Logic, Methodology and Philosophy of Science

Proceedings of the Fifteenth International Congress

Hannes Leitgeb
Ilkka Niiniluoto
Päivi Seppälä
Elliott Sober
Editors

14 Comparing causes – an information-theoretic approach to specificity, proportionality and stability

ARNAUD POCHEVILLE [*], PAUL E. GRIFFITHS [†] AND KAROLA STOTZ [‡]

Abstract. It would be useful if the interventionist account of causation, in addition to distinguishing causes from non-causes, could define various desirable properties of causal relationships. Amongst these properties are specificity, proportionality and stability. In earlier work we offered an information theoretic analysis of causal specificity, using an approach which parallels existing work in complex systems science. Here we extend this approach to proportionality and stability. First, we show that the interventionist criterion of causation, 'minimal invariance', is formally equivalent to non-zero specificity. We then show that there are natural, information theoretic ways to explicate the distinction between potential and actual causal influence. With these foundations in place we show that there is a natural information-theoretic approach to describing causal variables that explicates the idea that causes should be proportional to their effects. Then we draw a clear distinction between two ideas in the existing literature, the range of invariance of a causal relationship and its stability. Range of invariance is simply specificity. Stability concerns the effect of additional variables on the relationship between some focal pair of cause and effect variables. We show

[*]Department of Philosophy and Charles Perkins Centre, The University of Sydney, NSW2006, Australia
[†]Department of Philosophy and Charles Perkins Centre, The University of Sydney, NSW2006, Australia
[‡]Department of Philosophy, Macquarie University, NSW 2109, Australia

that in an information theoretic framework there is an important distinction between the extent to which these additional variables influence the effect and the extent to which they influence the relationship between the focal cause and effect variable. We show how to measure the influence of additional variables in both these respects. The overall result of this work is to provide precise explications of a whole family of intuitive notions associated with the interactionist account of causation. In principle, these properties can now all be measured on a causal graph. The information theoretic approach has substantial technical limitations, however, as we discuss towards the end of the paper. The real value of our work lies as much in the way it reveals the ambiguity and equivocation in earlier, qualitative discussions as in the actual measures we construct.

Keywords: causality, intervention, invariance, specificity, stability.

1 Invariance and causal explanation

The interventionist approach to causal explanation is based on the insight that "causal relationships are relationships that are potentially exploitable for purposes of manipulation and control" (Woodward, 2010, p. 314). Interventionists approach causation via the relationships between the variables that characterise an organised system. These relationships can be represented by an acyclic directed graph. In such a graph, variable C is a cause of variable E when a suitably isolated manipulation of C would change the value of E. With suitable restrictions on the idea of 'manipulation' this test provides a criterion of causation, distinguishing causal relationships between variables from merely correlational relationships (Woodward, 2003, pp. 94–107).

The interventionist account only applies to 'change-relating' generalisations, where at least one intervention upon C will produce some change in E. Generalisations which are not change-relating are not candidates to provide causal explanations. Non-change-relating generalizations may state the impossibility of certain affairs: nothing can be accelerated past the speed of light. Or they may relate an outcome to a reliable but irrelevant antecedent: men who take birth control pills will never become pregnant (Woodward 2000, 206f).

Change-relating generalisations provide causal explanations in virtue of being invariant under interventions rather than because they hold widely in nature, or have nomological force as traditionally conceived (Woodward, 2003, p. 16):

> [E]xplanation has to do with the exhibition of patterns of counterfactual dependence describing how the systems whose behavior we wish to explain would change under various conditions. ... Explanatory generalizations allow us to answer what-if-things-had-been different questions: they show us what the value of the explanandum variable depends upon. (Hitchcock & Woodward, 2003, pp. 182–183)

Invariance under intervention simply means that the relationship between variables C and E continues to hold when interventions are made on C.

> I will say that a generalization is invariant simpliciter if and only if (i) the notion of an intervention is applicable to or well-defined in connection with the variables figuring in the generalization (...) and (ii) the generalization is invariant under at least some interventions on such variables. ... To count as invariant it is not required that a generalization be invariant under all interventions. (Woodward, 2000, p. 206)

The idea of invariance is sometimes expressed in terms of the 'stability' of the generalization:

> A generalization is invariant if (i) it is ... change-relating and (ii) it is stable or robust in the sense that it would continue to hold under a special sort of change called an intervention. (Woodward, 2000, p. 198)

However, as we will shortly see, it is more convenient to reserve the term 'stability' for a different idea associated with the interventionist account.

Woodward makes a clear distinction between the actual criterion of causation and various desirable properties of causal relationships. The criterion of causation is 'minimal invariance' – invariance in the face of at least one possible intervention. A wider range of invariance is a desirable property of causal relationships: a relationship that holds for more values of C and E is a more powerful means of intervention. However, while a minimally invariant relationship may be less useful, it is not less causal.

'Specificity' is another desirable property of causal relationships. The intuitive idea behind specificity is that interventions on C can be used to produce any one of a large number of values of E, providing what Woodward terms "fine-grained influence" over the effect variable (Woodward, 2010, p. 302).

'Proportionality' is a further desirable feature of causal relationships, or, more accurately, of how causal relationships are described:

> ... causal description/explanation can be either inappropriately broad or general, including irrelevant detail, or overly narrow, failing to include relevant detail. (Woodward, 2010, pp. 296–7)

Woodward provides several striking example where a causal explanation is weakened because the choice of variables suffers from one of these vices. Saying that one person went bungy-jumping whilst another did not because only one has a 'gene for bungy-jumping' is less explanatory than saying that only one has a gene associated with risk-seeking behavior. The former explanation excludes important information that the latter provides.

'Stability' is a final desirable property of causal relationships. Whilst invariance concerns the relationship between C and E, stability concerns the relationship between other variables and that relationship. Intuitively, C is a stable cause of E if it continues to cause E across some range of values of other variables Z, W, etc. These other

variables are sometimes referred to as 'background' variables. There is much more to be said (and settled) about stability and its relationship to invariance, as we will see below.

In earlier work with other collaborators we have developed an information-theoretic approach to measuring the specificity of causal relationships within the interventionist framework (Griffiths et al., 2015). In this paper we extend that approach to (1) explore the relationship between invariance and specificity, (2) distinguish between potential and actual causal influence, (3) explicate the idea of proportionality, (4) distinguish invariance from stability, (5) draw a further distinction between the stability of an effect the stability of the relationship between cause and effect, and (6) show how to measure both forms of stability. We conclude by discussing the limitations of an information-theoretic approach.

