Measuring Causal Specificity

Paul E. Griffiths, Arnaud Pocheville, Brett Calcott,
Karola Stotz, Hyunju Kim, and Rob Knight*†

Several authors have argued that causes differ in the degree to which they are 'specific' to their effects. Woodward has used this idea to enrich his influential interventionist theory of causal explanation. Here we propose a way to measure causal specificity using tools from information theory. We show that the specificity of a causal variable is not well defined without a probability distribution over the states of that variable. We demonstrate the tractability and interest of our proposed measure by measuring the specificity of coding DNA and other factors in a simple model of the production of mRNA.

1. Causal Specificity. Several authors have argued that causes differ in the degree to which they are 'specific' to their effects. The existing literature on causal specificity is mostly qualitative and recognizes that the idea is not yet adequately precise (e.g., Weber 2006, 2013; Waters 2007; Woodward 2010). Marcel Weber has suggested that the next step should be a quantitative measure of specificity (2006, 606). In this article we examine how to measure specificity using tools from information theory.

Causal specificity is often introduced by contrasting the tuning dial and the on/off switch of a radio. Hearing the news is equally dependent on the dial (or digital tuner) taking the value '576' and on the switch taking the

Received October 2014; revised April 2015.

*To contact the authors, please write to: Paul E. Griffiths, Department of Philosophy, Quadrangle A14, University of Sydney NSW 2006, Australia; e-mail: paul.griffiths@usyd.edu.au.

†This publication was made possible through the support of a grant from the Templeton World Charity Foundation. The opinions expressed are those of the author(s) and do not necessarily reflect the views of the Templeton World Charity Foundation. Brett Calcott was supported by Joshua Epstein's NIH Director's Pioneer Award DP1OD003874 from the Office of the Director, National Institutes of Health. The article is the result of a workshop held at the University of Colorado, Boulder, with support from Templeton World Charity Foundation. BC, PG, AP, and, KS wrote the manuscript, and all authors agreed on the final content. We would like to thank two anonymous referees for their helpful comments.

Philosophy of Science, 82 (October 2015) pp. 529–555. 0031-8248/2015/8204-0001$10.00

value 'on'. But the dial seems to have a different kind of causal relationship with the news broadcast than the switch. The switch is a nonspecific cause, whereas the dial (or digital tuner) is a specific cause. The difference has something to do with the range of alternative effects that can be produced by manipulating the tuner, as opposed to manipulating the switch.

Another widely discussed example of specific and nonspecific causes contrasts a coding sequence of DNA with other factors involved in DNA transcription and translation (e.g., Waters 2007). But this example has to be carefully tailored to produce the desired intuition about specificity (Griffiths and Stotz 2013). In section 5 we show that the causal specificity of coding sequences of DNA differs dramatically in different cases.

Like most of the recent literature, our account of causal specificity makes use of Woodward's interventionist theory of causal explanation (Woodward 2003). We give only the briefest summary of Woodward's theory here, since it should be well known to the presumptive audience for this article, and Woodward has provided a succinct and readily accessible summary online (Woodward 2012). Woodward construes causation as a relationship between variables in a scientific representation of a system. There is a causal relationship between variables X and Y if it is possible to manipulate the value of Y by intervening to change the value of X. 'Intervention' here is a technical notion with various restrictions. For example, changing a third variable Z that simultaneously changes X and Y does not count as 'intervening' on X. Causal relationships between variables differ in how 'invariant' they are. Invariance is a measure of the range of values of X and Y across which the relationship between X and Y holds. But even relationships with very small ranges of invariance are causal relationships.

Both Waters (2007) and Woodward (2010) have suggested that causal specificity is related to 'causal influence' (Lewis 2000; and see sec. 2). A causal variable has 'influence' on an effect variable if a range of values of the cause produces a range of values of the effect, as in the example of the tuner. However, while Lewis proposed that 'influence' distinguishes causes from noncauses, for Woodward it merely marks out causes that are particularly apt for intervention.

Although Woodward (2010) gives the most complete account of specificity to date, there remains much to be done, as he recognizes. Marcel Weber has suggested that causal specificity is merely a variety of Woodward's invariance. A variable is a more specific cause of some other variable, Weber suggests, to the extent that the causal relationship between cause and effect variables is invariant across the range of values of both variables and to the extent that the two variables have large ranges of values (Weber 2006, 606). Woodward disagrees, arguing that a causal relationship with these properties may fail to meet some of the other conditions we discuss below, such as being a bijective function from cause variable to effect variable

(Woodward 2010, n. 17).[1] An attempt to quantify specificity is one obvious way to move the discussion forward. As we see below, the points that Weber and Woodward are making become much clearer when expressed using a quantitative measure.

A skeptical reader may wonder why the apparently elusive notion of causal specificity deserves such effort. Our motivation is the same as that of Waters and Weber: clarifying the notion of *causal* specificity may elucidate the notion of *biological* specificity and facilitate the study of specificity in actual biological systems. The term 'specificity' entered biology in the 1890s in response to the extraordinary precision of biochemical reactions, such as the ability to produce an immune response to a single infective agent or the ability of an enzyme to interact with just one substrate. By the 1940s biological specificity had come to be identified with the precision of stereochemical relationships between biomolecules. In 1958, however, Francis Crick's theoretical breakthrough in understanding protein synthesis introduced a complementary conception of specificity, sometimes referred to as 'informational specificity'. Stereochemical specificity results from the unique, complex three-dimensional structure of a molecule that allows some molecules but not others to bind to it and interact. In contrast, informational specificity is produced by exploiting combinatorial complexity within a linear sequence, which can be done with a relatively simple and homogenous molecule such as DNA (see Griffiths and Stotz 2013, chap. 3).

The notion of causal specificity in philosophy of science was not introduced with any a priori assumption that it is the same thing as biological specificity. However, Waters has used the idea of causal specificity to argue that DNA encodes biological specificity for gene products, unlike other factors involved in making those products (Waters 2007). In contrast, Stotz and Griffiths have used causal specificity to argue that the biological specificity for a gene product is distributed across several of these factors (Stotz 2006; Griffiths and Stotz 2013).

A merely intuitive approach to causal specificity is unlikely to be helpful in settling disputes like this. In section 5 we show that a quantitative approach may allow a more definitive resolution. At the very least, it makes clear which assumptions are driving the different conclusions reached by the protagonists.

2. Specificity and Information. Causal specificity has been characterized by Woodward as a property of the mapping between causes and effects:

1. A function mapping causes to effects will be injective if no effect has more than one cause, surjective if every effect has at least one cause, and bijective if it is both injective and surjective—every effect has one and only one cause and vice versa.

> My proposal is that, other things being equal, we are inclined to think of C as having more rather than less influence on E (and as a more rather than less specific cause of E) to the extent that it is true that:
>
> (INF) There are a number of different possible states of C ($c_1 \ldots c_n$), a number of different possible states of E ($e_1 \ldots e_m$) and a mapping F from C to E such that for many states of C each such state has a unique image under F in E (that is, F is a function or close to it, so that the same state of C is not associated with different states of E, either on the same or different occasions), not too many different states of C are mapped onto the same state of E and most states of E are the image under F of some state of C. (Woodward 2010, 305)

We propose to quantify Woodward's proposal that a cause becomes more specific as the mapping of cause to effect resembles a bijection.

We start from the simple idea that the more specific the relationship between a cause variable and an effect variable, the more information we have about the effect after we perform an intervention on the cause. Starting from this idea, we can apply the tools of information theory to measure some properties of causal mappings that relate values of the cause to values of the effect. For simplicity, we restrict ourselves to variables that take nominal values, with no obvious metric relating the diverse values.[2] One property we can measure in this way is Woodward's INF. Rather than describing a relationship as injective or bijective, information theory allows us to express the tendency toward a bijective relationship as a continuous variable. Thus, our informational measure of specificity will preserve the essence of Woodward's proposal while allowing this desirable flexibility.

