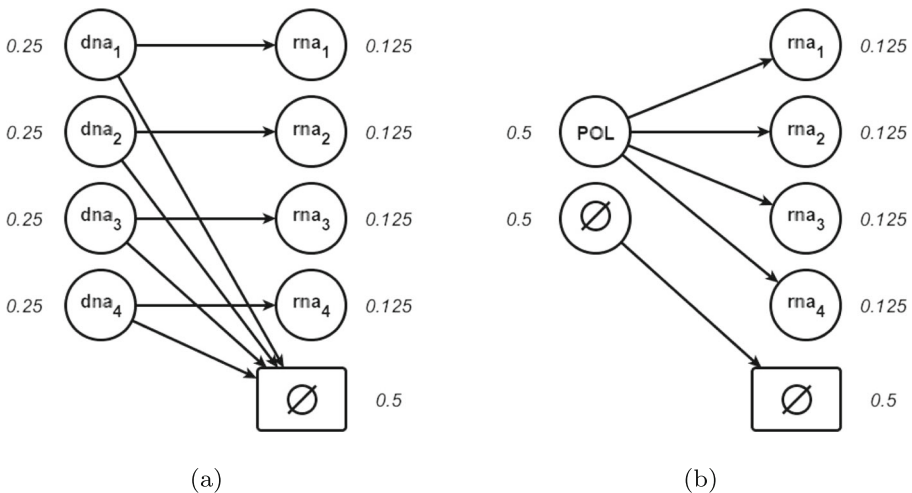


Fig. 3 Causal diagram of the mapping between DNA and RNA (a), and between POL and RNA (b) with probability distributions. Redrawn from Griffiths et al. (2015, p. 540)

the conclusion that \widehat{POL} is less causally specific than \widehat{DNA} . It can be shown that the one-to-one specificities of the two causal variables would be equal for a number of values of DNA sequences of about 9.4. Above that number the higher the number of equiprobable values of \widehat{DNA} , the higher the difference in one-to-one specificity between \widehat{POL} and \widehat{DNA} where \widehat{DNA} is the most specific cause with the theoretical limits bound for $NVI(\widehat{DNA}; RNA) = 0.5$ (maximally one-to-one specific under the assumptions of equiprobability) and for $NVI(\widehat{POL}; RNA) = 1$ (maximally one-to-one *unspecific* under the assumptions of equiprobability).

7 Conclusion

The aim of this paper was to provide a formal account of causal specificity that builds on the work of previous accounts. I started by showing that two legitimate meanings or dimensions of causal specificity exist. On the one hand, ‘causal specificity’ can mean ‘one-to-one causal specificity’ (which can itself be decomposed into specificity of the cause and specificity of the effect). On the other hand, it can mean ‘range of causal influence.’ I showed that mutual causal information only accounts for the range of causal influence dimension of causal specificity. From there, building on Griffiths et al. (2015)’s and Pocheville et al. (2017)’s work, I proposed a method to choose the right grain of description for causal relata. I then presented a causal measure of variation of information to account for the one-to-one dimension of causal specificity. I showed that variation of causal information does not account for the range of causal influence dimension of causal specificity. Finally, I showed in what sense one-to-one specificity and range of causal influence are related. This led me to the claim that causal specificity amounts to both mutual causal information and variation of causal information. Elsewhere (Bourrat *in press*), I show how variation of information can be used to give traction to the distinction between instructive and permissive causes, a distinction used by developmental biologists (see Woodward 2010, p. 317; Calcott 2017).

Acknowledgments I am thankful to the Theory and Method in Biosciences group at the University of Sydney, three anonymous reviewers, and the editors who provided useful feedback on previous versions of this manuscript. I am more particularly thankful to Arnaud Pocheville who introduced me to information theory and discussed it at length with me, and Stefan Gawronski who proofread the final manuscript. This research was supported by a Macquarie University Research Fellowship and a Large Grant from the John Templeton Foundation (Grant ID 60811).

References

- Bourrat, P. (in press). *On Calcott’s Permissive and Instructive Cause Distinction*. *Biology & Philosophy*.
- Brodeur, J. (2012). Host specificity in biological control: insights from opportunistic pathogens. *Evolutionary Applications*, 5(5), 470–480. <https://doi.org/10.1111/j.1752-4571.2012.00273.x>.
- Calcott, B. (2017). Causal specificity and the instructive–permissive distinction. *Biology & Philosophy*, 32(4), 481–505. <https://doi.org/10.1007/s10539-017-9568-0>.
- Cover, T.M., & Thomas, J.A. (2006). *Elements of information theory*. Hoboken: Wiley.
- Davidson, E.H. (2001). *Genomic Regulatory Systems: In Development and Evolution*. New York: Elsevier.