2 Minimal invariance and specificity

In earlier work we noted that specificity is not entirely independent of the criterion of causation (Griffiths et al., 2015). This is a straightforward consequence of our measure of specificity, which formalises the idea that, other things being equal, the more a cause specifies a given effect, the more knowing the value set for the cause variable will inform us about the value of the effect variable. This idea led us to propose a simple measure:

> Spec: the specificity of a causal variable is obtained by measuring how much mutual information interventions on that variable carry about the effect variable.[1]

The mutual information of two variables is simply the redundant information present in both variables. Where $H(X)$ is the entropy of X (see Appendix), the mutual information of X with another variable Y, or $I(X;Y)$, is given by:

$$I(X;Y) = H(X) - H(X|Y)$$

Mutual information is not in itself a suitable measure of causal influence. It is symmetrical, that is $I(X;Y) = I(Y;X)$, and variables can share mutual information without being related in the manner required by the interventionist criterion of causation. However, any variables that satisfy the interventionist criterion of causation will show some degree of mutual information between *interventions* and effects. If $C \rightarrow E$ is minimally invariant, that is, invariant under at least one intervention on C, then $I(do(C);E) > 0$, where $do(C)$ means that the value of C results from an intervention

[1] This measure has been independently proposed in cognitive sciences by Tononi et al. (1999) and in computational sciences by Korb et al. (2009). For related measures see also Ay & Polani (2008), Janzing et al. (2013). Ay and Polani's measure captures what we call SAD below.

on C (Pearl, 2009). To simplify writing, we will from now on represent the $do(\ )$ operator by a hat on the variable: $do(X) \equiv \widehat{X}$.[2] So our measure of specificity does not simply measure the mutual information between variables C and E. Instead, it measures the mutual information between interventions on the variable C and the variable E. This is not a symmetrical measure because the fact that interventions on C change E does not imply that interventions on E will change C: in general, $I(\widehat{C};E) \neq I(\widehat{E};C)$.[3] Furthermore, if any pair of variables $\{C,E\}$ satisfies the interventionist criterion of causation, with C being a cause of E, then C will have some degree of specificity for E. So minimal invariance is equivalent to non-zero specificity.

This raises the obvious further question of how the *degree* of specificity of a causal relationship relates to its *range* of invariance – the range of values of the variables across which a causal relationship holds. Marcel Weber has argued in qualitative terms that the degree of specificity is just the same thing as its range of invariance (Weber, 2006). Woodward questioned Weber's proposed equivalence because a causal relationship might hold across a large range of invariance but fail to be bijective, and thus to offer the sort of fine-grained control associated with the idea of specificity: "a functional relationship might be invariant and involve discrete variables but not be 1–1 [injective] or onto [surjective] " – that is, it might fail to be bijective (2010, p. 305 fn 17). In our earlier paper we argued that measuring the mutual information between two variables is a good way to formalize Woodward's idea that the mapping between the cause and effect may 'approximate a bijection'. We then showed that with a slight correction corresponding to Woodward's caveat, Weber is correct. He is correct because the mutual information between cause and effect variables will typically be greater when these variables have more values, simply because the entropy of both variables is higher. Woodward's caveat corresponds to the fact that it is not enough to increase the number of values of a cause variable unless the additional values of the cause map onto distinct values of the effect. Our measure of specificity captures both points. Increasing the entropy of the cause variable will not increase mutual information when no additional entropy in the effect variable is captured. We can see this by contrasting the cases in Figures 14.1 and 14.2.

In fact, we would argue, it does not really make sense to say that the relationship $C \rightarrow E$ in Figure 14.2 has a greater range of invariance than $C \rightarrow E$ in Figure 14.1. The variable C merely has a large number of nominal values. The appropriate way to divide a causal variable into discrete states for the purposes of an interventionist account of causal explanation is to group together states that make the same difference. A description of the variable that does not respect this constraint is effectively

[2] We take this convention from related work in computer sciences applying information theory to causal modeling (see fn above, Ay & Polani, 2008; Lizier & Prokopenko, 2010).

[3] These quantities can be equal if and only if the two variables are not causally connected. Indeed, at least one of these quantities is null since C and E are variables in a causal graph: if C causes E, E can't feed back on C (causal graphs are acyclic, see Pearl (2009)). Thus the two quantities can be equal if and only if they are both null.

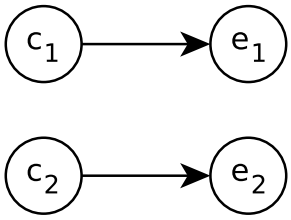

Figure 14.1: Causal mapping showing a bijection between causal values and effect values. Complete ignorance (maximum entropy) obtains when each value of the effect has a probability of $\frac{1}{2}$ before intervening on the value of the cause: $H(E) = -\sum_{j=1}^{2} p(e_j) \log_2 p(e_j) = -\sum_{j=1}^{2} \frac{1}{2} \log_2(\frac{1}{2}) = 1$ bit. After knowing the value set for the cause (c_1 or c_2), the effect is fully specified and the conditional entropy is: $H(E|\widehat{C}) = -\sum_{i=1}^{2} p(\widehat{c}_i) \sum_{j=1}^{2} p(e_j|\widehat{c}_i) \log_2 p(e_j|\widehat{c}_i) = -\sum_{i=1}^{2} \frac{1}{2} \sum_{j=1}^{2} 1 \log_2(1) = 0$ bit. The information gained by knowing the cause can be obtained by measuring the difference between the entropy before and the entropy after intervening to set the value of the cause. This quantity is the mutual information between E and $\widehat{C}$: $I(E;\widehat{C}) = H(E)$–$H(E|\widehat{C}) = 1$ bit.

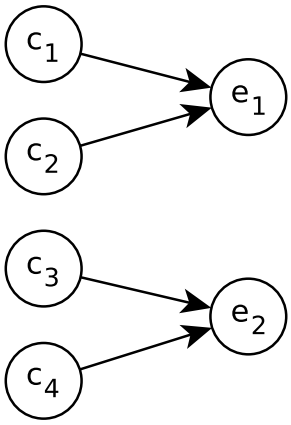

Figure 14.2: Here, different values of the cause lead to the same outcome. As in Figure 14.1, $H(E) = 1$bit. Although here two values of the cause can lead to the same effect, intervening to set the value of the cause fully specifies the value of the effect just as effectively as it does in Figure 14.1. Therefore, the difference in uncertainty about the effect between before and after intervening to set the value of the cause is the same: $I(E;\widehat{C}) = H(E)$–$H(E|\widehat{C}) = 1 - 0 = 1$bit.

a gerrymandered description, as we discuss in Section 4. So, in line with Weber's proposal, the range of invariance of a causal relationship is simply the specificity of that relationship (assuming for simplicity an equal weighting of values).

In this section we have argued that both the interventionist criterion of causation – minimal invariance – and the desirable property of having a greater range of invariance can be assessed using our measure of specificity (Spec). The information-theoretic framework we have adopted allows a quantitative formulation of these two key elements of the interventionist framework. However, the information theoretic approach requires us to specify a probability distribution over the cause, something that earlier, qualitative discussions seemed to be able to do without.

Our measure of specificity (Spec) depends on what probability distribution we choose to impose on the the causal variable C, as well as on the mapping from C to E. In our earlier paper we showed that this is very much a feature and not a bug of our measure. As we will now discuss, specificity measured with different distributions over C corresponds to different properties that are of interest to the philosophy of causation (Griffiths et al., 2015).

3 Actual and potential difference-making

Measuring Spec with different probability distributions over $\widehat{C}$ corresponds to different views of causal specificity in the existing, qualitative literature.

One way to measure specificity corresponds to Woodward's (2010) characterisation of fine-grained influence (INF). In his presentation the value of C depends only on interventions by an idealised agent. Since the aim is to characterise how one variable causally depends on another, we assume that this agent does not favour one value over another, so that every value is equiprobable. The distribution of values of C is therefore the maximum entropy distribution:

INF: $I(\widehat{C};E)$, where the distribution of $\widehat{C}$ has maximum entropy.

