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

Arnaud Pocheville[a], Paul E. Griffiths[a], Karola Stotz[b]

[a]*Department of Philosophy and Charles Perkins Centre, The University of Sydney, NSW 2006, Australia*
[b]*Department of Philosophy, Macquarie University, NSW 2109, Australia*

Abstract

The interventionist account of causation offers a criterion to distinguish causes from non-causes. It also aims at defining various desirable properties of causal relationships, such as specificity, proportionality and stability. Here we apply an information-theoretic approach to these properties. We show that the interventionist criterion of causation is formally equivalent to non-zero specificity, and that there are natural, information-theoretic ways to explicate the distinction between potential and actual causal influence. We explicate the idea that the description of causes should be proportional to that of their effects. Then we draw a distinction between two ideas in the existing literature, the range of invariance of a causal relationship and its stability. The range of invariance is related to specificity and range of causal values. Stability concerns the effect of additional variables on the relationship between some focal pair of cause and effect variables. We show how to distinguish and measure the direct influence of background variables on the effect variable, and their influence on the relationship between the focal cause and the effect variable. Finally, we discuss the limitations of the information-theoretic approach, and offer prospects for complementary approaches.

Keywords: Causality, Information theory, Interventionism

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 a directed acyclic 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

Pocheville, A., Griffiths, P. E., Stotz, K., 2017. **Comparing causes: an information-theoretic approach to specificity, proportionality and stability**, in: Proceedings of the 15th Congress of Logic, Methodology and Philosophy of Science. College Publications, London.

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 and 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: all other things being equal, 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 and 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 and offer prospects for complementary approaches.

2. Specificity and invariance

In earlier work we and our collaborators proposed a measure of specificity formalising 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:[1]

> Spec: the specificity of a causal variable is obtained by measuring how much mutual information interventions on that variable carry about the effect variable. (Griffiths et al. 2015)

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

$$\mathrm{I}(\mathsf{X};\mathsf{Y}) = \mathsf{H}(\mathsf{X}) - \mathsf{H}(\mathsf{X}|\mathsf{Y})$$

Mutual information is not in itself a suitable measure of causal influence. It is symmetrical, that is $\mathrm{I}(\mathsf{X};\mathsf{Y}) = \mathrm{I}(\mathsf{Y};\mathsf{X})$, and variables can share mutual information without being related in the manner required by the interventionist criterion of causation. However, our measure of specificity does not simply measure the mutual information between variables C and E, but the mutual information between *interventions* on the variable C and the variable E. Interventions can be written in equations by using the $do(\ )$ operator: $do(\mathsf{C})$ means that the value of C results from an intervention on C (Pearl 2009). To

[1] This measure has been independently proposed in cognitive sciences by Tononi, Sporns, and Edelman (1999) and in computational sciences by Korb, Hope, and Nyberg (2009). For related measures see also Ay and Polani (2008) and Janzing et al. (2013). Ay and Polani's measure captures what we call SAD below. See Pocheville (n.d.), for a review.

simplify writing, we will use a hat on the variable: $do(\mathsf{X}) \equiv \widehat{\mathsf{X}}$.[2] Specificity is thus measured by $\mathrm{I}(\widehat{\mathsf{C}}; \mathsf{E})$. 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, $\mathrm{I}(\widehat{\mathsf{C}}; \mathsf{E}) \neq \mathrm{I}(\widehat{\mathsf{E}}; \mathsf{C})$.[3] Furthermore, any variables that satisfy the interventionist criterion of causation in some context will show some mutual information between *interventions* and effects in this context. If $\mathsf{C} \rightarrow \mathsf{E}$ is causal, that is, invariant under at least one intervention on C, then $\mathrm{I}(\widehat{\mathsf{C}}; \mathsf{E}) > 0$. Causation is equivalent to non-zero specificity (see also Pocheville n.d.).

This raises the further question of how the 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 (Woodward 2010, 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. 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 sufficiently map onto distinct values of the effect (Fig. 1 and 2). Increasing the entropy of the cause variable will not increase mutual information when no additional entropy in the effect variable is captured. The range of invariance corresponds to the 'effective' entropy of the cause, that is, where all the values which make the same difference to the effect have been aggregated (Section 4).