We use the term 'information' in the classic sense of a reduction of uncertainty (Shannon and Weaver 1949). In information theory, the uncertainty about an event can be measured by the entropy of the probability distribution of events belonging to the same class (see app. A for a brief primer on information theory explaining the measures used in this article). Uncertainty about the outcome of throwing a die is measured by the entropy of the probability distribution of the six possible outcomes. Maximum entropy occurs when all six faces of the die have equal probabilities. If the die is loaded, the entropy is smaller, and there is less uncertainty about the outcome, because one side is more probable than the others.

2. Variants of our approach to causal specificity are possible for metric variables. The analysis of variance, for example, gives measures that are respectively equivalent to entropy, conditional entropy, and mutual information. The information theoretic approach taken here is more general, but the analysis of variance retains more information about the metric (see Garner and McGill [1956] for a comparison). Information theoretic variants have also been developed to deal with continuous variables (e.g., Reshef et al. 2011; Ross 2014).

Applying this framework to a causal relationship allows one to measure how much knowing the value set by an intervention on a causal variable reduces one's uncertainty about the value of an effect variable. We can measure this reduction of uncertainty by comparing the entropy of the probability distribution of the value of the effect before and after knowing the value of the cause set by an intervention. The more the difference in entropies, the more our uncertainty has been reduced. The maximum reduction of uncertainty occurs when we start from complete ignorance (i.e., maximum entropy) and when, after knowing the value of the cause set by an intervention, we end up with a completely specified value for the effect (null entropy—e.g., when a die is so heavily loaded that it always comes up 6).

These ideas can be illustrated with simple diagrams showing how different values of a causal variable (C) map to different values of an effect variable (E). We draw the reader's attention to the fact that these diagrams are causal mappings rather than conventional causal graphs. Nodes represent values of variables, rather than variables, as they would in a causal graph. Likewise, arrows do not represent causal connections between variables, as they would in a causal graph. An arrow connecting a value of a cause to a value of an effect means that interventions which set the cause to that value will lead to the effect having that value, with some probability. For instance, the arrow stemming from c_i and pointing to e_j corresponds to the joint event $(\widehat{c_i}, e_j)$ with probability $p(\widehat{c_i}, e_j)$. The hat in the formula means that the value c_i is fixed by an 'atomic' intervention (see app. B for a brief primer on causal modeling, explaining in particular the concept of an atomistic intervention).

For ease of presentation, we make some simplifying assumptions:

1. We consider only cases in which we start from complete ignorance about the effect (maximum entropy).
2. We assume that all causal values, arrows, and effect values are equiprobable.
3. We consider only cases relating one cause and one effect, ruling out the possibility of confounding factors. However, the same measures could be used in cases with confounding factors, as atomic interventions on the causal value will break the confounding influence of such factors on the association between values of the cause and values of the effect.

The simplest case is a bijection, where each value of the cause corresponds to one value of the effect and vice versa (see fig. 1). Here, complete ignorance (maximum entropy) obtains when each value of the effect has a probability of 1/2 before knowing the value set by the intervention on the cause:

https://doi.org/10.1086/682914 Published online by Cambridge University Press

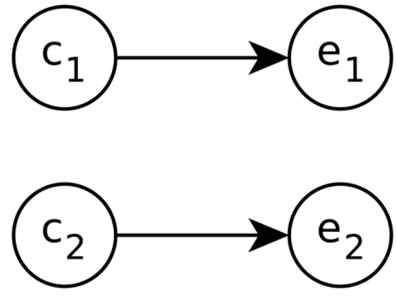

Figure 1. Bijection between causal values and effect values.

$$H(E) = -\sum_{j=1}^{2} p(e_j)\log_2 p(e_j)$$
$$= -\sum_{j=1}^{2} \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1[\text{bit}].$$

After knowing the value of the cause set by the intervention (say, $\widehat{c_1}$), the effect is now fully specified (it is e_1 with probability 1), and the conditional entropy is

$$H(E|\widehat{C}) = -\sum_{i=1}^{2} p(\hat{c}_i)\sum_{j=1}^{2} p(e_j \mid \hat{c}_i)\log_2 p(e_j \mid \hat{c}_i)$$
$$= -\sum_{1}^{2} \frac{1}{2}\{1\log_2(1) + 0\log_2(0)\} = 0[\text{bits}].$$

The information gained by knowing the cause can be obtained by measuring the difference between the entropy before and the entropy after knowing the value set for the cause by the intervention. This quantity is the mutual information between E and $\widehat{C}$:

$$I(E;\widehat{C}) = H(E) - H(E|\widehat{C}) = 1[\text{bit}].$$

These three quantities $H(E)$, $H(E|\widehat{C})$, and $I(E;\widehat{C})$ characterize interesting properties of the causal mapping above. The entropy, $H(E)$ measures how large and even the repertoire of possible effects is. It is the amount of information that can be gained by totally specifying an effect among a set of possible effects (here, this is 1 bit). The conditional entropy $H(E|\widehat{C})$ characterizes the remaining uncertainty about an effect when the value set for the cause is known (here it is fully specified, so the uncertainty is 0 bits). Finally, the mutual information $I(E;\widehat{C})$ measures the extent to which knowing the

https://doi.org/10.1086/682914 Published online by Cambridge University Press

value set for the cause specifies the value of the effect (here, knowing the value of the cause brings 1 bit of information).

Another simple case is when any value of the cause can lead to any value of the effect (see fig. 2). We only present this as a limiting case because manipulating the value of C between c_1 and c_2 would have no effect on the value of E, and so C is not a cause of E on the interventionist account. In this case, as in the previous case,

$$H(E) = -\sum_{1}^{2} \frac{1}{2} \log_2 \left(\frac{1}{2} \right) = 1[\text{bit}].$$

Because in this case knowing the value set by an intervention on C gives no information about the value of E, the conditional entropy $H(E|\widehat{C})$ is equal to $H(E)$ (our uncertainty is unchanged):

$$H(E \mid \widehat{C}) = -\sum_{1}^{2} \frac{1}{2} \sum_{1}^{2} \frac{1}{2} \log_2 \left(\frac{1}{2} \right) = 1[\text{bit}].$$

Thus, the information gained by knowing the value set for C is nil (C is entirely nonspecific):

$$I(E; \widehat{C}) = H(E) - H(E|\widehat{C}) = 0[\text{bits}].$$

Notice that we can approach this null mutual information as a limit of a genuine cause whose different values make decreasingly smaller differences as regards the value of the effect. This implies that specificity and the interventionist criterion of causation are not fully independent. These two cases, bijection (fig. 1) and exhaustive connection (fig. 2) illustrate limit cases of Woodward's 'degree of bijectivity' of causal mappings.

We can go further by examining two slightly more complicated cases. The first is when each value of a cause leads to a proper set of values of the

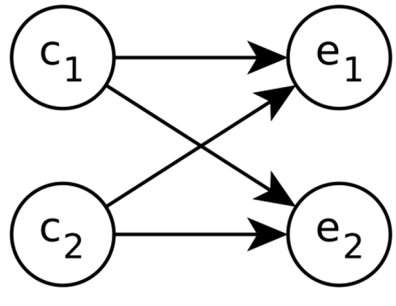

Figure 2. Any value of the cause can lead to any value of the effect.

https://doi.org/10.1086/682914 Published online by Cambridge University Press

effect (see fig. 3). In this case the maximum uncertainty about the effect is larger:

$$H(E) = -\sum_{1}^{4} \frac{1}{4} \log_2 \left(\frac{1}{4} \right) = 2[\text{bits}].$$

Furthermore, knowing the cause less than fully specifies the effect. Assuming equiprobability between the two effect values that can be produced by a single value of the cause, the conditional entropy $H(E|\widehat{C})$ is

$$\begin{aligned} H(E \mid \widehat{C}) &= -\sum_{i=1}^{2} p(\widehat{c}_i) \sum_{j=1}^{4} p(e_j|\widehat{c}_i) \log_2 p(e_j|\widehat{c}_i) \\ &= -\sum_{1}^{2} \frac{1}{2} \{2 \left(\frac{1}{2} \log_2 \left(\frac{1}{2} \right) \right) + 2(0 \log_2(0))\} = 1[\text{bit}]. \end{aligned}$$