In our earlier work we suggested that INF was a good measure of the potential of C to causally affect E (Griffiths et al., 2015). However, another non-arbitrary choice is to construct a distribution which maximizes specificity. Such a distribution does not necessarily maximize the entropy of the cause variable (see Korb, Hope, & Nyberg, 2009).

MaxSpec: $I(\widehat{C};E)$, where the distribution of $\widehat{C}$ maximises Spec.

One formal advantage of MaxSpec is that it is insensitive to finer redescription of the variables. MaxSpec is unaffected if we divide C or E into a greater number of nominal values.

Whereas INF measures how much influence C exerts on E in an unbiased set of intervention experiments, MaxSpec measures how much influence C exerts on E under

ideal conditions. This is the 'causal power' of C with respect to E (Korb et al., 2009) and can also be thought of as a measure of C's *potential* influence on E. We are inclined to think MaxSpec a better explication than INF of the intuitive idea that a system has an intrinsic causal structure and that this structure is independent of how the system operates on some particular occasion. Understanding causal connections in this sense is a central aim of science – seeking to understand how the parts of a mechanism are connected to one another, rather than how often each connection is used or whether they are used on a particular occasion.

A different view of causal specificity has been advocated by Kenneth Waters (2007). Waters draws attention to contexts in which scientists are only interested in the actual causes of differences in some population, situations in which, he argues, they seek to characterise the causes which are 'specific actual difference makers' in that population (SADs). In earlier work we argued that this amounts to measuring *Spec* when C takes the distribution it has in the actual population. Although Water's stresses the *observed* distribution of properties in a population, his discussion makes it clear that he intends SAD to be a conception of causation, not merely of correlation, so rather than measuring the mutual information between the actual distributions of C and E, we need to imagine a set of interventions that create the same distribution of values of C that we see in the population, hence:

> SAD: $I(\widehat{C};E)$ where the distribution of $\widehat{C}$ is identical to the actual distribution of C in some population.[4]

If we take Spec to correspond to SAD rather than MaxSpec we get a rather different picture of causation from the one we sketched in Section 2. First, it will no longer be true that all causes have some degree of specificity for their effects. It may simply be that the range of C within which there is a relationship between C and E does not occur in the population from which we derive the distribution of C. In other words, the 'experiment of nature' does not include the experiment that reveals how E depends on C. For the same reason, under the SAD interpretation, the degree of specificity of C for E may not correspond to the range of invariance of the relationship $C \rightarrow E$.

We interpret SAD as a measure of a complementary idea to potential causal influence, namely actual causal influence – how much difference a cause *actually* makes to an effect. For example, in a causal graph representing a firing squad, the potential causal influence of the variable SHOOT with respect to the variable DIE, as measured by MaxSpec, will be greater than that of the variable SAY BOO, but SAY BOO will

[4] In addition, Marcel Weber (2013) has argued that in the biological sciences specificity should be assessed using a wider range of values of C than actually occur in any given population, but not all possible values of C. He suggests we should restrict ourselves to 'biologically normal' values of C. We interpret this to mean that C should be restricted to the range of variation that could be produced by known mechanisms operating on the timescale of whatever process we are trying to study. We have suggested that within that range, $\widehat{C}$ should conform to the maximum entropy distribution and named this additional flavour of specificity REL for relevant specificity (Griffiths et al., 2015). But it is also possible construct a version of relevant specificity based on the MaxSpec measure.

have greater actual causal influence on DIE than SHOOT does in a population where more prisoners die from fright than from bullets. The same idea has been termed 'information flow' (Ay & Polani, 2008). Ay and Polani explicitly conditionalise on a set of background variables S. We intervene to set S equal to what we observe in nature, derive a distribution for A and then ask what part of the correlation between A and B results from a causal relationship between them. To do this we need to import causal information derived from intervening on A, but the distribution over A whose effect we are assessing is observed, not imposed by intervention. Information flow is intended to measure the causal impact of variables on one another in a specific set of data, or what we have called actual causal influence (Lizier & Prokopenko, 2010).

In Section 2 we analysed the relationship between specificity and invariance. In this section we showed that specificity can be used to measure both potential and actual causal influence. In Section 4 we move on from our examination of specificity to examine a second desirable property of causal relationships, proportionality.

4 Proportionality

The proportionality of cause to effect is a matter of "whether the cause and effect are characterized in a way that contains irrelevant detail" (Woodward, 2010, p. 287) This idea has been discussed extensively in the philosophy of causation, where it has been explained via examples and qualitative characterisations:

> Yablo suggests that causes should "fit with" or be "proportional" to their effects—proportional in the sense that they should be just "enough" for their effects, neither omitting too much relevant detail nor containing too much irrelevant detail. (Woodward, 2010, p. 297)

In an effort to characterise the idea more precisely, Woodward has characterised it as a 'proportionality constraint' on the mapping between value of the cause and values of the effect.

> (P) There is a pattern of systematic counterfactual dependence (with the dependence understood along interventionist lines) between different possible states of the cause and the different possible states of the effect, where this pattern of dependence at least approximates to the following ideal: the dependence (and the associated characterization of the cause) should be such that (a) it explicitly or implicitly conveys accurate information about the conditions under which alternative states of the effect will be realized and (b) it conveys only such information – that is, the cause is not characterized in such a way that alternative states of it fail to be associated with changes in the effect. (Woodward, 2010, p. 298)

We stress that Woodward is not adding an additional condition to his criterion of causation. Like specificity, proportionality is meant to enrich the theory of causation

by capturing why some causal facts may legitimately be of more interest to us than others, and thus may be highlighted in our explanations whilst other causal facts are omitted. Highly specific causes provide more precise control over an effect, and explain outcomes with greater precision. Proportional descriptions of causes provide us with all and only the information relevant to intervening or explaining with those causes.

We are now in a position to spell out the relationship between proportionality and specificity, and by doing so to define proportionality more precisely. If we choose a set of values for a causal variable, and a probability distribution over those values, which maximizes specificity, then, by definition, we cannot have omitted any relevant detail, since we have explained as much of the differences in the effect variable as possible. How can we make sure not to include any irrelevant detail? This is performed by minimizing the entropy of the cause variable by aggregating values which make the same difference, whilst maintaining its specificity: the less the entropy of the cause, the less information about the cause we have included in our explanation. Ideal proportionality is thus achieved when the cause is described in a way which minimizes its entropy and still maximizes specificity.

We can see how this works with Yablo's original example (1992, p. 4). A pigeon called Sophie has been trained to peck in response to any stimulus which is some shade of red. Yablo contrasts two explanations:

1. Sophie pecked because she was exposed to a red stimulus
2. Sophie pecked because she was exposed to a scarlet stimulus

Yablo suggests that 1 is a better causal explanation than 2. Like many philosophical thought experiments, this one is underspecified. We have two variables, P, with the values 'peck' and '$\sim$peck', and S. What values should S take? The combined probability of all values of a random variable must sum to one, so let us take the values of S to be the actual colour chips available in the laboratory, which neatly avoids the problem that birds do not perceive human spectral colours like red and scarlet. We stipulate that there are colour chips of more than one shade of red, and of some non-red shades. Finally, we stipulate that Sophie has been trained to peck at each of the colour chips that is a shade of red, giving us a causal graph in which P has the value 'peck' if and only if S has one of the values which is a shade of red.