3. Actual and potential difference-making

Our measure of specificity depends on what probability distribution we choose to impose on the causal variable C (as well as on the mapping from C to E). This is something that earlier, qualitative discussions seemed to be able to do without. In our earlier paper we showed that this is a feature and not a bug of our measure. As we will now discuss, measuring specificity with different probability distributions over $\widehat{\mathsf{C}}$ corresponds to different views of causal specificity in the existing, qualitative literature.

One way to measure specificity corresponds to Woodward's characterisation of fine-grained influence (INF) (Woodward 2010). In his presentation the

[2] We take this convention from related work in computer sciences applying information theory to causal modeling (e.g. Ay and Polani 2008; Lizier and Prokopenko 2010).

[3] These quantities can be equal if and only if both are null, i.e. iff 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, thus acyclic, graph: if C causes E, E cannot feed back on C.

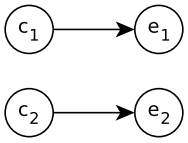

Figure 1: Causal mapping showing a bijection between causal values and effect values. Complete ignorance (maximum entropy) obtains when each value of the effect has a probability of $\frac{1}{2}$ before intervening on the value of the cause: $\mathsf{H}(\mathsf{E}) = -\sum_{j=1}^{2} p(e_j) \log_2 p(e_j) = -\sum_{j=1}^{2} \frac{1}{2} \log_2(\frac{1}{2}) = 1$ bit. After knowing the value set for the cause (c_1 or c_2), the effect is fully specified and the conditional entropy is: $\mathsf{H}(\mathsf{E}|\widehat{\mathsf{C}}) = -\sum_{i=1}^{2} p(\widehat{c}_i) \sum_{j=1}^{2} p(e_j|\widehat{c}_i) \log_2 p(e_j|\widehat{c}_i) = -\sum_{i=1}^{2} \frac{1}{2} \sum_{j=1}^{2} 1 \log_2(1) = 0$ 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 $\widehat{\mathsf{C}}$: $\mathrm{I}(\mathsf{E}; \widehat{\mathsf{C}}) = \mathsf{H}(\mathsf{E}) - \mathsf{H}(\mathsf{E}|\widehat{\mathsf{C}}) = 1$ bit.

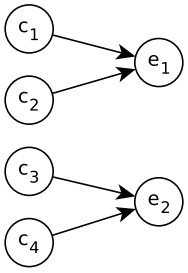

Figure 2: Here, different values of the cause lead to the same outcome. As in Figure 1, $\mathsf{H}(\mathsf{E}) = 1$ bit. Although here two values of the cause can lead to the same effect, intervening to set the value of the cause fully specifies the value of the effect just as effectively as it does in Figure 1. Therefore, the difference in uncertainty about the effect between before and after intervening to set the value of the cause is the same: $\mathrm{I}(\mathsf{E}; \widehat{\mathsf{C}}) = \mathsf{H}(\mathsf{E}) - \mathsf{H}(\mathsf{E}|\widehat{\mathsf{C}}) = 1 - 0 = 1$ bit.

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.

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 Pocheville n.d.):

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, Hope, and Nyberg 2009), and can also be thought of as a measure of C's potential influence on E. We are inclined to think MaxSpec is 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.

A different view of causal specificity has been advocated by 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.

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.[4] 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 have greater actual causal influence on DIE than SHOOT does in a population where more prisoners die from fright than from bullets. It may also be the case 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. The 'experiment of nature' does not necessarily include the experiment that reveals how E depends on C, leading to no actual influence.[5]

[4]The same idea has been termed 'information flow' (Ay and Polani 2008; see Pocheville n.d.)

[5]In addition, Weber (2013) has argued that in the biological sciences specificity should

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. 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

---

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 to construct a version of relevant specificity based on the MaxSpec measure.