Thus, the information about the effect gained by knowing the cause is

$$I(E; \widehat{C}) = H(E) - H(E|\widehat{C}) = 1[\text{bit}].$$

Notice that knowing the value of the cause provides as much information about the effect as in figure 1, but because the repertoire of effects is larger, the remaining uncertainty—$H(E|\widehat{C})$—is not null anymore. The repertoire of effects will be larger if, for instance, we increase the level of detail when

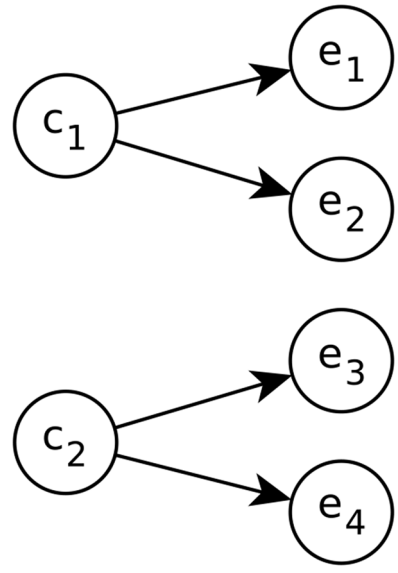

Figure 3. Single value of the cause can lead to more than one value of the effect.

https://doi.org/10.1086/682914 Published online by Cambridge University Press

describing effects (compare a game of dice based on odd vs. even outcomes to a game based on the values of the six individual faces).

Let us now consider the symmetric case (see fig. 4). As in figures 1 and 2, if we suppose complete ignorance of the effects

$$H(E) = 1[\text{bit}].$$

Although in figure 4 two values of the cause can lead to the same effect, knowing the value of the cause fully specifies the value of the effect just as effectively as it does in figure 1. Thus

$$H\left(E|\widehat{C}\right) = 0[\text{bits}].$$

Therefore, the difference in uncertainty about the effect between not knowing the value of the cause and knowing it is

$$I\left(E;\widehat{C}\right) = 1[\text{bit}].$$

Here again, knowing the cause provides as much information about the effects as in figure 1, but because the repertoire of states of the causal variable is now larger, some values lead to the same effects (this can happen if we increase the level of detail in our description of the cause). Notice that

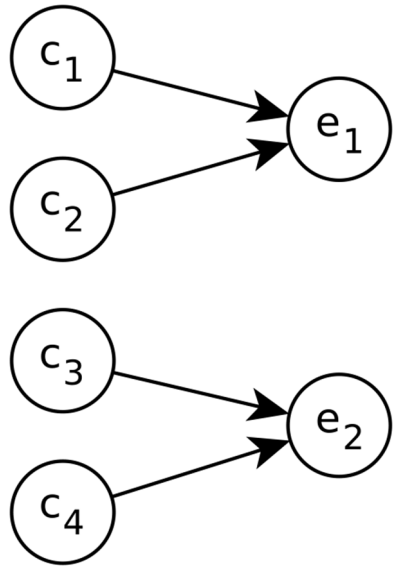

Figure 4. Different values of the cause lead to the same outcome.

https://doi.org/10.1086/682914 Published online by Cambridge University Press

this will not matter if we are interested in controlling the value of the effect: applying c_1 or c_2 will deterministically lead to e_1.

Furthermore, we can distinguish between figures 1 and 4 if we introduce a fourth quantity, that is, the entropy characterizing the repertoire of the cause, which in these two cases is the maximum entropy. In figure 1 the entropy $H(\widehat{C}) = 1$[bit], whereas in figure 4, $H(\widehat{C}) = -\sum_1^4 1/4 \log_2(1/4) = 2$[bits].

Thus, both the conditional entropy $H(E|\widehat{C})$ and the mutual information $I(E;\widehat{C})$ capture aspects of the intuition that causes differ in 'specificity'. Because the prior uncertainty $H(E)$ is not constant—it depends in particular on the size of the repertoire of effects—both measures are needed. The mutual information $I(E;\widehat{C})$ measures how much a cause specifies an effect. The conditional entropy $H(E|\widehat{C})$ measures how much an effect is determined when knowing the value set for the cause.

In the cases considered here, if $H(E|\widehat{C}) = 0$, then manipulating C provides complete control over E. This corresponds to Woodward's observation (2010, 305) that it is more important that the mapping from C to E is a surjective function than that it is also bijective. Woodward's notion of a *fine-grained control*, however, would be better represented using $H(E)$ and $I(E;\widehat{C})$. That is, fine-grained control requires that the repertoire of effects is large and that a cause screens off many of them (recall that we are currently dealing only with nominal variables). In the ideal case, $H(E)$ would tend toward infinity, and $I(E;\widehat{C})$ would tend toward $H(E)$.

3. Comparing Two Variables. We now have a proposal for a measure of causal specificity:

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

It is important to note that, while mutual information is a symmetric measure—$I(X;Y) = I(Y;X)$—the mutual information between an intervention and its effect is not symmetrical because the fact that interventions on C change E does not imply that interventions on E will change C: in general, $I(\widehat{C};E) \neq I(\widehat{E};C)$.

Recall that the aim of producing a measure of causal specificity was to use it to compare different causes of the same effect. So we need to look at a case in which an effect depends on more than one upstream causal variable and compare the mutual information they carry. To do so we explore some increasingly complex cases involving gene transcription. In each case we focus on (messenger) RNA as the effect variable and look at the relative specificity of different upstream causal variables.

We begin with a simple case that has already been discussed in the literature, namely, comparing the causal contributions of RNA polymerase and

https://doi.org/10.1086/682914 Published online by Cambridge University Press

DNA coding sequences to the structure of a messenger RNA (Waters 2007). Both are causes of RNA, since manipulating either makes a difference to the RNA. Polymerase is like the radio on/off button, and the DNA is like the channel tuner, with a number of settings.[3]

We can formalize this in the following way (fig. 5). There are two causal variables, DNA and POL, and one effect variable, RNA. Each variable can take on a number of values. Assume, for now, that there are four possible DNA sequences (d_1, d_2, d_3, d_4) and that the RNA polymerase is either '*present*' or '*absent*'. Our effect variable can thus take on five values—four correspond to the RNA sequences (r_1, r_2, r_3, r_4) transcribed from the DNA, and one is a state we call r_0, which occurs when there is no transcription. In order to calculate the mutual information, we need to assign each of the values a probability, and these must sum to 1. We begin by simply assigning uniform probabilities over the causal variables, DNA and POL. What does our specificity measure tell us about the two causal variables in this simple scenario?

When we do the calculation (see the supplementary online companion piece, Griffiths et al. [2015], sec. 1), interventions on either DNA or POL carry the same amount of mutual information:

$$I(\widehat{\text{DNA}}; \text{RNA}) = p(\widehat{\text{POL}}) \times H(\widehat{\text{DNA}}) = 0.5 \times 2 = 1[\text{bit}]$$
$$I(\widehat{\text{POL}}; \text{RNA}) = H(\widehat{\text{POL}}) = 1[\text{bit}].$$

They are (given our working assumptions) equally causally specific. That might seem odd, as the DNA sequences can take on four different values, and the polymerase is simply 'present' or 'absent'. Our measurement seems to be saying that there is no difference between on/off switches and tuning knobs. What has gone wrong?

To understand why this happens, recall that mutual information measures how much information on average we get by looking at a causal variable. Notice that the value of $\widehat{\text{DNA}}$ is irrelevant if $\widehat{\text{POL}}$ = absent, and our uniform distribution sets the probability of this at 0.5. So half the time, when we look at the value of $\widehat{\text{DNA}}$, we learn nothing about the system. When $\widehat{\text{POL}}$ = present, knowing the value of $\widehat{\text{DNA}}$ is useful: it delivers 2 bits of information. In short, half the time, $\widehat{\text{DNA}}$ gives us 0 bits of information, and the other half of the time 2 bits. Hence, 1 bit on average.