We now construct the maximum specificity distribution, in this case making the combined weight of probablity on the red values equal to that on the non-red values. The graph we have described resembles that in Figure 14.2 above, and is exactly that graph if there are just two red and two non-red values. If we coarse-grain the values of our variable, so that S now has just two values, red and $\sim$ red, then we get the graph in Figure 14.1. S now has the same specificity as before, but the entropy of S has been reduced from 2 bits to 1 bit. This is is the optimally proportional way to divide the variable S into discrete values. No more specificity can be obtained by fine-graining and any further coarse-graining will reduce the specificity.

The artificiality of the example produces some problems. Whilst this is the optimal way to discretise the variable S for this single experiment with Sophie, it is not optimal for a wider experimental program! A better example of proportionality might be an experimentalist who sets her values for S to correspond to the distinctions in the pigeon's own tetrachromatic spectrum, since this would make S express only the 'differences that can make a difference' to the pigeon's behavior.

Woodward's other example of a failure of proportionality is taken from psychiatric geneticist Kenneth Kendler:

> To illustrate how this issue of the appropriateness of level of explanation may apply to our evaluation of the concept of "a gene for..." consider these two "thought experiments":
>
> Defects in gene X produce such profound mental retardation that affected individuals never develop speech. Is X is a gene for language?
>
> A research group has localized a gene that controls development of perfect pitch ... Assuming that individuals with perfect pitch tend to particularly appreciate the music of Mozart, should they declare that they have found a gene for liking Mozart?
>
> For the first scenario, the answer to the query is clearly "No." Although gene X is associated with an absence of language development, its phenotypic effects are best understood at the level of mental retardation, with muteness as a nonspecific consequence. X might be a "gene for" mental retardation but not language.
>
> Although the second scenario is subtler, if the causal pathway is truly gene variant $\rightarrow$ pitch perception $\rightarrow$ liking Mozart, then it is better science to conclude that this is a gene that influences pitch perception, one of the many effects of which might be to alter the pleasure of listening to Mozart. It is better science because it is more parsimonious (this gene is likely to have other effects such as influencing the pleasure of listening to Haydn, Beethoven, and Brahms) and because it has greater explanatory power. (Kendler 2005 p. 1249–50, quoted in Woodward 2010 p. 300–301)

The grain of description of the cause variable in these cases is fixed by the technology used to detect the genetic variant. The failure of proportionality is supposedly the result of describing the *effect* in too fine-grained a manner. But 'proportionality' here is not the same phenomena that we identified in the pigeon-pecking case, and nor is it really a matter of fine- versus coarse-graining. There are two alternatives to saying that the genetic variant is a gene for language or a gene for liking Mozart. The first is to say that it is a gene for an intervening variable, a variable which is linked to a host of behavioral and cognitive deficits in the former case or a host of musical preferences and abilities in the later. The second is to that say it is a pleiotropic gene with effects on many phenotypes. The first option corresponds to redrawing the causal graph by

inserting an intervening variable, not to redescribing the effect variable. 'Perfect pitch' and 'Liking Mozart' do not stand to one another as variable and value but as cause and effect. The second alternative, pleitropy, also amounts to adding connections between the cause and additional variables, not coarse-graining the original variable: liking Mozart is not an instance of liking Haydn.

In these two later examples it is the causal graph itself that is too 'coarse grained' rather than one of the variables. That is consistent with Woodward's original characterisation of proportionality as 'neither omitting too much relevant detail nor containing too much irrelevant detail', but it reveals an ambiguity in that description. Choosing which variables to include in the graph and choosing how finely or coarsely to discretize a variable are clearly very different problems and it is better to keep them distinct. We therefore prefer to define 'Proportionality' more narrowly[5]:

> Proportionality constraint: Given an effect variable E that is a target of intervention or causal explanation, a causal variable C should be discretised so as to minimise the entropy of C whilst maximising specificity for E.

The main philosophical dispute in which the notion of proportionality has figured concerns whether lower-level, reductive explanations of phenomena are always superior to high-level explanations of the same phenomena. Carl Craver, for example, has argued that in some cases the lower-level explanation merely recognises additional differences that make no difference (e.g. Craver, 2007, Ch 6). Whether this argument is successful or not, our version of the proportionality constraint seems suitable to capture the intention behind it.

In this section we have added an account of proportionality to our earlier account of specificity.[6] In the remaining sections we tackle the concept of 'stability'.

5 Stability: Some clarifications

The interventionist account of causation aims to identify causes that "are likely to be more useful for many purposes associated with manipulation and control" (Woodward, 2010, p. 315). One aspect of this is the 'stability' of causal relationships.

> Among change-relating generalizations, it is useful to distinguish several sorts of changes that are relevant to the assessment of invariance. First, there are changes in the background conditions to the generalization. These are changes that affect other variables besides those that fig-

[5]The other issues being discussed under the heading of 'proportionality' seem to concern what statisticians call 'model selection'.

[6]Our discussion in this section is indebted to conversations with Jun Otsuka, Pierrick Bourrat and Brett Calcott

> ure in the generalization itself. ... Second, there are changes in those variables that figure explicitly in the generalization itself... (Woodward, 2003, p. 248)

At this point some terminological stipulation is needed. We will reserve the term 'invariance' strictly for the properties of Woodward's "variables that figure explicitly in the generalization itself." Invariance characterizes the relationship between two variables, one of which can be used to intervene on the other. The invariance of that relationship is the range of values of those two, focal variables across which one can be used to intervene on the other. We showed above that the range of invariance can be captured by the degree of specificity. We will refrain from using the term 'stability' in connection with the relationship between those two focal variables. Instead, we use it strictly to describe how the relationship between those two variables is related to other variables. Stability is about of whether a causal relationship continues to hold across a range of background conditions.

Hitchcock and Woodward distinguish between two senses in which a causal generalization may be said to hold against a 'background' of other factors. In their first sense the 'background' to a causal generalization is simply everything not mentioned in the generalization. Most of the background, in this sense, is causally irrelevant. In their second sense, the 'background' consists of variables that are causally relevant to the effect but not explicitly represented in the model (Hitchcock & Woodward, 2003, p. 187). In our terms, causally relevant background conditions are additional variables that have some degree of specificity for the effect variable.

Sandra Mitchell has also made extensive use of something she calls 'stability' in an account of causal generalisations. Using 'invariance' in a broader sense, rather than in the restricted sense we have stipulated, she writes that:

> Stability for me is a measure of the range of conditions that are required for the relationship described by the law to hold, which I take to include the domain of Woodward's invariance. ... Stability does just the same work [as Woodward's invariance], however it is weaker and includes what might turn out to be correlations due to a non-direct causal relationship. But for there to be a distinction between stability and invariance, then we would have to already know the causal structure producing the correlation. (Mitchell, 2002, pp. 346–347))

Mitchell's 'stability' is a matter of whether a generalization holds across a range of values of other variables that are statistically relevant to the effect, either because they are causally relevant to it or due to confounding factors. Her treatment of stability is thus very different from Woodward's, and from ours. Mitchell's work is centrally concerned with complex systems for which there may be no practical way to reliably and fully document their causal structure. Hence she emphasises the scientific and pragmatic value of generalisations that are stable in her sense irrespective of what other, more stringent requirements they may satisfy. She also doubts the value, in her chosen context, of the distinction between the range of invariance of a relationship and

its stability.