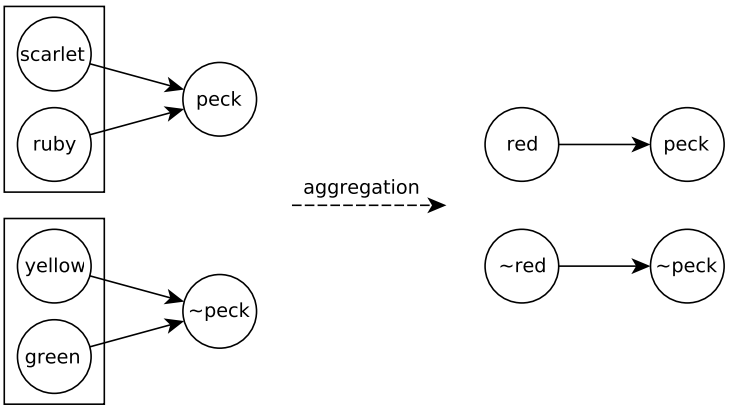

Figure 3: The causal mapping in Yablo's example. The re-description of the causal variable in terms of red/~red reduces the entropy from 2 bits to 1 bit (assuming equiprobability of colors).

way which minimizes its entropy and still maximizes specificity.[6]

We can see how this works with Yablo's original example (Yablo 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 however, this one is underspecified. We have two variables: P, with the values 'peck' and '~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. 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 probability on the red values equal to that on the non-red values. The graph we have described resembles that in Figure 2 above, and is exactly that graph if there are just two red and two non-red values (Fig. 3). If we coarse-grain the values of our variable, so that S now has just two values, red and ~red, then we get the graph in Figure 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 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.

[6]This means aggregating values $\{c_i, c_j\}$ for which $p(e_k|\widehat{c_i}, \widehat{z_l}) = p(e_k|\widehat{c_j}, \widehat{z_l})$ for any e_k and z_l of interest, where C is the cause, E the effect and Z a set of background variables. We suppose here that the effect E is already described in the fashion of interest. Its values could be similarly aggregated.

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 (57) [(Alfred 2000)]. 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 → pitch perception → 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, pp. 1249-50, his emphasis)

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 phenomenon that we identified in the pigeon-pecking case, nor is it really a matter of fine- versus coarse-graining values. The alternative to saying that the genetic variant is a gene for language or a gene for liking Mozart is to say that it is a gene for a variable which explains a host of cognitive effects. This corresponds to redrawing the causal graph by inserting a variable, not to redescribing the effect variable. Similarly, adding connections between the cause and additional effect variables does not amount to fine-graining the values of the original effect variable: liking Mozart is not an exclusive alternative to liking Haydn (Fig. 4).

In these two later examples it is the causal graph itself that is too 'coarse grained' rather than one of the variables. Choosing which variables to include in the graph and choosing how finely or coarsely to discretize a variable are different problems.[7] We therefore prefer to define 'proportionality' more narrowly:

> 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

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

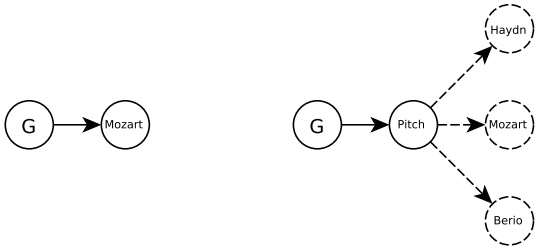

Figure 4: Causal graphs for Kendler's example, where a gene is supposed to influence music preferences.

> 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. Craver (2007, chap. 6), for example, has argued that in some cases the lower-level explanation merely recognises additional differences that make no difference. Whether this argument is successful or not, our version of the proportionality constraint seems suitable to capture the intention behind it.

5. Stability

The philosophical landscape

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 figure 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. 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 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 and Woodward 2003, p. 187). In our terms, causally relevant background conditions are additional variables that, under some conditions, can 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. 346-7)

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:

> 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 and 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 and Woodward 2003, p. 188, italics in original)

One further distinction is needed to think clearly about the relationship between causal generalisations and background conditions. When we speak of the 'stability' of some relationship $\mathsf{C} \rightarrow \mathsf{E}$ we may have in mind, not the