What this shows is that our proposed measure for causal specificity is sensitive to the probability distribution of the causal variables. This means that either our specificity measure is incorrect, or Woodward's INF (sec. 2) is missing something, because that condition makes no mention of the prob-

3. Because we do not impose an order on the values of the DNA variable, it is more like a digital tuner, to which any combination of digits can be entered, than an analogue tuning dial.

https://doi.org/10.1086/682914 Published online by Cambridge University Press

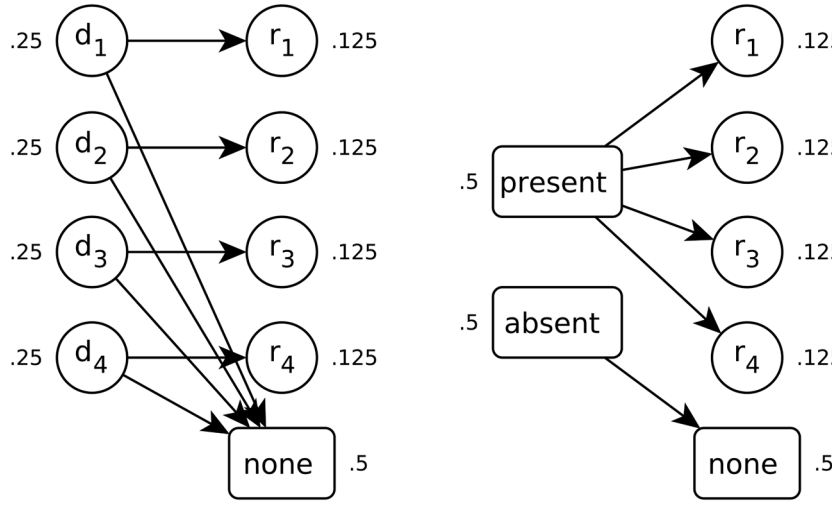

Figure 5. Causal mapping and probability distributions for DNA and RNA (*left*) and POL and RNA (*right*).

ability distributions over the variables. In the next section, we see that this dilemma corresponds to two different approaches to causal specificity.

4. Specific Actual Difference Making. The suggestion that the actual probability distributions of the causal variables matters when assessing which causes are significant is an idea we have heard before. Waters argues that in order to pick out the significant causes, you need to know the actual difference makers. For example, even when it is possible to manipulate POL (which identifies it as a potential cause), if there is no actual difference in POL in a population of cells, as Waters assumes, then it is not a significant cause. Waters's notion of an "actual difference maker" (2007, 567) can be related to our specificity measure.

Waters treats the question whether a variable exhibits actual variation as though it were a binary choice, but it makes sense to treat it as continuous. The 'actual variation' is the entropy of the variable.

To show how this idea fits into our specificity measure, consider how the mutual information (specificity) of each of our two variables $\widehat{\text{DNA}}$ and $\widehat{\text{POL}}$ with RNA changes as we vary the probability distribution of $\widehat{\text{POL}}$ (which, in turn varies its entropy). In figure 6, each value on the *X*-axis represents a different case. These range from cases in which the probability of present is 0 (polymerase is never around) to systems in which the probability of present is 1 (polymerase is always around). In these extreme cases, the variable has become a fixed background factor and doesn't actually vary, and thus

https://doi.org/10.1086/682914 Published online by Cambridge University Press

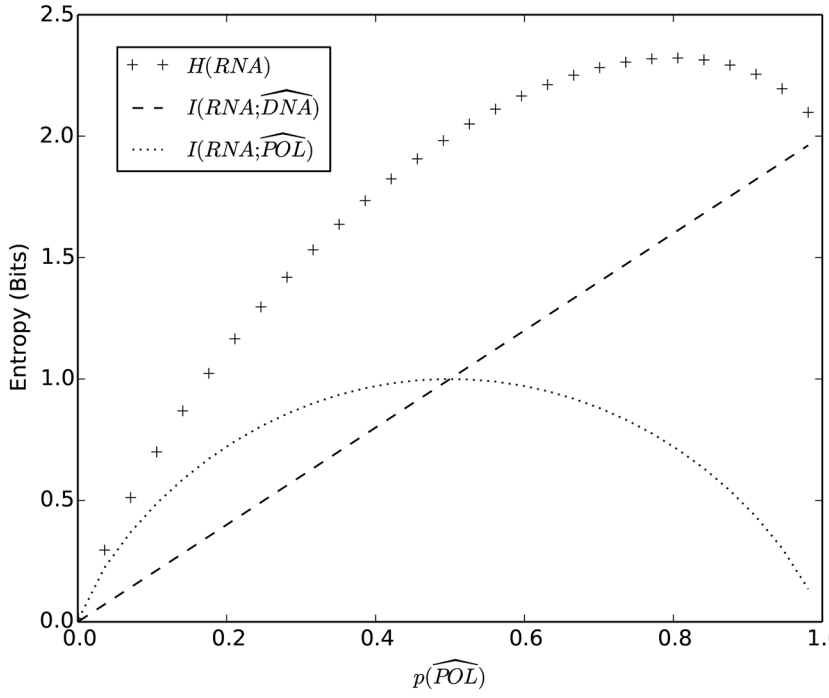

Figure 6. Effects of changing probability of $\widehat{\text{POL}}$ = present on several informational measures: the entropy of RNA (the effect), the mutual information between RNA and $\widehat{\text{DNA}}$, and the mutual information between RNA and intervening on the presence of polymerase. It can be shown that $H(\text{RNA}) = I\big(\text{RNA}, \widehat{\text{DNA}}\big) + I(\text{RNA}, \widehat{\text{POL}}) = p\big(\widehat{\text{POL}}\big) \times H\big(\widehat{\text{DNA}}\big) + H\big(\widehat{\text{POL}}\big)$ (see Griffiths et al. 2015, sec. 1). The variation in the effect can thus here be decomposed into the respective contributions of the causes.

the entropy $H\big(\widehat{\text{POL}}\big)$ is 0. When the probability of present is 0.5, $\widehat{\text{POL}}$ is maximally variable and has maximum entropy. The mutual information between $\widehat{\text{POL}}$ and RNA is also maximized at this point. Notice also, that as we increase $p(\text{present})$ to 1, the mutual information between $\widehat{\text{DNA}}$ and RNA increases. When $\widehat{\text{POL}}$ = present all the time, the full 2 bits of information about RNA can be found in $\widehat{\text{DNA}}$. Our proposed measure of specificity captures two things: the extent to which a relationship approaches a bijection (Woodward's INF) and the degree to which the cause is an actual difference maker (i.e., the cause also has high entropy). So the mutual information measure appears to capture the degree to which a cause is a 'specific actual difference maker' (SAD; Waters 2007).

Within our information theoretic framework there is a clear difference between the SAD concept and Woodward's INF. SAD uses the actual prob-

https://doi.org/10.1086/682914 Published online by Cambridge University Press

ability distribution over the values of a causal variable in some population. INF makes no distinction between the states of a causal variable. We represent this by supposing that the variable has maximum entropy: all its states are equiprobable. This makes sense when we recall that for Woodward causal variables are sites of intervention. For idealized external agents intervening on the system, the value of a causal variable is whatever they choose to make it.

It is possible to find different scientific contexts in which biologists seem to approach causal relationships in ways that correspond to SAD and INF respectively. Waters argues that classical genetics of the Morgan school was only concerned to characterize causes that actually varied in their laboratory populations (2007). Griffiths and Stotz argue that some work in behavioral developmental and much work in systems biology sets out to characterize the effect on the system of forcing all causal variables through their full range of potential variation (2013, 198–99). This kind of research, they argue, is done with the aim of discovering new ways to intervene in complex systems. The information theoretic framework allows us to distinguish between the specificity of potential (INF) and actual (SAD) difference makers. Our measure of causal specificity sheds light on another issue that we discussed in our introduction. Weber proposed that the specificity of a causal relationship is simply the range of values of the variables across which a causal relationship holds, or what Woodward calls the "range of invariance" (Woodward 2003, 254). Woodward rejected this idea because a causal relationship might hold across a large range of invariance but fail to be bijective. Our information theoretic framework captures both why Weber makes this suggestion and why Woodward's additional condition is needed. Weber's point corresponds to the fact that 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 will not do to increase the number of values of a cause variable unless the additional values of the cause map onto distinct values of the effect. Increasing the entropy of the cause variable will not increase mutual information when no additional entropy in the effect variable is captured. This is why the mutual information between the variables is the same in figures 1 and 4. In terms of figure A1, such an increase in the size of region $H(X)$ would be confined to the subregion $H(X|Y)$ with no increase in subregion $I(X;Y)$. The same point, of course, holds mutatis mutandis for the effect variable.