Despite the different foci of their work, there is real disagreement between Woodward and Mitchell about what distinguishes causally explanatory relationships between variables from mere correlations. Mitchell argues that causal generalisations are explanatory to the extent that they are stable. Woodward's criterion of causation was outlined above – causally explanatory generalisations need to be minimally invariant. Nothing more is needed to make them causally explanatory, and without this property no amount of stability in Mitchell's sense will make a generalization causally explanatory. The role of stability in Woodward's account is not to provide a criterion of causation, but to identify more *useful* causal relationships. Hitchcock and Woodward remark, using invariance in the same, wider sense as Mitchell, that,

> Invariance under changes in background conditions does not render a generalization explanatory; yet *greater* invariance [our stability] under changes in background conditions can render one generalization *more* explanatory than another. ... Briefly, if G is sensitive to changes in background conditions, that is because it has left out some variable(s) upon which the explanandum variable depends. (Hitchcock & Woodward, 2003, p. 187, italics in original)

As Hitchcock and Woodward emphasise, genuine background conditions are factors that could, and often should, be explicitly represented in a causal model:

> [C]laims about the invariance of a relationship under changes in *background conditions* are transformed into claims about invariance under interventions *on variables figuring in the relationship* through the device of explicitly incorporating additional variables into the relationship. (Hitchcock & Woodward, 2003, p. 188, italics in original)

One further distinction is needed to think clearly about the relationship between causal generalisations and background conditions. There are two different things with respect to which a variable may act as a (genuine) background condition. A variable may be a background condition with respect to the outcome of a causal process – the effect variable. Or it may be a background condition with respect to a causal relationship in the model. So if C is a cause of E, a third variable Z may affect the value of E, but it may also affect the way in which C is related to E. These two effects of Z are closely connected, and it may not be clear to the reader that they are distinct, but they are in fact importantly different (see Section 6 and Appendix).

6 Stability of causal relationships

When we speak of the 'stability' of some relationship $C \rightarrow E$ we often have in mind, not the influence of background variables on E, but whether the relationship $C \rightarrow E$ itself changes across a range of background conditions. For example, alternative

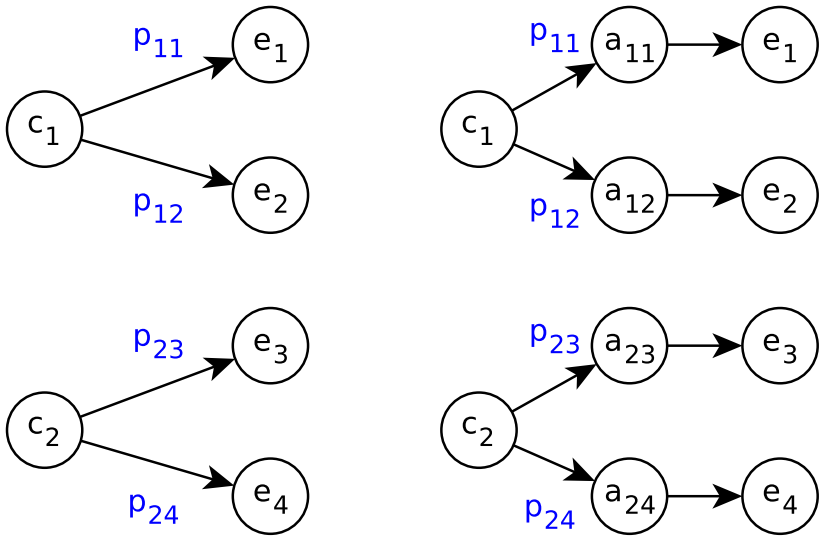

Figure 14.3: On the left, a causal mapping relates values of a nominal causal variable to values of a nominal effect variable. In this example, each causal value can lead to two proper (and incompatible) effect values, each arrow being associated with a probability $p_{ij} = p(e_j|\widehat{c_i})$. On the right, 'arrows' are now explicitly represented as values of a new variable, A, which represents the mapping between C and E.

splicing of genes depends on splicing regulatory elements (SREs), short nucleotide sequences in the pre-mRNA that bind protein factors that either activate or repress the use of adjacent splice sites. The causal relationship between the presence of an SRE and binding of its protein can be affected by the surrounding RNA sequence, because the shape of the whole RNA molecule can render the SRE more or less accessible to the factors for which it has an intrinsic binding affinity. Hence the same sequence can act as an SRE in one organism, but not in the orthologous gene of another organism, due to changes elsewhere in the gene (Wang & Burge, 2008). The molecular facts in these cases are very naturally represented as a focal causal relationship in which C is the sequence of the SRE and E is whether the protein binds or not, plus one or more background variables representing the structure of rest of the gene, which can interfere with that focal causal relationship.

It is stability and instability in this sense that we now proceed to analyse. Our aim in this section is not to come up with a definitive measure of causal stability for every purpose, but rather to show how to relate the idea of stability of causal relationships to our measure of causal specificity.

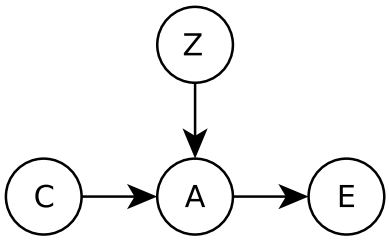

Figure 14.4: Causal graph with a variable representing the arrows A mapping C to E as they are affected by Z. (We draw reader's attention to the fact that this diagram is a causal graph relating variables, not a mapping relating values.)

To start with, let's consider a causal relationship $C \rightarrow E$ represented by a mapping between values of a (nominal) causal variable C to values of a (nominal) effect variable E (Fig. 14.3). Each causal value c_i can lead to one or several effect values e_j. To look at how the mapping can be influenced by a third variable, we will focus on the arrows connecting the values c_i and e_j. Each arrow a_{ij} can be defined as a couple of one causal value and one effect value. In formal terms, $a_{ij} \equiv (\widehat{c}_i, e_j)$.

When an intervention which sets C to c_i leads to e_j, we will say that the causal arrow a_{ij} has been instantiated, or that the variable A (for arrow – Figure 14.4) has taken the value a_{ij}.[7] The mapping between the values of the causal variable and the values of the effect variable is the set of these causal arrows, together with their associated conditional probabilities.

Now, let's consider that the mapping between C and E is somehow unstable with respect to a background variable Z. That is, Z makes the instantiation of some arrows more or less probable than it would be otherwise. We now treat the instantiations of the arrows a_{ij} as the events that are to be explained, and Z as the variable explaining them.