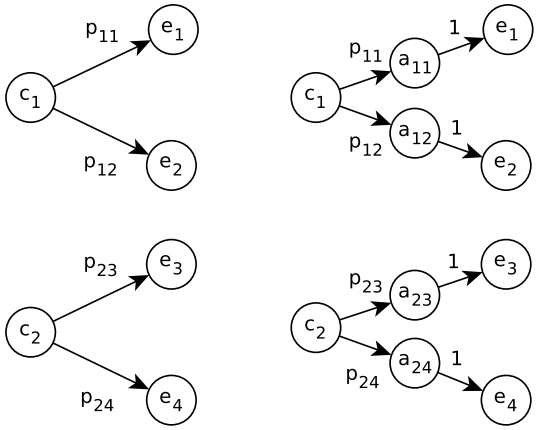

Figure 5: On the left, a causal mapping relates values of a nominal causal variable to values of a nominal effect variable. In this example, each causal value can lead to two proper (and incompatible) effect values, each arrow being associated with a probability $\mathsf{p}_{ij} = \mathsf{p}(\mathsf{e}_\mathsf{j}|\widehat{\mathsf{c}}_\mathsf{i})$. On the right, 'arrows' are now explicitly represented as values of a new variable, A, which represents the mapping between C and E. By definition each arrows leads to a given effect value with probability 1. (Note that it is possible to draw three-variable mappings in two dimensions only under certain conditions.)

influence of background variables on E, but whether the relationship $\mathsf{C} \rightarrow \mathsf{E}$ itself changes across a range of background conditions. For example, alternative 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 and 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.

An information-theoretic treatment

It is stability and instability in this last 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, in an information-theoretic framework, the idea of stability of causal relationships to our measure of causal specificity.

To start with, let us consider a causal relationship $\mathsf{C} \rightarrow \mathsf{E}$ represented by a mapping between values of a (nominal) causal variable C to values of a (nominal) effect variable E (Fig. 5). 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.

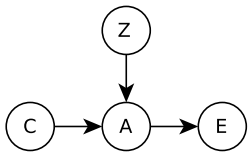

Figure 6: Causal graph with a variable representing the arrows A mapping C to E as they are affected by Z. (Note that this diagram is a causal graph relating variables, not a mapping relating values.)

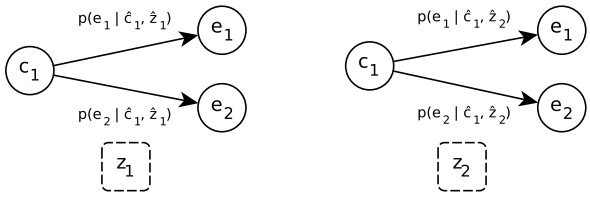

Figure 7: Diagram showing how interventions on Z can modify the mapping from C to E. For simplicity, only a single value of the causal variable is considered.

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 – Fig. 6) has taken the value a_{ij}.[8] 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 us 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 to be explained, and Z as the variable explaining them. How much Z explains the arrows can be measured by $I(A; \widehat{Z})$, as we explain now.

We first consider the arrows stemming from one causal value. Let us 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.[9] 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, $I(A; \widehat{Z}|\widehat{c}_1)$.[10]

Figure 7 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 mapping between C and E is then maximally stable with respect

[8] 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 variable A is also a nominal variable.

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

[10] 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 manipulating Z can be a cause of manipulating C. We let this case aside here (see Appendix).

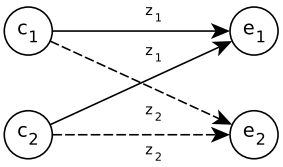

Figure 8: Causal mapping where Z is the only cause of E.

to $\widehat{\mathsf{Z}}$ (in this limiting case, Z is an irrelevant background condition). 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 $\widehat{z}_2$.[11] 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 $I(A; \widehat{Z}|\widehat{C})$. When Z and C are independently manipulated, which will be our hypothesis from now on, $I(A; \widehat{Z}|\widehat{C}) = I(A; \widehat{Z})$ (see Appendix). This quantity characterizes how much Z affects the mapping between C and E or, in other words, the instability of the mapping with respect to Z.

Stability, i.e. how much of the mapping does not depend on the background, is represented by the difference between the entropy of the mapping, and how much of it is explained by the background, that is $H(A) - I(A; \widehat{Z})$. Of course, this quantity does not necessarily correspond to variation in the causal arrows that increases the specificity of C: a cause can have almost null specificity and the causal relationship be maximally stable with respect to a given background condition.[12] Specificity answers a different question: it looks at how much *stable influence* the cause can exert, irrespective of the background (when uncontrolled) (Table 1).

