Quantifying Proportionality and the Limits of Higher-Level Causation and Explanation

Alexander Gebharter and Markus I. Eronen

Abstract

Supporters of the autonomy of higher-level causation (or explanation) often appeal to proportionality, arguing that higher-level causes are more proportional than their lower-level realizers. Recently, measures based on information theory and causal modelling have been proposed that allow one to shed new light on proportionality and the related notion of specificity. In this article we apply ideas from this literature to the issue of higher versus lower-level causation (and explanation). Surprisingly, proportionality turns out to be irrelevant for the question of whether higher-level causes (or explanations) can be autonomous; specificity is a much more informative notion for this purpose.

1 Introduction

2 Proportionality, Specificity, and Their Relevance for Higher-Level Causation and Explanation

3 Specificity and Proportionality Quantified

4 On the Limits of Higher-Level Causation and Explanation
- *4.1 On proportionality as an empirical matter*
- *4.2 On how common proportional higher-level causes are*
- *4.3 On the autonomy of higher-level causes*
- *4.4 On the autonomy of higher-level causal explanations*

5 Conclusion

Appendix

1 Introduction

The status of higher-level causation is of key importance for many philosophical and scientific issues, including explanation in the special sciences, mechanistic explanation, emergence, non-reductive physicalism, and mental causation. The question is whether higher-level properties

can in some sense be causally autonomous or indispensable with regard to their lower-level realizers, or whether lower-level (physical) causes always trump the corresponding higher-level causes. Sometimes the issue is also formulated in terms of which causal explanations are better or preferable: Should we always prefer lower-level causal explanations when they are available, or are there situations or structures where higher-level causal explanations are better?[1] On the one hand, higher-level causes and explanations (such as psychological causes and explanations) seem to be *prima facie* necessary for explaining the behaviour of complex systems (such as human behaviour). On the other hand, there are strong arguments that suggest that lower-level causes (or explanations) always trump higher-level causes (or explanations), the most famous of which is the causal exclusion argument (for example, Kim [2005]).

One important strategy, going back at least to Yablo ([1992]), has been to appeal to proportionality to argue that higher-level causes can be autonomous (see, for example, Shoemaker [2000]; List and Menzies [2009]; Zhong [2014]), or provide better or equally good explanations than lower-level causes (see, for example, Woodward [2010], [2018]; McLaughlin [2007]; Weslake [2013]). The idea is roughly that causes should be somehow commensurate to their effects (for details, see Section 2). Although this idea is quite intuitive, it has turned out to be difficult to spell out in a way that is clear and consistent. Many authors have also argued that it is a problematic or ill-defined notion that is not helpful for defending higher-level causation (see, for example, Bontly [2005]; Franklin-Hall [2016]; McDonnell [2017]). However, in recent years, advances in the causal modelling literature have led to opportunities to precisely define proportionality and the related notion of specificity. Several authors have proposed to characterize proportionality or specificity based on information theory (see, for example, Griffiths *et al.* [2015]; Pocheville *et al.* [2017]; Bourrat [2019]). In this article, we draw from this literature and apply ideas about how to quantify proportionality and specificity to the philosophical debate about higher-level causation. Based on these measures, we analyse causal structures involving higher- and lower-level causes.

Though the concepts of proportionality and specificity are neutral in nature, they can be

[1] Throughout the article we are only interested in non-pragmatic explanatory superiority (cf. Woodward [2018]). Basically everyone agrees that higher-level causal explanations can be superior for pragmatic reasons such as epistemic inaccessibility, lack of knowledge, or computational limitations.

applied to epistemic as well as to metaphysical contexts. Up to now information-theoretical measures have primarily been used in epistemic contexts, for example, as guides to choose the right or best level (or grain) of description (see, for example, Pocheville *et al.* [2017]). In contrast, in this article we apply them to the more metaphysical issue of the possibility of higher-level causal autonomy and to the question of whether higher-level explanations can be superior when all pragmatic considerations are bracketed. For answering these questions, it will be crucial how contexts where higher- and lower-level variables stand in relationships of supervenience constrain the behaviour of these measures.