In addition to the SAD and INF conceptions of specificity, there is a third option corresponding to a suggestion by Weber that causal specificity should be assessed on the assumption that causal variables are neither restricted to their actual variation in some population nor allowed to vary freely but instead restricted to their 'biologically normal' range of variation: "What we need is a distinction between *relevant* and *irrelevant counterfactuals*, where

relevant counterfactuals are such that they describe *biologically normal possible interventions*" (Weber 2013, 7). We call this REL. Weber tells us that a biologically normal intervention must (1) involve a naturally occurring causal process and (2) not kill the organism. More work is obviously needed to make this idea precise, but we see in section 5 that even in this crude form REL provides a useful framework for modeling actual cases. At a practical level, we interpret REL as assessing causal specificity with a uniform probability distribution within the range of variation in the variable that would be produced by known mechanisms acting on relevant timescales for the causal processes we are trying to model.

5. Distributed Causal Specificity. We have suggested that causal specificity can be measured by the amount of mutual information between variables representing cause and effect. This implies that the degree of specificity of a causal relationship depends on the probability distributions over the two variables, and we have argued that this relates to Waters's claim that significant causes are specific actual difference makers. We have also taken on board Weber's point that it may be more interesting to explore, not the strictly actual variation but the 'biologically normal' variation (REL). In this section we apply our measure to a more complex case than the roles of RNA polymerase and DNA in the production of RNA, namely, the role of splicing factors and DNA in the production of alternatively spliced mRNA. Importantly, we also attempt to fill out these measures with realistic values.

In contemporary molecular biology the image of the gene as a simple sequence of coding DNA with an adjacent promoter region is very much a special case. This image remains important in the practice of annotating genomes with 'nominal genes'—regions that resemble reasonably closely the textbook image (Fogle 2000; Burian 2004; Griffiths and Stotz 2007, 2013). But a more representative image of the gene, at least in eukaryotes, is a complex region of DNA whose structure is best understood top down in light of how that DNA can be used in transcription and translation to make a range of products. Multiple promoter regions allow transcripts of different lengths to be produced from a single region. This and other mechanisms allow the same region to be transcribed with different reading frames. mRNA editing allows single bases in a transcript to be changed before translation. *Trans*-splicing allows different DNA regions to contribute components to a single mRNA. Here, however, we concentrate on the most ubiquitous of these mechanisms, alternative *cis*-splicing, a process known to occur, for example, in approximately 95% of human genes (nominal genes).[4]

4. For more detail on all these processes, see Griffiths and Stotz (2013). It may be useful to know that the prefix *trans-* denotes processes involving a different region of the DNA, while the prefix *cis-* denotes processes involving the same or an immediately adjacent region.

Genes are annotated with two kinds of regions, exons and introns. The typically much larger introns are cut out of the corresponding mRNA and discarded. In alternative *cis*-splicing (hereafter just 'splicing') there is more than one way to do this, giving rise to a number of different proteins or functional RNAs. For simplicity, we ignore mechanisms such as exon repetition or reversal, and the fact that exon/intron boundaries may vary, and treat this process as if it were simply a matter of choosing to include or omit each of a determinate set of exons in the final transcript.

With alternative splicing, the final product is codetermined by the coding region from which the transcript originates and some combination of *trans*-acting factors that bind to the transcript to determine whether certain exons will be included or excluded. These factors are transcribed from elsewhere in the genome, and their presence at their site of action requires the activation of those regions and correct processing, transport, and activation of the product. The entire process thus exemplifies the themes of 'regulated recruitment and combinatorial control' characteristic of much recent work on the control of genome expression (Ptashne and Gann 2002; Griffiths and Stotz 2013). We simplify this by representing alternative splicing as a single variable, each of whose values correspond to a set of *trans*-acting factors sufficient to determine a unique splice variant.

The role of alternative splicing is well known, but recent work on causal specificity does not treat this issue with much care. Weber states that, "depending on what protein factors are present, a cell can make a considerable variety of different polypeptides from the same gene. Thus we have some causal specificity, but it is no match for the extremely high number of different protein sequences that may result by substituting nucleic acids" (Weber [2006] endorsed by Waters [2007, n. 28]). Here Weber seems to be making a problematic comparison of the actual range of splicing variants present in a single organism with the possible genetic variants that could be produced by mutation. Recently, Weber has explicitly argued for this comparison, stating that only 'biologically normal' interventions should be considered and that variation in DNA coding sequences is biologically normal. He concludes that DNA and RNA deserve a unique status among biological causes because their biologically normal ability to vary in a way that influences the structure of gene products is "vastly higher (i.e., many orders of magnitude) than that of any other causal variables that bear the relation INF to protein sequences (e.g., splicing agents)" (Weber 2013, 31).

We are not convinced that it is a meaningful comparison to take, for example, the *Drosophila* DSCAM gene,[5] with 38,016 splice variants all or

5. In the *Drosophila* receptor DSCAM (Down syndrome cell adhesion molecule), 4 of the 24 exons of the *Dscam* gene are arranged in large tandem arrays, whose regulation is an example of mutually exclusive splicing. One block has 2 exons (leading to one of

most of which are found in any actual population of flies, and say that alternative splicing has negligible causal specificity because this number of variants is much lower than the number of variants possible by mutation of the DSCAM coding sequence with no limit on the number of mutational steps away from the actual sequence (Weber 2013, 19). This seems to be a classic example of the way in which philosophers are unable to sustain parity of reasoning (Oyama 2000, 200ff.) when thinking about DNA. The principle that only 'biologically normal' variation should be counted is rigorously enforced for nongenetic causes but not for genetic causes. An anonymous reviewer has pointed out that even when variation in the coding DNA sequence is restricted to a small (and thus 'biologically normal') number of mutational steps, the number of possible variants expands very rapidly because of the sheer number of nucleotides (about 6,000 in DSCAM). Which ranges of variation in splicing agents and coding sequences it is meaningful to compare will depend on the biological question being addressed, as we now discuss.

To make a meaningful comparison between splicing agents and coding sequences it is also necessary to specify a population of entities across which they produce variation. Waters (2007) focuses on two examples in which most of the actual variation is caused by variation in DNA. The first is the population of phenotypic *Drosophila* mutants in a classical genetics laboratory. The second is the population of RNA transcripts at one point in time in a bacterial cell in which there is no alternative splicing. Obviously, neither of these cases is a useful one with which to evaluate the causal specificity of splicing agents, but they do exemplify two important classes of comparisons we might make. First, we might compare the variation between individuals in an evolving population and seek to determine whether variation in DNA coding sequences is the sole or main specific difference maker. Second, we might consider the transcriptome (population of transcripts) in a single cell, either at a time or across time, and ask whether variation in DNA coding sequences is the sole or main specific difference maker between these transcripts. Weber also considers examples of these two kinds. However, neither Waters nor Weber considers a third important case, which is the variation between cells in an organism, both spatial and temporal. This is the kind of variation that needs to be explained to understand develop-

two alternative transmembrane segments), the others contain respectively 12, 48, and 33 alternative exons (leading to 19,008 different ecto-domains). A neuronal cell differs not only with respect to which one of the 38,016 variants (in a genome of about 15,000 genes) it expresses but in the exact ratio in which it expresses up to 50 variants at a time. Each block of exons seems to possess a unique mechanism that ensures that exclusively only one of the alternative exons is included in the final transcript. For details and references, see Griffiths et al. (2015), sec. 3.

ment, the context in which controversy over the causal roles of genes and other factors most often arises.