We first consider the arrows stemming from one causal value. Let's intervene on C to set it to value $\widehat{c}_1$. Given $\widehat{c}_1$, we look at how intervening on Z changes the probability of the arrows $a_{1j} : \widehat{c}_1 \rightarrow e_j$ that will be instantiated.[8] The amount of change can be measured by the mutual information between $\widehat{Z}$ and the variable A given c_1, that is, in formal terms, $I(A;\widehat{Z}|\widehat{c}_1)$.[9]

Figure 14.5 illustrates this idea. An intervention on Z has no effect on the mapping when the causal probabilities are unchanged, in which case $I(A;\widehat{Z}|\widehat{c}_1) = 0$ bit. The

[7] Because both C and E are sets of alternative events, it is axiomatic that one and only one arrow is instantiated in every intervention on the cause C. Also, because C and E are nominal variables, the composite

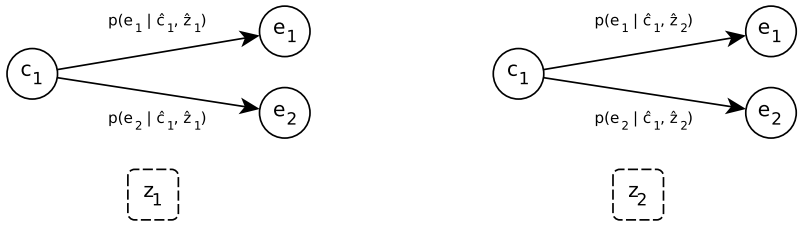

Figure 14.5: Diagram showing how interventions on Z can modify the mapping from C to E . For simplicity, only a single value of the causal variable is considered.

mapping between C and E is then maximally stable with respect to $\widehat{Z}$ (in this limiting case, Z is an irrelevant background condition, see Section 5). An intervention on Z has a maximum effect when it completely specifies the causal arrows being instantiated, that is, when one of the causal arrows is always instantiated when the intervention results in $\widehat{z}_1$ and reciprocally when $\widehat{z}_2$. In this case $I(A;\widehat{Z}|\widehat{c}_1) = 1$ bit, which is the maximum possible instability for this mapping. In between these two limiting cases stability will come in degrees.

When more than one value of C is considered (which is a necessary condition to be able to speak of C as a putative cause of E), it is reasonable to average the conditional mutual information $I(A;\widehat{Z}|\widehat{c}_i)$ over all the values of the causal variable C. The rationale for this is that causal arrows stemming from causal values that are themselves improbable (or impossible) should count less in characterizing the properties of the mapping. Calculating this average is equivalent to computing the conditional mutual information $I(A;\widehat{Z}|\widehat{C})$. This quantity characterizes how much Z affects the mapping between C and E when C is given in the background, or, in other words, the instability of the mapping with respect to Z.

However, not all mappings between C and E represent causal relationships. If C is not a causally relevant variable with respect to E, then the mapping between them is one where any value of C maps to all values of E (Fig. 14.6). The method we just outlined may nevertheless detect an effect of Z on the arrows being instantiated, but this will be due solely to the direct effect of Z on E. What we are after, however, is not just the mere effect of the variable Z on the effect E, it is rather how much the cause C and the background Z interact when they are *both* causes of E (Fig. 14.7). This, in our view,

variable A is also a nominal variable.

[8]Given $\widehat{c}_1$, the probabilities $p(a_{1j}|\widehat{c}_1)$ sum to 1.

[9]We condition on $\widehat{c}_1$ for pedagogical reasons, but it also makes philosophical sense. If $\widehat{Z}$ and $\widehat{C}$ are not independent, then we want to control for $\widehat{C}$ before assessing any effect of $\widehat{Z}$ on the arrows, as Z can be a cause C. If they are independent, conditioning makes no difference (see Appendix).

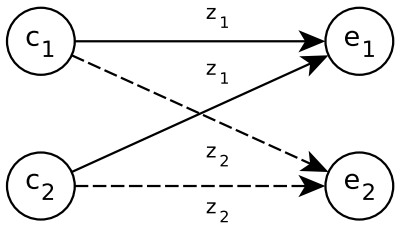

Figure 14.6: Causal mapping where Z is the only cause of E.

is what it means to talk of the causal relationship $C \rightarrow E$ depending on Z.

To measure our real target, the extent to which C and Z interact when they are both causes of C, we can use a bit of calculus and remark that the (conditional) specificity of Z for the mapping is equal to the conditional specificity of Z for the effect. In formal terms, $I(A;\widehat{Z}|\widehat{C}) = I(E;\widehat{Z}|\widehat{C})$ (see Appendix). This term embeds both the proper information coming from $\widehat{Z}$ alone, which is here equal to $I(E;\widehat{Z})$, and the information coming from the interaction between $\widehat{Z}$ and $\widehat{C}$, which is what we are after (Fig. 14.7).[10] To measure this interaction we compute the quantity $I(E;\widehat{Z};\widehat{C}) = I(E;\widehat{Z}|\widehat{C}) - I(E;\widehat{Z})$. This quantity is called the interaction information between the three variables.[11] The interaction information represents the portion of the effect of Z on the relationship between C and E that is not merely a consequence of the direct effect of Z on E.

7 Stability: Some conclusions

In Section 5, we remarked that the stability of the causal relationship $C \rightarrow E$ with respect to a background variable Z must be distinguished from the stability of E with respect to Z. We can now make this point more precise. The stability of the relationship $C \rightarrow E$ in response to changes in Z can only be reduced to the effect of Z on E if C and Z have entirely non-interactive effects on E, that is, if the interaction information is zero (see Appendix). Another way to look at this condition is that there is no inter-

[10] These components are often referred to as the unique information and the synergistic information, respectively. Another component of information is often considered: the redundant information (e.g. Williams and Beer, 2010). Decomposing multivariate information into such components is a currently debated topic (e.g. Bertschinger et al., 2013a, 2013b; Rauh et al., 2014). Here we assume that C and Z are independently manipulated and do not share any redundant information with respect to E.

[11] The interaction information is symmetrical: $I(E;\widehat{Z};\widehat{C}) = I(E;\widehat{C}|\widehat{Z}) - I(E;\widehat{C}) = I(\widehat{Z};\widehat{C}|E) - I(\widehat{Z};\widehat{C})$. In philosophical terms, there is parity, in our framework, between the causal variable C and the background variable Z: both C and Z are causal variables in the mapping from $\{C,Z\}$ to E.

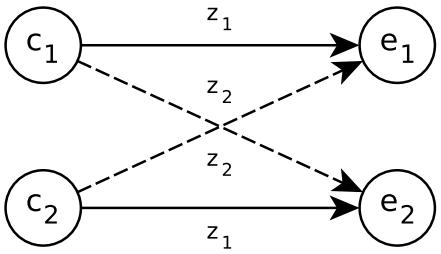

Figure 14.7: Example of interacting causes C and Z with respect to E. If the background Z is not controlled, the cause C is entirely not specific (assuming, for the ease of presentation, equiprobability between $\widehat{z}_1$ and $\widehat{z}_2$). Indeed, any intervention $\widehat{c}_1$ or $\widehat{c}_2$ can equiprobably lead to e_1 or e_2 . Thus, $I(\widehat{C};E) = 0$ bit. However, once we know the background, C is entirely specific: $I(\widehat{C};E|\widehat{Z}) = 1$ bit (assuming, for the ease of presentation, equiprobability between $\widehat{c}_1$ or $\widehat{c}_2$). The interaction information in this case is $I(\widehat{C};E;\widehat{Z}) = I(\widehat{C};E|\widehat{Z}) - I(\widehat{C};E) = 1$ bit. (By design, the same holds when Z is the focal cause variable and C is as a background variable.)

action if and only if, given that we know which value E has taken, learning the value of the background variable Z gives us no additional information about which causal arrow from C has been instantiated: interventions on Z do not cause the same result in E to be produced in a different way. This is the case for instance in Figure 14.3 but not in Figure 14.7.

It is probably worth emphasizing that a relationship can be unstable with respect to a background variable but nevertheless have a stable conditional specificity under each background condition. This comes from the fact that the background variable affects the mapping between C and E but not necessarily the properties of the mapping, of which specificity is one. In other words, changing the background may produce a new mapping, but one that is exactly as specific as the original (e.g. Fig. 14.7).