In this article, we assume an interventionist approach to causation. This approach has two strands. The more philosophical strand, developed by Woodward ([2003]), analyses causation in terms of interventions.[2] The basic idea of this approach is that a variable C is causally relevant for another variable E if there are possible interventions on C that would lead to a change in E. Interventions can be roughly understood as ideal experimental manipulations that change C without influencing any other causes of E (for details, see Woodward [2003], Section 3.1.3). In our analysis, we apply the more general and more formal framework of causal Bayes nets (Spirtes *et al.* [1993]; Pearl [2000]), where causal structures are represented as causal graphs that are based on conditional independence relationships between variables (see the appendix). There are subtle differences between these two forms of interventionism (see, for example, Gebharter [2017b], Chapter 5), but they are not essential for our results.

Similarly, we rely on the interventionist approach to explanation and explanatory power (Hitchcock and Woodward [2003]; Woodward [2003]; Woodward and Hitchcock [2003]). According to this approach, explanatory power is a matter of providing answers to what-if-things-had-been-different questions, also known as w-questions. More precisely, a generalization is explanatory insofar as it can answer w-questions, and generalization A has more explanatory power than generalization B if it can answer a wider range of w-questions. As the w-questions approach is conceptually clear and widely held, we adopt it in this article, but at the same time we fully acknowledge that there are alternative accounts and further dimensions of explanatory power that may lead to different results (for example, Ylikoski and Kuorikoski [2010]).

[2] Note that this analysis uses causal terms and is, thus, not reductive (Woodward [2003]).

The structure of the article is as follows: In Section 2, we introduce proportionality and specificity in a more detailed way and sketch how they have been used to argue for the autonomy of higher-level causation and non-pragmatic higher-level explanation. In Section 3, we discuss how to quantify these notions. Based on Griffith *et al.*'s ([2015]) work and Bourrat's ([2019]) classification of different kinds of specificity we discuss three ways that cause and effect could line up ideally and specify measures for each case. We then identify one of these measures as an indicator for proportionality and another one as an indicator for causal power or influence. In Section 4 we then apply these measures to higher-level causation and explanation. We come up with a simple but representative causal model, which we then use together with the measures to study contexts where there is a supervenience relationship between the higher and lower levels. In particular, we will focus on the following interrelated questions arising for such contexts:

(1) Are there cases where higher-level causes are more proportional with respect to their effects than their lower-level realizers and if so, are higher-level causes always more proportional? (Section 4.1)

(2) How common are cases where higher-level causes are more proportional and can this question be answered *a priori*? (Section 4.2)

(3) Can higher-level causes be autonomous and can proportionality be used to support the autonomy of higher-level causation? (Section 4.3)

(4) Can higher-level causal explanations provide information about the explanandum phenomenon that goes beyond the information provided by their lower-level realizers and can proportionality be used to support the autonomy of higher-level causal explanations? (Section 4.4).

We conclude in Section 5. The appendix provides basics in causal modelling and information theory.

2 Proportionality, Specificity, and Their Relevance for Higher-Level Causation and Explanation

Many philosophers hold the view that the properties studied by special sciences such as biology, neuroscience, psychology, and sociology, cannot be fully explained by or identified with properties of (fundamental) physics. These philosophers can be roughly divided into two types: those who claim that the systems studied by the special sciences posses some kind of causal autonomy, and those who believe that they have at least some kind of explanatory autonomy. The causal autonomy camp consists mainly of emergentists and non-reductive physicalists (for example, Yablo [1992]; Shoemaker [2000]; List and Menzies [2009]; Zhong [2014]), while the explanatory autonomy camp consists mainly of a broad range of philosophers of science (for example, Shapiro [2010]; Woodward [2010], [2018]; Weslake [2013]). The former are committed to the metaphysical view that higher-level properties are autonomous in the sense that they have causal powers that are to some extent independent of the causal powers of their corresponding lower-level properties. The latter are only committed to autonomy in the sense that higher-level explanations are to some extent independent of the details of competing lower-level explanations.[3] The causal autonomy view is stronger than explanatory autonomy, as it is widely agreed that the existence of autonomous higher-level causes implies that there are autonomous higher-level explanations (but not vice versa). In this article, we discuss both forms of the autonomy view.