Both actual and relevant ('biologically normal') variation in genes or splicing agents will be different in each of these three cases. In the case of an evolving population, mutation is a biologically normal source of variation, but without any limit on the number of mutational steps from the current sequence, let alone variation in genome size or ploidy, the values of the DNA variable would simply be every possible genome, which would be both unmanageable and biologically meaningless. It might seem natural to exclude any other sources of variation on the grounds that they are not heritable, but a number of evolutionary theorists would hotly dispute this (e.g., Jablonka and Lamb 2005; Bonduriansky 2012; Uller 2012). Furthermore, the machinery of splicing also changes over evolutionary time, so in the evolutionary case the 'biologically normal' variation in splicing is greater than the amount of variation observed in any actual population. These are very complex issues, and we cannot undertake the extensive work of establishing the relevant ranges of variation of genetic and other variables in the evolutionary case in this article.

Instead, we examine the simpler case suggested by Waters, the population of RNA transcripts in a single cell at one time. But while Waters considers only cells with no splicing, we consider cells with splicing, so as to make a comparison possible. For the transcriptome of a single cell at a time, the relevant values of the DNA variable are the different sequences that can be transcribed by the polymerase. If we ignore complexities such as multiple promoters, we can set this equal to the nominal gene count in the genome, so that realistic figures are available. The values of the DNA variable will be weighted by the probability of each gene being expressed. The values of the splicing variable can be set equal to the number of splicing variants from each gene, weighted by the probability of each splice variant.

We now propose a quantification of the respective causal specificity of the DNA and splicing variables for this very simple case. To further simplify the exposition, we assume that the polymerase is always present (an assumption that can be relaxed easily; see Griffiths et al. 2015, sec. 2). We focus on the mutual information measure outlined above, but we need to take a slightly different approach to compare the specificity of splicing with the specificity of DNA, for we assume that splicing factors are recruited only after a given strand of DNA has been transcribed. We do this because, in reality, it is not the case that any set of splicing factors can be combined with any gene. If we were to model splicing in this way, then the outcome of most combinations of genes and sets of splicing factors would be that the system fails to produce any biologically meaningful outcome. So it is both simpler and more biologically realistic to represent the process sequentially, as the transcription of an mRNA followed by the recruitment of a set of

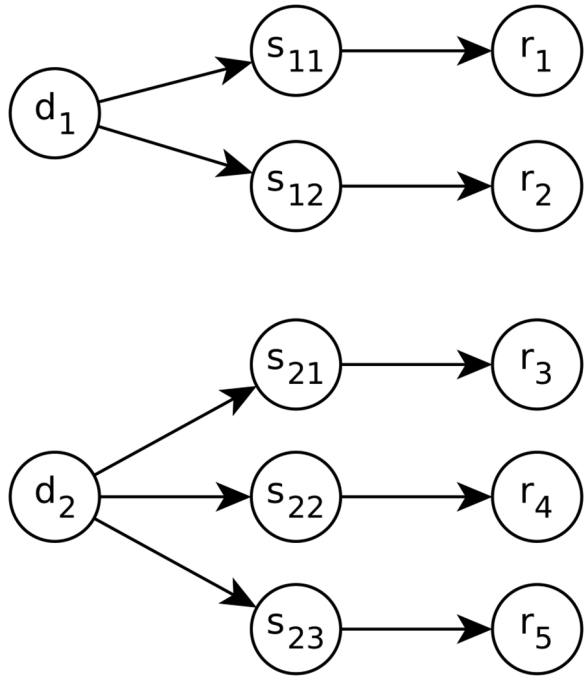

Figure 7. Simplified relationship between DNA (D), splicing (S), and RNA (R) variables, assumed in the models in section 5. Selection of a value for DNA opens a proper set of possibilities of splicing. There is a bijective relationship between splicing and RNA.

splicing factors. In other words, the transcription of a given DNA strand opens a set of possibilities among a proper set of the possible combinations of splicing factors (fig. 7). This entails that the information in splicing factors, measured by $H(\widehat{S})$, contains all the information in DNA, measured by $H(\widehat{D})$:[6]

$$H(\widehat{D}, \widehat{S}) = H(\widehat{S}).$$

Because the entropy in the DNA variable is conserved in the entropy of the splicing variable, the mutual information between RNA and splicing will also conserve the mutual information between RNA and DNA. Thus, we need a way to decompose our causal specificity measure into two components, isolating the separate contributions of DNA and splicing.

6. In the following equations, D and R are the variables DNA and RNA (see fig. 6), and S is the splicing variable.

https://doi.org/10.1086/682914 Published online by Cambridge University Press

As mentioned above, we treat the splicing process as if it were simply a matter of choosing to include or omit each of a determinate set of exons in the final transcript. Each value of our splicing variable corresponds to a set of *trans*-acting factors sufficient to determine a unique splice variant of the RNA. In other words, we consider a bijective relationship between sets of splicing factors (once recruited) and RNA variants. This bijection entails that the mutual information between RNA and interventions on splicing $I(R;\widehat{S})$ is simply equal to the so-called self-information of splicing, $I(\widehat{S};\widehat{S})$, which is itself equal to the entropy of splicing $H(\widehat{S})$. We can then decompose the entropy of splicing according to well-known chain rules:

$$I(R;\widehat{S}) = I(\widehat{S};\widehat{S}) = H(\widehat{S}) = H(\widehat{D},\ \widehat{S}) = H(\widehat{S} \mid \widehat{D}) + H(\widehat{D}).$$

Noting that $I(R;\widehat{D}) = H(\widehat{D})$ when the polymerase is always present (see sec. 5) and that $I(R;\widehat{S}|\widehat{D}) = H(\widehat{S}|\widehat{D})$ (see Griffiths et al. 2015, sec. 2), we can rewrite the equation as

$$I(R;\widehat{S}) = I(R;\widehat{S}|\widehat{D}) + I(R;\widehat{D}).$$

This equation provides a decomposition of the mutual information between RNA and splicing, $I(R;\widehat{S})$, into two components, the mutual information between RNA and DNA, $(R;\widehat{D})$, and the mutual information between RNA and splicing conditional on DNA, $I(R;\widehat{S}|\widehat{D})$. Because $I(R;\widehat{S}|\widehat{D}) \geq 0$, this entails that $I(R;\widehat{S}) \geq I(R;\widehat{D})$. If we simply proceed as before, taking mutual information as a measure of causal specificity, we find that the specificity of splicing is always greater than or equal to the specificity of DNA. As we mentioned above, however, we need to account for the fact that all the information contributed by DNA to RNA is conserved in the splicing variable. Fortunately, we can decompose the mutual information in splicing to obtain two terms that represent the contribution from the DNA and the contribution from the splicing process. The term $H(\widehat{D})$ in the decomposition of $I(R;\widehat{S})$ represents the amount of information that is preserved in the splicing process but originates in the DNA. The variation in RNA properly coming from the splicing process is represented by the term $H(S|\widehat{D})$—a term that, roughly, reflects the number of splicing variants per DNA strand. Thus, if one wants to compare the causal specificity of splicing and DNA, one needs to know which of these two terms, $H(\widehat{D})$ and $H(\widehat{S}|\widehat{D})$, makes the greatest contribution to $(R;\widehat{S})$.

The answer will crucially depend on the biological system. In *Drosophila*, an important determinant of neuronal diversity is the single *Dscam* gene with 38,016 splice variants (see Griffiths et al. 2015, sec. 3). This gives a maximum entropy of approximately $\log_2(38,016) = 15.2$ bits for

https://doi.org/10.1086/682914 Published online by Cambridge University Press

$H(\widehat{S}|\widehat{D})$, compared with 0 bits for $H(\widehat{D})$. The diversity of this class of transcripts in *Drosophila* is entirely explained by posttranscriptional processing.[7]

The homologues of this gene in humans, *Dscam* and *Dscam*-like, present a very different picture. The number of splicing variants per gene appears to be no greater than three. Assuming that the transcription of each of these two DNA regions is equiprobable, this gives a maximum entropy of approximately 1.6 bits for $H(\widehat{S}|\widehat{D})$, to be compared with 1 bit for $H(\widehat{D})$. DNA and splicing are roughly equal determinants of diversity in this class of transcripts.