Mappings between C and E do not all 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. 8). The method we just outlined may 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 is not this direct influence 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. 9). This, in our view, is what it means to talk of the causal relationship $\mathsf{C} \rightarrow \mathsf{E}$ depending on Z.

[11] Making an arbitrary choice on the indices, we can write the condition as $p(a_{11}|\widehat{z}_1) = p(a_{12}|\widehat{z}_2) = 1$.

[12] If the description of values is gerrymandered (Section 4), the entropy and stability of the mapping can be artificially inflated. This measure has been put to work in biology in a forthcoming paper by Calcott (2017).

Concepts	**Relationship**	**Measure**
Specificity	$\mathsf{C} \rightarrow \mathsf{E}$	$\mathrm{I}(\widehat{\mathsf{C}}; \mathsf{E})$, $\mathrm{I}(\widehat{\mathsf{C}}; \mathsf{E}\vert\widehat{\mathsf{Z}})$
Stability	$\mathsf{Z} \rightarrow (\mathsf{C} \rightarrow \mathsf{E})$	$\mathrm{H}(\mathsf{A}) - \mathrm{I}(\mathsf{A}; \widehat{\mathsf{Z}})$
Interaction	$\mathsf{Z} \rightarrow (\mathsf{C} \rightarrow \mathsf{E})$	$\mathrm{I}(\widehat{\mathsf{Z}}; \mathsf{E}\vert\widehat{\mathsf{C}}) - \mathrm{I}(\widehat{\mathsf{Z}}; \mathsf{E})$

Table 1: Information-theoretic measures for specificity, stability, and interaction.

To measure our real target, a solution would be to follow the proportionality constraint and aggregate the values of C *and* E accordingly.[13] Here we propose to look at another interesting quantity, called the interaction information. Let us first remark that the (conditional) specificity of Z for the mapping is equal to the conditional specificity of Z for the effect. That is, $\mathrm{I}(\mathsf{A}; \widehat{\mathsf{Z}}|\widehat{\mathsf{C}}) = \mathrm{I}(\mathsf{E}; \widehat{\mathsf{Z}}|\widehat{\mathsf{C}})$ (see Appendix). This term embeds both the information coming from $\widehat{\mathsf{Z}}$ alone, which is here equal to $\mathrm{I}(\mathsf{E}; \widehat{\mathsf{Z}})$, and the information coming from the interaction between $\widehat{\mathsf{Z}}$ and $\widehat{\mathsf{C}}$, which is what we are after (Fig. 9).[14] To measure this interaction we compute the quantity $\mathrm{I}(\mathsf{E}; \widehat{\mathsf{Z}}; \widehat{\mathsf{C}}) = \mathrm{I}(\mathsf{E}; \widehat{\mathsf{Z}}|\widehat{\mathsf{C}}) - \mathrm{I}(\mathsf{E}; \widehat{\mathsf{Z}})$.

This is the interaction information between the three variables. It 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. The quantity is zero if C and Z have entirely non-interactive effects on E. There is no interaction 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. That is, 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 5 but not in Figure 9.[15]

We remarked above that the instability of the causal relationship $\mathsf{C} \rightarrow \mathsf{E}$ with respect to a background variable Z must be distinguished from the direct effect of Z on E. We can now make this point more precise. The first would be measured by $\mathrm{I}(\mathsf{E}; \widehat{\mathsf{Z}})$ and the second by $\mathrm{I}(\widehat{\mathsf{C}}; \mathsf{E}; \widehat{\mathsf{Z}})$. 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 (Fig. 9).

[13] This would be done without controlling for Z, to ignore its direct effects on E.

[14] 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 (Bertschinger et al. 2013a; Bertschinger et al. 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.

[15] The interaction information is symmetrical: $\mathrm{I}(\mathsf{E}; \widehat{\mathsf{Z}}; \widehat{\mathsf{C}}) = \mathrm{I}(\mathsf{E}; \widehat{\mathsf{C}}|\widehat{\mathsf{Z}}) - \mathrm{I}(\mathsf{E}; \widehat{\mathsf{C}}) = \mathrm{I}(\widehat{\mathsf{Z}}; \widehat{\mathsf{C}}|\mathsf{E}) - \mathrm{I}(\widehat{\mathsf{Z}}; \widehat{\mathsf{C}})$. 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 $\{\mathsf{C}, \mathsf{Z}\}$ to E.