Regardless of whether the focus is on causation or explanation, supporters of the autonomy of higher levels often appeal to proportionality. Intuitively, the idea is that the cause should convey enough information about the conditions under which the effect occurs, and that it should not convey irrelevant information about conditions that do not make a difference for the effect (cf. Woodward [2010]). This can be illustrated with Yablo's ([1992]) pigeon example: When a pigeon is trained to peck at red objects, and is presented with a scarlet object, it will peck at said object. There are two ways of characterizing the cause: (1) The cause of the pecking was

[3] Although we refer to levels in this article, which is the convention in the debate, our arguments do not require any substantive idea of levels of organization. They are also compatible with a deflationary reading of levels, where levels are identified locally and case-by-case, based on composition or supervenience. For more, see (Eronen [2015]; Eronen and Brooks [2018]).

a scarlet object, or (2) the cause of the pecking was a red object. The first characterization conveys the information that the specific shade of red made the difference for the occurrence of the effect, but this is too much irrelevant information, as different shades of red all lead to the same outcome. The second characterization implies that it was the redness of the object that made the difference. Hence, it conveys enough but not too much information about the conditions that make the effect occur. For this reason, the second characterization captures a proportional cause, whereas the first one does not. More precisely, Woodward ([2010], p. 298) proposes to characterize proportionality as follows:

> **(P)** There is a pattern of systematic counterfactual 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 [...] 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.

Proportionality is closely related to the notion of specificity (cf. Woodward [2010]). Whereas proportionality concerns the extent to which the cause contains the relevant, and only the relevant, information about how the effect will change, specificity is about the extent to which the cause gives precise control over the effect. The standard example is DNA and the cellular machinery surrounding it, and their roles in the development of an organism (Davidson [2001]; Waters [2007]; Woodward [2010]). The activation of a segment of DNA results in a developmental change, but this also crucially involves the activity of the cellular machinery, such as transcriptor factors or RNA molecules. Thus, both the DNA and the other components (control proteins and RNA molecules) should be seen as causes for the developmental change. However, DNA seems to be a more specific cause: Intervening on RNA or transcriptor factors leads to general and wide-ranging changes in the developmental process, whereas intervening on the DNA leads to distinct changes that depend on what precisely was changed in the DNA. Thus, the DNA sequence seems to be causally specific for developmental changes, in contrast to RNA or transcriptor factors. However, characterizing specificity in a precise way that captures this intuitive idea has turned out to be challenging. Woodward suggests that the intuitive idea of specificity is very close to Lewis's ([2004]) notion of influence. The idea is that a cause C is

specific with regard to an effect E to the extent that there is a range of changes to C that result in different states of E. In this sense, a specific cause C can be seen as having influence over E. This captures one sense in which the DNA is a specific cause for developmental changes: There is a broad range of changes to the DNA that result in different developmental outcomes. Thus, specificity indicates causal influence as well as control.

However, as Woodward ([2010]) points out, there is another notion of specificity that also plays an important role in the scientific literature, namely, one-to-one specificity: For example, when a specific type of antibody only interacts with a specific type of antigen, and vice versa, there is a one-to-one correspondence between cause and effect variables. Similarly, in Mendelian inheritance, a gene A affects one trait (for example, colour) but not others, whereas another gene B affects another trait (for example, size), but not others. Here the idea is that a cause C is (maximally) specific with regard to an effect E if each value of C corresponds to exactly one value of E, and vice versa—in other words, the mapping between C's and E's values is a bijection, or closely approximates a bijection.

In this section we introduced the notions of proportionality and specificity as they have been used in the philosophical literature and sketched how these notions can be used to support the causal or the explanatory autonomy view. In the next section, we will draw from the literature on causal modelling and information theory in order to make these notions more precise and to provide tools for quantifying them. Once these tools are available, we will come back to the question of whether proportionality and specificity can be used to support higher-level causal or explanatory autonomy.