A more meaningful comparison to the *Dscam* case in *Drosophila*, however, may be other classes of vertebrate cell-surface proteins. Generalizing from real cases,[8] we might imagine a class of transcripts that derives from, say, 100 related genes, each of which has 150 splicing variants. Assuming once again that the transcription of any of these DNA regions is equiprobable, this gives approximately 7.2 bits for $H(\widehat{S}|\widehat{D})$, to be compared with approximately 6.6 for $H(\widehat{D})$. Both DNA and splicing variables are important determinants of diversity in this class of transcripts.

Assigning specificity to the causes of transcript diversity in a single cell at a time is relatively tractable. The analyses just given could, in principle, be extended to the entire transcriptome at one stage in the life cycle of a well-studied system such as yeast. But this would be of limited interest. What is at stake in disagreements over the relative causal roles of coding regions of DNA and other factors in gene expression would be better represented by comparing the transcriptome in a cell at different times in its life cycle, or comparing transcriptomes between different cell types in an organism. These comparisons are both ways of thinking about development—the process by which regulated genome expression produces an organism and its life cycle. In comparing the same cell across times, a critical feature is that which genes are transcribed and how their products are processed depends on transcription and processing at earlier times. For the population of cells in an

7. Our decision to use actual figures for genes and isoforms but assume equiprobability (maximum entropy) for each variable can be justified in this particular case on both the INF and REL approaches (sec. 4). The data required for Waters's SAD approach are not available, but there is no reason to suppose it would give qualitatively different results.

8. *Dscam* is homologous between almost all animals, but in vertebrates the two homologous genes, *Dscam* and *Dscam*L1, do not encode multiple isoforms. There are, however, several hundred cell adhesion and surface receptor genes in vertebrates: the Ig superfamily, as well as integrins, cadherins, and selectins. This genetic diversity is combined with complex regulatory patterns, albeit not on the scale of the *Dscam* expression in *Drosophila*. The three neurexin genes display extensive alternative splicing, a process that can potentially generate thousands of neurexin isoforms alone. For details and references, see Griffiths et al. (2015), sec. 3.

organism, somatic mutations that could arise during development become relevant, leading to the need to say something about the number of mutational steps that counts as a 'biologically realistic' intervention on this variable. We hope to confront these complexities in future work.

6. Conclusion. Causal specificity is the label given to an intuitive distinction among the many conditions that are necessary to produce an effect. The specific causes are those variables that can be used for fine-grained control of an effect variable. It has been suggested that a specific relationship between two variables is one that resembles a bijective mapping between the values of the two variables (Woodward 2010). The concept of causal specificity can be clarified considerably by going a step further and attempting to measure it.

Our quantitative measure of specificity starts from the simple idea that the more specific the relationship between a cause variable and an effect variable, the more information we have about the effect after we perform an intervention on the cause. Section 2 used information theoretic measures to express this idea. We found that if the conditional entropy of the effect on interventions on the cause $H(E|\widehat{C} = 0)$, then manipulating C provides complete control over E. We argued, however, that the idea of sensitive manipulation, or fine-grained influence (Woodward 2010), would be better represented by measuring the entropy of the effect $H(E)$ and the mutual information between cause and effect $I(E;\widehat{C})$. Fine-grained influence requires both that the repertoire of effects is large and that the state of the cause contains a great deal of information about the state of the effect. In the ideal case, $H(E)$ would tend toward infinity, and $I(E;\widehat{C})$ would tend toward $H(E)$.

Section 3 examined the behavior of $I(E;\widehat{C})$ as a measure of causal specificity (SPEC). The behavior of the measure depends on the probability distributions over the states of the variables, as well as the structure of the causal graph. Other things being equal, a variable with many states that are rarely or never occupied is a less specific cause than one equally likely to be in any of its states, that is, one with higher entropy. Section 4 showed that this feature is a strength of our proposed measure. It is in line with the qualitative reasoning of Waters (2007), who argues that the property which justifies singling out one cause as more significant that another can be its specificity with respect to the actual variation seen in some population, and of Weber (2013), who suggests that we focus on the somewhat wider class of 'biologically normal' variation.

The sensitivity of our measure to the underlying probability distributions contrasts with presentations of causal specificity in which it is assumed that the value can be inferred from the structure of a causal graph. Our attempt to quantify specificity forces this assumption to become explicit. The least ar-

https://doi.org/10.1086/682914 Published online by Cambridge University Press

bitrary way to represent this assumption in our models would seem to be to make all values of the causal variables equiprobable. Making this assumption is probably not appropriate for settling the disputes about the relative significance of various causal factors in biology with which Waters and Weber are concerned. However, in the broader context of the interventionist account of causation, it may be entirely appropriate, because causal variables are the sites of voluntary intervention by an idealized agent.

Section 5 used our measure to assess the relative specificity of different causes that contribute to the same effect. The idea of specificity has been used to argue that DNA sequences are the most significant causes because of their supposedly unrivalled degree of specificity. Our discussion revealed that this is completely premature. First, it is necessary to specify the causal process in question. The causes of individual differences in an evolving population are quite different from the causes of transcript diversity in a single cell, and different again from the causes of spatial and temporal diversity among the cells of a single organism. We constructed a simple model with which we were able to quantify the specificity of a DNA coding sequence and of splicing factors with respect to transcript diversity in a single cell at a time. We showed that the relative specificity of these two variables can be very different for different classes of transcripts. The idea that DNA obviously has an unrivalled degree of specificity seems to arise because earlier qualitative discussions implicitly compared the actual variation in the splicing variable within cells to the possible variation in the DNA variable on an evolutionary timescale.

While it seems plausible to us that the specificity of coding DNA as a cause of evolutionary change is very high, we pointed out that proper exploration of this would require serious thought about which range of variation in the DNA variable can be meaningfully compared with which range of variation in other cellular mechanisms. Similar work would be needed before our measure can be applied to what is arguably the most pressing case, namely, the relative specificity of different causes in development. We hope to focus on this case in future work.

We believe that the work reported here amply demonstrates the philosophical payoff of developing quantitative measures of causal specificity. However, a great deal remains to be done. First, although our measures provide information about causal specificity rather than the presence of causation per se, in future work we hope to provide an information theoretic statement of the interventionist criterion of causation. Second, our measure of specificity is only one of several information theoretic measures that can be used to characterize causal relationships. In future work we hope to explore the potential of these other measures for the philosophy of causation. Third, and perhaps most urgently, we gave only minimal attention in this article (in sec. 4) to the ways in which the relationship between two vari-

ables can be affected by additional variables. In a forthcoming paper we extend our framework to deal with these interactions.

Appendix

A Primer on Information Theory

Information theory provides us with tools to measure uncertainty and to measure the reduction of that uncertainty. Importantly, for our purposes, it tells us how information about the value of one variable can reduce the uncertainty about the value of another, related, variable.

The simplest case occurs when a discrete variable has only two values, which can then be known by answering a single question (e.g., by yes or no). The answer is said to convey 1 unit of information (a *bit*). If the set of possible values for the variable now contains 2^n equally likely elements, we can remark that n dichotomous questions (n bits) are needed to determine the actual value of the variable. The quantity of information contained in knowing the actual value is thus $n = \log_2(2^n)$. If we adopt a probabilistic framework in which each possible value has equal probability $p = 1/2^n$, we can say that knowing any actual value of the variable brings $-\log_2 p$ bits of information. When the values are not equiprobable, the average information gained by knowing an actual value of the variable is measured as an average over the probabilities of the different values. This quantity is the *entropy* of the probability distribution of the variable, defined as

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i),$$

where x_i represent values of the variable X, and N is the number of different values. Entropy measures the uncertainty about the value of the variable and is always nonnegative. Uncertainty is maximized (*maximum entropy*) when each value is equiprobable. Departing from uniformity will always make one (or more) values more probable, and so decrease uncertainty. In a similar way, increasing the number of possible values will increase uncertainty. All of the above can be generalized to cases in which the number of possible values is not a power of 2.