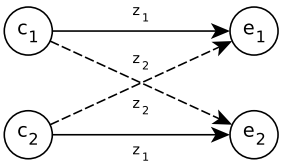

Figure 9: Example of interacting causes C and Z with respect to E. If the background Z is not controlled, the cause C is entirely not specific (assuming, for the ease of presentation, equiprobability between $\widehat{z}_1$ and $\widehat{z}_2$). Indeed, any intervention $\widehat{c}_1$ or $\widehat{c}_2$ can equiprobably lead to e_1 or e_2. Thus, $\mathrm{I}(\widehat{\mathsf{C}};\mathsf{E}) = 0$ bit. However, once we know the background, C is entirely specific: $\mathrm{I}(\widehat{\mathsf{C}};\mathsf{E}|\widehat{\mathsf{Z}}) = 1$ bit (assuming, for the ease of presentation, equiprobability between $\widehat{c}_1$ or $\widehat{c}_2$). The interaction information in this case is $\mathrm{I}(\widehat{\mathsf{C}};\mathsf{E};\widehat{\mathsf{Z}}) = \mathrm{I}(\widehat{\mathsf{C}};\mathsf{E}|\widehat{\mathsf{Z}}) - \mathrm{I}(\widehat{\mathsf{C}};\mathsf{E}) = 1$ bit. (By design, the same holds when Z is the focal cause variable and C is as a background variable.)

6. Conclusion

Our information-theoretic framework was originally developed for thinking about causal specificity within the interventionist approach to causation. In this paper we have used it to analyse several other key elements of the interventionist account. In Section 2 we showed that the criterion of causation in Woodward's interventionist account is equivalent to a non-zero specificity in the relationship between a cause and its effect. We suggested that the range of invariance of a causal relationship can be measured by the effective entropy of the cause for its effect. In Section 3 we argued that different qualitative discussions of specificity correspond to different probability distributions over the causal variable, leading to measure respectively fine-grained influence, potential causal control, and actual difference-making. In Section 4 we proposed to make more precise the controversial idea that the description of the cause should be 'proportional' to its effects. Ideal proportionality is achieved by simultaneously minimising the entropy of the cause whilst maximising its specificity. This amounts to discretising the cause variable so as to mark all and only differences that make a difference to the effect variable. We suggested that some questions in the literature about proportionality concern which variables to include in a causal graph, rather than the grain of description of a given variable. In Section 5 we suggested that the 'stability' of a causal relationship is the extent to which that relationship is not affected by additional, background variables. We offered an information-theoretic analysis of the stability of causal relationships, where background variables affect the mapping between the focal cause and the focal effect.

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. It goes without saying that the strength and simplicity of the information theoretic formalism come with 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. 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.

Acknowledgements

We thank Maël Montévil for reading a previous version of the manuscript. The current version has been presented in a workshop to Jun Otsuka, Pierrick Bourrat, Brett Calcott, and Wei Fang, whom we warmly thank for their feedback. Thanks to Stefan Gawronski for designing an early version of Figure 3. 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.

Appendix

Here we provide a quick primer in information theory, proofs of equations cited in the text and expand on some ideas of Section 5.

6.1. Entropy, conditional entropy, and mutual information

We recall basic formulas of information theory. For a primer on information theory, see Cover and 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:

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

Our measure of specificity of C to E is defined as the mutual information between $\widehat{C}$ and E: $Spec(C \to E) \equiv I(\widehat{C}; E)$. In the conditional form, it reads:

$$I(\widehat{C}; E|\widehat{Z}) = \sum_k p(\widehat{z}_k) \sum_i \sum_j p\left(\widehat{c}_i, e_j \middle| \widehat{z}_k\right) \log_2 \left(\frac{p(\widehat{c}_i, e_j|\widehat{z}_k)}{p(\widehat{c}_i|\widehat{z}_k)p(e_j|\widehat{z}_k)} \right).$$

6.2. Proofs

• We want to prove: $I(A; \widehat{Z}|\widehat{C}) = I(A; \widehat{Z})$ when $\widehat{C}$ and $\widehat{Z}$ are independent. We start with the plain formula for $I(A; \widehat{Z}|\widehat{C})$:

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

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:

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

• We want to prove: that 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 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.