3 Specificity and Proportionality Quantified

In this section we introduce Griffiths *et al.*'s ([2015]) and Bourrat's ([2019]) proposals for measuring causal specificity and discuss their connection to proportionality. We also relate these proposals to Hope and Korb's ([2005]) causal power theory and provide reasons for choosing specific measures for the endeavour of this article. All of these measures combine basic concepts from causal modelling and information theory. (For a primer on causal modelling and information theory, see the appendix.) In some way or another all of them rely on a causal

version of conditional entropy (cf. Griffiths *et al.* [2015], p. 534):

Conditional Causal Entropy:

$$H(Y|\hat{X}) = -\sum_{x\in X} P(\hat{x}) \sum_{y\in Y} P(y|\hat{X}) \cdot log_2 P(y|\hat{X}) \tag{1}$$

$$H(\hat{X}|Y) = -\sum_{y\in Y} P(y) \sum_{x\in X} P(\hat{x}|y) \cdot log_2 P(\hat{x}|y) \tag{2}$$

$H(Y|\hat{X})$ can be interpreted as one's average degree of uncertainty[4] about Y's value if one would learn some X-value, and $H(\hat{X}|Y)$ as measuring the average of one's uncertainty about X's value if one would learn some Y-value. The difference to ordinary conditional entropy (see the appendix) is that conditional causal entropy is computed on the basis of the probability distribution one gets from intervening on X.[5] When intervening on X (denoted by $\hat{X}$), one makes X independent of its causes; one breaks the causal arrows into X. This guarantees that only information propagated over causal paths from X to Y is taken into account when computing $H(Y|\hat{X})$ and $H(\hat{X}|Y)$; additional information due to possible confounders is filtered out. Note that we do not use the standard notion of an intervention often used in the philosophical literature (for example, Woodward [2003]). Our interventions screen off the variable intervened on from its direct causes, but do not set it to a specific value. For this article we assume that $P(\hat{x})$ and $P(y)$ in conditional causal entropy are the probabilities from the pre-intervention distribution instead; they represent how frequently the different values of X and Y are instantiated.[6]

Based on this notion of conditional causal entropy, Griffiths *et al.* ([2015]) propose mutual causal information as a basis for measuring specificity:

Mutual Causal Information:

$$I(Y;\hat{X}) = H(Y) - H(Y|\hat{X}) \tag{3}$$

[4] Because interpreting entropy as a measure for the degree of one's uncertainty about a variable's value is intuitive, we use this interpretation to motivate the information-theoretic measures used in this article. Later on, however, we will discard the uncertainty interpretation and rather use the measures as indicators for how nicely causes and effects line up.

[5] Since we are interested in metaphysical issues about higher-level causal autonomy and non-pragmatic explanations in this article, we can ignore epistemic limitations and interpret probabilities as the true population-level frequencies.

[6] For alternative possibilities, see, for example, (Pocheville *et al.* [2017], Section 3).

Mutual causal information can be interpreted as measuring how much learning the cause's value after an intervention reduces the uncertainty about the value of the effect, and vice versa. The 'and vice versa' part is justified by the fact that mutual causal information, just like ordinary mutual information, is symmetric simply because one and the same post-intervention distribution is used for computing both $I(Y;\hat{X})$ and $I(\hat{X};Y)$. In some sense, however, mutual causal information adds an asymmetric element into the mix: While manipulating the cause will typically provide a certain amount of information about the effect (that is, $I(Y;\hat{X}) > 0$), wiggling the effect will tell us nothing about the cause (that is, $I(X;\hat{Y}) = 0$).

Mutual causal information can be used as a basis for measuring different things. To get an overview about what it can be used for, a brief excursion to a recent article by Bourrat will be helpful. Bourrat ([2019], p. 4) distinguishes between the following kinds of specificity[7]:

> Specificity of the Cause for the Effect: To what extent is a value of E caused by values of C that do not cause other values of E? In other words, for each value of C, to what extent does this value determine a single value of E?
>
> Specificity of the Effect for the Cause: To what extent does a value of C cause values of E that are different from values of E caused by other values of C? In other words, for each value of E, to what extent is this value determined by a single value of C?
>
> One-to-One Specificity: The specificity of the cause for the effect as well as for the specificity of the effect for the cause. In other words, to what extent is there a one-to-one mapping from the values of the cause C to the values of the effect E?