The measures of stability described in Section 6 can be reconfigured in line with the 'specific actual difference making' (SAD) version of specificity favoured by Waters (2007) using the procedure described in Section 3. The intervention distribution of a background condition, $\widehat{Z}$, is forced to correspond to the actual distribution of Z in some population. The cost will be the same as for other applications of SAD any conclusions about stability will be relevant only for the population from which the actual distribution is derived.

In Section 6 we quantified how much a causal relationship depends on interactions of

Concepts	Relationship	Measure
Specificity, Stability	$C \to E$	$I(\widehat{C};E)$
Interaction	$Z \to (C \to E)$	$I(\widehat{Z};E\|\widehat{C}) - I(\widehat{Z};E)$

Table 14.1: Information theoretic measures for specificity, stability, and interaction.

the cause with background variables.[12] But what about the reverse of this – how much influence the cause can exert irrespective of the background background variables? In information theoretic terms, the answer is quite simple. The amount of causal information which is independent from the background variables is just the mutual information between interventions on the cause and the effect, without controlling for the background variables. In other terms, a positive notion of stability is automatically captured by the idea of specificity (Table 14.1).

It goes without saying that the strengths of the information theoretic formalism, its simplicity, and the way it helps us clarify our concepts, come with limitations. The fact that we can quantify specificity and stability by the same measure in part reflects the conceptual overlap between the two: both deal with how much a cause can affect an effect across a range of background conditions. In part, however, this formal homogeneity is a by-product of using a highly constrained theoretical framework, a theme we return to in our conclusion.

8 Conclusions

Our information-theoretic framework was developed for thinking about causal specificity within the interventionist approach to causation (Griffiths et al., 2015). In this paper we have used it to analyse several other key elements of the interventionist account. In Section 2 we showed that the property of 'minimal invariance', which provides the criterion of causation in Woodward's (2003) interventionist account, is equivalent to a non-zero degree of specificity in the relationship between a cause and its effect. The 'range of invariance' of a causal relationship can be measured by the degree of specificity of the cause for its effect. Our proposed measure of specificity is the mutual information between interventions on a causal variable and observations of an effect variable:

$$\text{Spec} = I(\widehat{C};E)$$

In our earlier work we suggested that the potential of C to causally influence E should be measured by Spec with a maximum entropy distribution over $\widehat{C}$. This seems to be

[12] Z can be any set of several variables.

the natural interpretation of Woodward's conception of fine-grained influence (INF), ultimately derived from David Lewis. Here, however, we have argued that that the potential causal influence of $\widehat{C}$ on E, considered in the abstract, is better measured by constructing the distribution over $\widehat{C}$ that maximises the value of Spec (MaxSpec, see also Korb et al. (2009)).

In Section 3 we examined Water's proposal to assess specificity using only the actual variation in a cause in some population, or 'specific actual difference making' (SAD). This conception of specificity can also be expressed information-theoretically and has useful applications, as we have argued elsewhere (Griffiths et al., 2015). We showed that SAD specificity corresponds to Spec when we intervene on $\widehat{C}$ to mimic the actual distribution of $\widehat{C}$ in some population. It is instructive that different qualitative discussions of specificity correspond to different probability distributions over the causal variable. However, SAD behaves very differently from INF or MaxSpec and we interpret it as a measure of the *actual* influence of $\widehat{C}$ on E in some population, rather than of the *potential* influence of $\widehat{C}$ on E. We also suggested that another information-theoretic measure, information flow (Ay & Polani, 2008), is an alternative way to measure actual causal influence that has some advantages over SAD.

In Section 4 we argued that the controversial idea that causes should be described in a more or less fine-grained way so as to render the description of the cause 'proportional' to its effects could be made more precise in our framework. Ideal proportionality is achieved by simultaneously minimising the entropy of $\widehat{C}$ whilst maximising the specificity $I(\widehat{C};E)$. This amounts to discretising the variable $\widehat{C}$ so as to mark all and only differences that make a difference to E. We suggested that some features referred to in the qualitative literature as 'proportionality' but not captured by our proposal concern which variables to include in a causal graph in the first place, rather than the grain of description of a given variable.

In Sections 5 we suggested that the 'stability' of a causal relationship is the extent to which that relationship is not affected by additional variables, often termed background variables. We distinguished two ways in which a background variable could have an effect on a causal relationship. It might affect the value of the effect variable, or it might affect the relationship between the causal variable and the effect variable. To the best of our knowledge, this distinction has not been clearly drawn in any earlier discussions.

In Section 6 we offered an information-theoretic analysis of the instability of causal relationships. The effect of a third variable, Z on the causal relationship $C \rightarrow E$ is the effect of interventions on Z on the mapping from C to E. The greater this effect, the more unstable $C \rightarrow E$ is relative to the background variable Z. The amount of instability can be measured by the interaction information between C, Z and E. We showed that the impact of Z on $C \rightarrow E$ needs to be distinguished from the impact of Z on the value of E. The causal mutual information between Z and these two will only be equal under special conditions. The opposite of instability, the *in*sensitivity of the relationship $C \rightarrow E$ to background variables, is simply specificity of that relationship.

We believe that the work presented here adds precision to some important elements of the interventionist approach to causation and opens up many potential lines for further research. However, our use of information theory as a formal tool introduces some very severe limitations. Most importantly, we are restricted to using nominal variables. Individual values are different from one another, but not different by any amount. We are thus unable to capture the idea that highly specific relationships are 'smooth'. This might mean that the size of changes in the cause corresponds to the size of changes in the effect, for which we would need metric variables. Alternatively, it might mean that adjacent values of causes produce adjacent values of the effect, for which we would need at least ordinal variables. A related blind-spot for our approach to stability is whether changes to background variables have large, small, or negligible, impacts on a causal relationship. We can only measure *how many* changes in a background variable have *an* impact.

There are two possible responses to the intrinsic limitations of some formal framework. One is to return to a qualitative approach which can encompass the full richness of the relevant concepts, but at the price of being less clear about what constitutes that richness. That strikes us as a very high price. The other is to seek to approach different aspects of the topic using different formalisms. The interventionist framework would benefit very greatly from being given a treatment in an entirely different formalism, such as dynamical systems theory, but that is a project for another day.

9 Appendix

Here we provide a quick primer in information theory, proofs of equations cited in the text and expand on some of the ideas in Section 6.