References

Alfred, Jane (2000). "Tuning in to perfect pitch". In: *Nature Reviews Genetics* 1.3.

Ay, Nihat and Daniel Polani (2008). "Information flows in causal networks". In: *Advances in complex systems* 11.1, pp. 17–41.

Bertschinger, Nils, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost (2013a). "Shared information—New insights and problems in decomposing information in complex systems". In: *Proceedings of the European Conference on Complex Systems 2012*. Springer, pp. 251–269.

Bertschinger, Nils, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay (2013b). "Quantifying unique information". In: *arXiv preprint arXiv:1311.2852*.

Calcott, Brett (2017). "Causal Specificity and the Instructive-Permissive Distinction". In: *Biology and Philosophy*.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. John Wiley & Sons.

Craver, Carl F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Clarendon Press. 329 pp.

Griffiths, Paul E., Arnaud Pocheville, Brett Calcott, Karola Stotz, Hyunju Kim, and Rob Knight (2015). "Measuring Causal Specificity". In: *Philosophy of Science* 82.4, pp. 529–555.

Hitchcock, Christopher and James Woodward (2003). "Explanatory generalizations, part II: Plumbing explanatory depth". In: *Noûs* 37.2, pp. 181–199.

Janzing, Dominik, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf (2013). "Quantifying causal influences". In: *The Annals of Statistics* 41.5, pp. 2324–2358.

Kendler, Kenneth S. (2005). ""A gene for...": the nature of gene action in psychiatric disorders". In: *American Journal of Psychiatry* 162.7, pp. 1243–1252.

Korb, Kevin B, Lucas R Hope, and Erik P Nyberg (2009). "Information-Theoretic Causal Power". In: *Information Theory and Statistical Learning*. Ed. by Frank Emmert-Streib and Matthias Dehmer. Boston, MA: Springer US, pp. 231–265.

Lizier, Joseph T. and M. Prokopenko (2010). "Differentiating information transfer and causal effect". In: *The European Physical Journal B* 73.4, pp. 605–615.

Mitchell, Sandra D. (2002). "Ceteris paribus—an inadequate representation for biological contingency". In: *Ceterus Paribus Laws*. Springer, pp. 53–74.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. 2nd Edition. New York: Cambridge University Press. 486 pp.

Pocheville, Arnaud. "Causal specificity, information flow, and causal independence". In:

Rauh, Johannes, Nils Bertschinger, Eckehard Olbrich, and Jürgen Jost (2014). "Reconsidering unique information: Towards a multivariate information decomposition". In: *2014 IEEE International Symposium on Information Theory*. IEEE, pp. 2232–2236.

Tononi, Giulio, Olaf Sporns, and Gerald M. Edelman (1999). "Measures of degeneracy and redundancy in biological networks". In: *Proceedings of the National Academy of Sciences* 96.6, pp. 3257–3262.

Wang, Z. and C. B. Burge (2008). "Splicing regulation: From a parts list of regulatory elements to an integrated splicing code". In: *RNA* 14.5, pp. 802–813.

Waters, C. Kenneth (2007). "Causes that make a difference". In: *The Journal of Philosophy*, pp. 551–579.

Weber, Marcel (2006). "The central dogma as a thesis of causal specificity". In: *History and philosophy of the life sciences*, pp. 595–609.

— (2013). "Causal Selection Versus Causal Parity in Biology: Relevant Counterfactuals and Biologically Normal Interventions".

Williams, Paul L. and Randall D. Beer (2010). "Nonnegative decomposition of multivariate information". In: *arXiv preprint arXiv:1004.2515*.

Woodward, James (2000). "Explanation and invariance in the special sciences". In: *The British Journal for the Philosophy of Science* 51.2, pp. 197–254.

— (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

— (2010). "Causation in biology: stability, specificity, and the choice of levels of explanation". In: *Biology & Philosophy* 25.3, pp. 287–318.

Yablo, Stephen (1992). "Mental Causation". In: *The Philosophical Review* 101.2, pp. 245–280.