- Franklin-Hall, L.R. (2015). Explaining causal selection with explanatory causal economy: Biology and beyond. In Braillard, P.A., & Malaterre, C. (Eds.) *Explanation in biology, history, philosophy and theory of the life sciences* (pp. 413–438). Dordrecht: Springer.
- Griffiths, P.E., & Stotz, K. (2013). *Genetics and Philosophy: An introduction*. New York: Cambridge University Press.
- Griffiths, P.E., Pocheville, A., Calcott, B., Stotz, K., Kim, H., Knight, R. (2015). Measuring causal specificity. *Philosophy of Science*, 82(4), 529–555. <https://doi.org/10.1086/682914>.
- Höfler, M. (2005). The Bradford Hill considerations on causality: a counterfactual perspective. *Emerging Themes in Epidemiology*, 2, 11. <https://doi.org/10.1186/1742-7622-2-11>.
- Hill, A.B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, 58(5), 295–300.
- Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, 156(3), 563–585. <https://doi.org/10.1007/s11229-006-9138-5>.
- Jishage, M., & Ishihama, A. (1997). Variation in RNA polymerase sigma subunit composition within different stocks of Escherichia coli W3110. *Journal of Bacteriology*, 179(3), 959–963.
- Kendler, K.S. (2005). "A gene for...": The nature of gene action in psychiatric disorders, (Vol. 162. <https://doi.org/10.1176/appi.ajp.162.7.1243>.
- Korb, K.B., Nyberg, E.P., Hope, L. (2011). A new causal power theory. In Illari, P.M., Russo, F., Williamson, J. (Eds.) *Causality in the Sciences* (pp. 628–652). Oxford: Oxford University Press.
- Langman, R.E. (2000). The specificity of immunological reactions. *Molecular Immunology*, 37(10), 555–561. [https://doi.org/10.1016/S0161-5890\(00\)00083-3](https://doi.org/10.1016/S0161-5890(00)00083-3).
- Machado, C.A., Robbins, N., Gilbert, M.T.P., Herre, E.A. (2005). Critical review of host specificity and its coevolutionary implications in the fig/fig-wasp mutualism. *Proceedings of the National Academy of Sciences*, 102(suppl 1), 6558–6565. <https://doi.org/10.1073/pnas.0501840102>.
- Meilă, M. (2003). Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines* (pp. 173–187). Berlin: Springer.
- Pearl, J. (2009). *Causality: Models, reasoning and inference*, 2nd edn. New York: Cambridge University Press.
- Pearl, J., Glymour, M., Jewell, N.P. (2016). *Causal Inference in Statistics: A Primer*. New York: Wiley.
- Pocheville, A., Griffiths, P.E., Stotz, K. (2017). Comparing causes – an information-theoretic approach to specificity, proportionality and stability. In Leitgeb, H., Niiniluoto, I., Sober, E., Seppälä, P. (Eds.) *Proceedings of the 15th Congress of Logic, Methodology and Philosophy of Science*. London: College Publications.
- Rothman, K.J., & Greenland, S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health*, 95(S1), S144–S150. <https://doi.org/10.2105/AJPH.2004.059204>.
- Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423.
- Stone, J.V. (2015). *Information Theory: A Tutorial Introduction*. UK: Sebtel Press.
- Suzuki, H. (2015). *How Enzymes Work: From Structure to Function*. Boca Raton: CRC Press.
- Waters, C.K. (2007). Causes that make a difference. *The Journal of Philosophy*, 104(11), 551–579.
- Weber, M. (2006). The central dogma as a thesis of causal specificity. *History and philosophy of the life sciences*, 28(4), 595–609.
- Weber, M. (in press). Causal Selection versus Causal Parity in Biology: Relevant Counterfactuals and Biologically Normal Interventions. In Waters, C.K., & Woodward, J. (Eds.) *Philosophical Perspectives on Causal Reasoning in Biology, no. XXI in Minnesota Studies in Philosophy of Science*. Minneapolis: University of Minnesota Press.
- Weiss, N.S. (2002). Can the "Specificity" of an Association Be Rehabilitated as a Basis for Supporting a Causal Hypothesis? *Epidemiology*, 13(1), 6–8.
- Woodward, J. (2003a). *Making things happen: a theory of causal explanation*. New York: Oxford University Press.
- Woodward, J., & Hitchcock, C. (2003b). Explanatory generalizations, part i: a counterfactual account. *Noûs*, 37(1), 1–24.
- Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, 25(3), 287–318.
- Woodward, J. (2013). Causation and Manipulability. In Zalta, E.N. (Ed.) *The Stanford Encyclopedia of Philosophy, winter 2013 edn*.