9.1 Entropy, conditional entropy, and mutual information

We recall basic formulas of information theory. For a primer on information theory, see (Cover & Thomas, 2006). The Shannon entropy of a variable X is defined as:

$$H(X) \equiv -\sum_i p(x_i) \log_2 p(x_i)$$

The conditional entropy of a variable X knowing Y is defined as:

$$H(X|Y) \equiv -\sum_j p(y_j) \sum_i p\left(x_i \left| y_j \right.\right) \log_2 p(x_i|y_j).$$

The mutual information of two variables X and Y can be computed as:

$$I(X;Y) = H(X) - H(X|Y) = \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \left(\frac{p(x_i, y_j)}{p(x_i) p(y_j)} \right).$$

The conditional mutual information of two variables X and Y knowing a third variable Z can be computed as:

$$I(X;Y|Z) = \sum_k p(z_k) \sum_i \sum_j p(x_i, y_j|z_k) \log_2 \left(\frac{p(x_i, y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)} \right).$$

Our measure of specificity of C to E is defined as the mutual information between $\widehat{C}$ and E: $Spec(C \to E) \equiv I(\widehat{C};E)$. The formula of $I(\widehat{C};E)$ reads (see Griffiths et al., 2015):

$$I(\widehat{C};E) = \sum_i \sum_j p(\widehat{c_i}, e_j) \log_2 \left(\frac{p(\widehat{c_i}, e_j)}{p(\widehat{c_i})p(e_j)} \right).$$

9.2 $I(A;\widehat{Z}|\widehat{C}) = I(A;\widehat{Z})$ when $\widehat{C}$ and $\widehat{Z}$ are independent:

We start with the plain formula for $I(A;\widehat{Z}|\widehat{C})$:

$$\begin{aligned}
I(A;\widehat{Z}|\widehat{C}) &= \sum_i p(\widehat{c_i}) I(A;\widehat{Z}|\widehat{c_i}) \\
&= \sum_i p(\widehat{c_i}) \sum_j \sum_k p(a_{ij}, \widehat{z_k}|\widehat{c_i}) \log_2 \left(\frac{p(a_{ij}, \widehat{z_k}|\widehat{c_i})}{p(a_{ij}|\widehat{c_i})p(\widehat{z_k}|\widehat{c_i})} \right) \\
&= \sum_i \sum_j \sum_k p(\widehat{c_i})\, p(a_{ij}, \widehat{z_k}|\widehat{c_i}) \log_2 \left(\frac{p(\widehat{c_i})p(a_{ij}, \widehat{z_k}|\widehat{c_i})}{p(\widehat{c_i})p(a_{ij}|\widehat{c_i})p(\widehat{z_k}|\widehat{c_i})} \right) \\
&= \sum_i \sum_j \sum_k p(a_{ij}, \widehat{z_k}, \widehat{c_i}) \log_2 \left(\frac{p(a_{ij}, \widehat{z_k}, \widehat{c_i})}{p(a_{ij}, \widehat{c_i})p(\widehat{z_k}|\widehat{c_i})} \right)
\end{aligned}$$

Now we use $p(a_{ij}, \hat{c}_i) = p(a_{ij})$ and $p(a_{ij}, \widehat{z_k}, \widehat{c_i}) = p(a_{ij}, \widehat{z_k})$ ($\widehat{c_i}$ is necessary to obtain a_{ij}), as well as $p(\widehat{z_k}|\widehat{c_i}) = p(\widehat{z_k})$ (independence of $\widehat{C}$ and $\widehat{Z}$). We obtain:

$$I(A;\widehat{Z}|\widehat{C}) = \sum_i \sum_j \sum_k p(a_{ij}, \widehat{z_k}) \log_2 \left(\frac{p(a_{ij}, \widehat{z_k})}{p(a_{ij})p(\widehat{z_k})} \right) = I(A;\widehat{Z})$$

9.3 Conditional specificity about the mapping is conditional specificity about the effect

We can transform $I(A;\widehat{Z}|\widehat{C})$, using the bijection (by construction) between the events (a_{ij}) and $(\widehat{c_i}, e_j)$:

$$I(A;\widehat{Z}|\widehat{C}) = I\left((\widehat{C}, E); \widehat{Z}|\widehat{C} \right) = I(E;\widehat{Z}|\widehat{C}).$$

Curious readers might wonder what would yield a reciprocal approach to computing $I(A;\widehat{Z}|\widehat{C})$, which would be to look at how C influences the mapping A, holding Z in

the background. This actually amounts to computing the entropy of the cause:

$$I(A;\widehat{C}|\widehat{Z}) = I\left((E,\widehat{C});\widehat{C}|\widehat{Z}\right) = H(\widehat{C}|\widehat{Z}) = H(\widehat{C}).$$

The last equality obtains by hypothesis of independence between $\widehat{C}$ and $\widehat{Z}$. This reduction to the entropy of the cause comes from the fact that, by construction $\widehat{c}_i$ is necessary to obtain a_{ij} (recall that by definition $a_{ij} \equiv (\widehat{c}_i, e_j)$), while there is no such condition with respect to Z.

Acknowledgments. We thank Maël Montévil for reading a previous version of the manuscript. This publication was made possible through the support of a grant from the Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Templeton World Charity Foundation.

Bibliography

Ay, N., & Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems*, *11*(01), 17–41.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). John Wiley & Sons.

Craver, C. F. (2007). *Explaining the brain*. New York and Oxford: Oxford 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.

Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, part II: Plumbing explanatory depth. *Nos*, *37*(2), 181–199.

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., & Schölkopf, B. (2013). Quantifying causal influences. *The Annals of Statistics*, *41*, 2324–2358.

Kendler, K. S. (2005). 'A gene for...': The nature of gene action in psychiatric disorders. *American Journal of Psychiatry*, *162*(7), 1243–1252.

Korb, K., Hope, L., & Nyberg, E. (2009). Information-theoretic causal power. In F. Emmert-Streib, & M. Dehmer (Eds.), *Information theory and statistical learning* (pp. 231–265). Boston, MA: Springer US.

Lizier, J. T., & Prokopenko, M. (2010). Differentiating information transfer and causal effect. *The European Physical Journal B*, *73*(4), 605–615. doi:10.1140/epjb/e2010-00034-5

Mitchell, S. D. (2002). Ceteris paribus – an inadequate representation for biological contingency. *Erkenntnis*, *57*(3), 329–350.

Pearl, J. (2009). *Causality; models, reasoning and inference*. New York: Cambridge University Press.

Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences*, *96*(6), 3257–3262.

Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, *14*(5), 802–813. doi:10.1261/rna.876308

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*, 595–609.

Weber, M. (2013). Causal selection versus causal parity in biology: Relevant counterfactuals and biologically normal interventions. In *What if? On the meaning, relevance and epistemology of counterfactual claims and thought experiments* (pp. 1–44). Konstanz: University of Konstanz.

Williams, P. L., & Beer, R. D. (2010). *Nonnegative decomposition of multivariate information*. arXiv Preprint arXiv:1004.2515. Retrieved from http://arxiv.org/abs/1004.2515

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British Journal for the Philosophy of Science*, *51*(2), 197–254.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*(3), 287–318.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, *101*(2), 245–280.

Author biography. Arnaud Pocheville is a theoretical biologist and a philosopher of biology. He received his PhD thesis in theoretical biology from the Ecole Normale Supérieure Paris , and pursued his research as a postdoctoral fellow at the Center for Philosophy of Science, University of Pittsburgh. Arnaud is now a research fellow in the Theory and Methods in Biosciences group, University of Sydney, funded by the Templeton World Charity Foundation project "Causal foundations of biological information."

Paul E. Griffiths is a philosopher of science, in the Department of Philosophy, University of Sydney. He is a Fellow of the American Association for the Advancement of Science, of the Australian Academy of the Humanities and former President of the International Society for History, Philosophy and Social Studies of Biology. He leads the Theory and Methods in Bioscience group and the Templeton World Charity Foundation project "Causal foundations of biological information."

Karola Stotz is senior lecturer and a Templeton World Charity Foundation Fellow at the Department of Philosophy, Macquarie University. She has worked at the Konrad Lorenz Institute for Evolution and Cognition Research in Austria, the University of Pittsburgh, Indiana University, and the University of Sydney. She is the co-author of Griffiths, P.E. and K. Stotz (2013) Genetics and Philosophy: An Introduction. CUP.