According to Bourrat ([2019]), the basic version of mutual causal information suggested by Griffiths *et al.* ([2015]) (see the definition of mutual causal information, above) does not measure one-to-one specificity, but rather the range of causal influence. (Since this notion will play no role for our endeavour in this article, we ignore it from now on.) Bourrat proposes to measure one-to-one specificity by using the variation of causal information instead:

[7] We use the labels 'specificity of the cause for the effect' and 'specificity of the effect for the cause' differently than Bourrat ([2019]). We switched them because this terminology fits—in our view—nicer to the intuition that a more specific cause (for the effect) allows for more (or more fine-grained) control of the effect, while a more specific effect (for the cause) provides more information about the cause (cf. Woodward [2010]).

Variation of Causal Information:

$$VI(Y;\hat{X}) = H(Y|\hat{X}) + H(\hat{X}|Y) \tag{4}$$

Note that the closer $VI(Y;\hat{X})$ is to zero, the closer the pattern of causal dependencies of the effect Y's values on the cause X's values is to a bijection (cf. Bourrat [2019]). In order to compare causal relationships between variables with different numbers of possible values, Bourrat suggests to normalize the variation of causal information to the interval $[0, 1]$ based on the entropy of Y and X:

Normalized Variation of Causal Information:

$$NVI(Y;\hat{X}) = \frac{VI(Y;\hat{X})}{H(Y,\hat{X})} \tag{5}$$

However, as already suggested by Griffiths *et al.* ([2015], Section 2), there are different ways that the entropy of Y and X, in addition to mutual causal information, can be relevant for answering questions concerning specificity. In particular, mutual causal information can be compared to the entropy of the effect, the entropy of the cause, or their joint entropy. In other words, mutual causal information can be normalized to the interval $[0, 1]$ in at least three different ways that are all relevant for issues concerning specificity:

Normalized Mutual Causal Information:

$$SPEC_c(Y;\hat{X}) = \frac{I(Y;\hat{X})}{H(Y)} \tag{6}$$

$$SPEC_e(Y;\hat{X}) = \frac{I(Y;\hat{X})}{H(\hat{X})} \tag{7}$$

$$PROP(Y;\hat{X}) = \frac{I(Y;\hat{X})}{H(Y,\hat{X})} \tag{8}$$

Note that because

$$1 - NVI(Y;\hat{X}) = \frac{H(Y,\hat{X}) - H(Y|\hat{X}) - H(\hat{X}|Y)}{H(Y,\hat{X})} = \frac{I(Y;\hat{X})}{H(Y,\hat{X})} = PROP(Y;\hat{X})$$

holds (see Figure 3 in the appendix for illustration), $PROP(Y;\hat{X})$ turns out to be the complement of the normalized variation of causal information (normalized variation of causal information). Hence, it follows that $PROP(Y;\hat{X})$ can actually be used to measure one-to-one specificity if and only if $NVI(Y;\hat{X})$ can, which somewhat undermines Bourrat's ([2019]) claim that (normalized) variation of information can be used to measure one-to-one specificity while a measure based on mutual causal information cannot.