If X and Y are two random variables (with respectively N and M different values, noted x_i, y_j), we can define the entropy of the couple X, Y:

$$H(X, Y) = -\sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \log_2 p(x_i, y_j).$$

https://doi.org/10.1086/682914 Published online by Cambridge University Press

This enables us to define the *conditional entropy*, representing the amount of uncertainty remaining on Y when we already know X:

$$\begin{aligned} H(Y\,|\,X) &= H(X,Y) - H(X) \\ &= -\sum_{i=1}^{N} p(x_i) \sum_{j=1}^{M} p(y_j\,|\,x_i) \log_2 p(y_j\,|\,x_i). \end{aligned}$$

In a similar way, the *mutual information*, that is, the amount of redundant information present in X and Y, is obtained by

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i,\, y_j) \log_2 \frac{p(x_i,\, y_j)}{p(x_i)p(y_j)}. \end{aligned}$$

Mutual information can be thought of as the amount of information that one variable, X, contains about the other, Y (normalized variants of mutual information are available).

Conditional entropy is null, and mutual information is maximal, when Y is completely determined by X. Note that conditional entropy is generally asymmetric, while mutual information is always symmetric:

$$\begin{aligned} H(X\,|\,Y) &\neq H(Y\,|\,X) \\ I(X;Y) &= I(Y;X). \end{aligned}$$

The relationships between these three different measures are represented in figure A1. See Cover and Thomas (2012) for more detail.

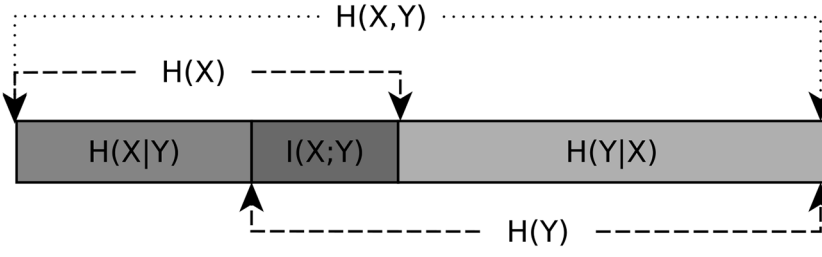

Figure A1. Relationships between the different informational measures, entropy $H(X)$, conditional entropy $H(X|Y)$, and mutual information $I(X;Y)$.

https://doi.org/10.1086/682914 Published online by Cambridge University Press

Appendix B

Causal Modeling

Causal modeling provides us with the tools to track the effects of interventions on a system. Where statistical modeling would look at statistical associations between supposed causes and supposed effects, causal modeling introduces the requirement of intervening on the system to compute the causal effect. More precisely, consider a causal model consisting of

1. a set of functional relationships $x_i = f(pa_i, u_i)$, $i = 1 \ldots n$, where x_i is the value of the variable X_i being caused by X_i's parent variables pa_i, according to some function f, given some background conditions u_i
2. a joint distribution function $P(u)$ on the background factors.

Then the simplest 'atomic' intervention consists in forcing X_i to take some value x_i irrespective of the value of the parent variables pa_i, keeping everything else unchanged. Such an intervention can be written formally with the $do()$ operator. As Pearl writes: "Formally, this atomic intervention, which we denote by $do(X_i = x_i)$ or $do(x_i)$ [or $\widehat{x_i}$] for short, amounts to removing the equation $x_i = f(pa_i, u_i)$ from the model and substituting $X_i = x_i$ in the remaining equations. The new model when solved for the distribution of X_j, yields the causal effect of X_i on X_j, which is denoted $P(x_j \mid \widehat{x_i})$" (2009, 70).

The causal effect $P(x_j \mid \widehat{x_i})$ is to be contrasted with the observational conditional probability $P(x_j \mid x_i)$, which can be affected by confounding factors leading to spurious associations or spurious independence. Other recent works in mathematics and computer science have brought information theory together with causal modeling to study information processing in complex systems (Ay and Polani 2008; Lizier and Prokopenko 2010). These works also build on Pearl (2009) and are consistent with the work presented here. However, our approach and measures are significantly different, reflecting the fact that we start from a concern with 'causal selection' in a context of intervention and control. The differences between these approaches will be explored in a future paper. See Pearl (2009, esp. chap. 3) for more details.

REFERENCES

Ay, N., and D. Polani. 2008. "Information Flows in Causal Networks." *Advances in Complex Systems* 11 (1): 17–41.

Bonduriansky, R. 2012. "Rethinking Heredity, Again." *Trends in Ecology and Evolution* 27 (6): 330–36.

Burian, R. M. 2004. "Molecular Epigenesis, Molecular Pleiotropy, and Molecular Gene Definitions." *History and Philosophy of the Life Sciences* 26 (1): 59–80.

Cover, T. M., and J. A. Thomas. 2012. *Elements of Information Theory.* Hoboken, NJ: Wiley.
Fogle, T. 2000. "The Dissolution of Protein Coding Genes in Molecular Biology." In *The Concept of the Gene in Development and Evolution*, ed. P. J. Beurton, R. Falk, and H.-J. Rheinberger, 3–25. Cambridge: Cambridge University Press.
Garner, W. R., and W. McGill. 1956. "The Relation between Information and Variance Analyses." *Psychometrika* 21 (3): 219–28.
Griffiths, P. E., A. Pocheville, B. Calcott, K. Stotz, K. Karola, H. Kim, and R. Knight. 2015. "Measuring Causal Specificity: Supplementary Online Materials." PhilSci Archive, http://philsci-archive.pitt.edu/11593/.
Griffiths, P. E., and K. Stotz. 2007. "Gene." In *Cambridge Companion to Philosophy of Biology*, ed. M. Ruse and D. Hull, 85–102. Cambridge: Cambridge University Press.
———. 2013. *Genetics and Philosophy: An introduction*. New York: Cambridge University Press.
Jablonka, E., and M. J. Lamb. 2005. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge, MA: MIT Press.
Lewis, D. K. 2000. "Causation as Influence." *Journal of Philosophy* 97:182–97.
Lizier, J. T., and M. Prokopenko. 2010. "Differentiating Information Transfer and Causal Effect." *European Physical Journal* B 73 (4): 605–15. doi:10.1140/epjb/e2010-00034-5.
Oyama, S. 2000. *The Ontogeny of Information: Developmental Systems and Evolution*. 2nd rev. ed. Durham, NC: Duke University Press.
Pearl, Judea. 2009. *Causality.* Cambridge: Cambridge University Press.
Ptashne, M., and A. Gann. 2002. *Genes and Signals*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh et al. 2011. "Detecting Novel Associations in Large Data Sets." *Science* 334 (6062): 1518–24.
Ross, B. C. 2014. "Mutual Information between Discrete and Continuous Data Sets." *PLoS ONE* 9 (2): e87357.
Shannon, C. E., and W. Weaver. 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.
Stotz, K. 2006. "Molecular Epigenesis: Distributed Specificity as a Break in the Central Dogma." *History and Philosophy of the Life Sciences* 28 (4): 533–48.
Uller, T. 2012. "Parental Effects in Development and Evolution." In *The Evolution of Parental Care*, ed. N. J. Royle, P. T. Smiseth, and M. Kölliker, 247–66. Oxford: Oxford University Press.
Waters, C. K. 2007. "Causes That Make a Difference." *Journal of Philosophy* 104 (11): 551–79.
Weber, M. 2006. "The Central Dogma as a Thesis of Causal Specificity." *History and Philosophy of the Life Sciences* 28 (4): 595–609.
———. 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*, 1–44. Konstanz: University of Konstanz.
Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
———. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology and Philosophy* 25 (3): 287–318.
———. 2012. "Causation and Manipulability." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University. http://plato.stanford.edu/archives/win2012/entries/causation-mani/.