So far, we have introduced four different measures for three different kinds of specificity. However, we are interested in quantifying specificity and proportionality. So do the measures introduced also tell us something about proportionality? Let us first recall that the closer $NVI(Y;\hat{X})$ is to zero, the closer the pattern of dependencies between values of the cause X and the effect Y is to a bijection. In light of the finding above, this is also the case the closer $PROP(Y;\hat{X})$ is to one. If we now compare $NVI(Y;\hat{X})$ (or its complement $PROP(Y;\hat{X})$) to the characterization of proportionality **(P)** in Section 2, it is clear that $NVI(Y;\hat{X})$ (and, thus, also $PROP(Y;\hat{X})$) is not only a measure for one-to-one specificity, but also a good candidate for measuring proportionality (thus the label '*PROP*' in conditional mutual information). Recall that $NVI(Y;\hat{X})$ is defined as the sum of $H(Y|\hat{X})$ and $H(\hat{X}|Y)$. Minimizing $H(Y|\hat{X})$ amounts to maximizing informativeness about the conditions under which alternative states of the effect are realized—condition (a) of **(P)**—and minimizing $H(\hat{X}|Y)$ amounts to favoring causes not characterized in such a way that alternative states of the cause fail to be associated with changes in the effect—condition (b) of **(P)**. Consequently we propose to use $NVI(Y;\hat{X})$ (and, thus, also $PROP(Y;\hat{X})$) as a measure for proportionality.[8]

Before we turn to the remaining two kinds of normalized mutual causal information, let us briefly compare the proposal to measure proportionality via $NVI(Y;\hat{X})$ to what Pocheville *et al.* ([2017]) say about proportionality. They too suggest that proportionality is closely related to specificity. In particular, they suggest that proportionality of the cause X with respect to the effect Y can be maximized by maximizing $I(Y;\hat{X})$ and minimizing $H(X|Y)$. The idea is that the most proportional causes X are those that are as informative as possible while, at the same time, each effect value corresponds to as few as possible cause values. In other words, a cause

[8] Because the definition of $NVI(Y;\hat{X})$ as the sum of $H(Y|\hat{X})$ and $H(\hat{X}|Y)$ so nicely corresponds to conditions (a) and (b) in **(P)**, we will most of the time use $NVI(Y;\hat{X})$ rather than its complement $PROP(Y;\hat{X})$.

is the more proportional to its effect the closer the pattern of dependencies between cause and effect values resembles a bijection. So the goal is a pattern of dependencies that is as close as possible to a surjection as well as to an injection.[9] On the basis of Figure 3 (see the appendix), it is easy to see that this nicely fits the idea of using $PROP(Y;\hat{X})$ as a measure for proportionality: Maximizing $I(Y;\hat{X})$ while, at the same time, minimizing $H(\hat{X}|Y)$ amounts to shifting $I(Y;\hat{X})$ as close as possible to $H(Y,\hat{X})$. How close $I(Y;\hat{X})$ comes to $H(Y,\hat{X})$, however, is exactly what $PROP(Y;\hat{X})$ and, thus, also its complement $NVI(Y;\hat{X})$ measures.

Let us now turn to the remaining two kinds of normalized causal information in conditional mutual information. We propose that $SPEC_c(Y;\hat{X})$ measures the specificity of the cause for the effect (or more shortly, the specificity of the cause) and that $SPEC_e(Y;\hat{X})$ measures the specificity of the effect for the cause (or the specificity of the effect). This can be further motivated by a brief look at Figure 1. For simplicity, we assume that the probabilities are equally distributed over the values of the cause variable Y in (a)–(c). Intuitively, (a) should be more proportional than (b) and (c) because the pattern of dependencies is closer to a bijection. Our proposed measure of proportionality captures this intuition: For (a), $NVI(Y;\hat{X}) = \frac{0+0}{2} = 0$, while for (b) $NVI(Y;\hat{X}) = \frac{0+1}{2} = 0.5$ and for (c) $NVI(Y;\hat{X}) = \frac{1+0}{2} = 0.5$. Since in (a) and (b) the cause fully determines the effect, but not in (c), we would expect the cause in (a) and (b) to be more specific for the effect than in (c). This also fits what our measure for the specificity of the cause yields: For (a), $SPEC_c(Y;\hat{X})^{\frac{2-0}{2}} = 1$, while for (b) $SPEC_c(Y;\hat{X}) = \frac{1-0}{1} = 1$ and for (c) $SPEC_c(Y;\hat{X}) = \frac{2-1}{2} = 0.5$. Finally, the specificity of the effect for the cause should be maximal in (a) and (c), but not in (b). Also this intuition is captured by our measures: the measure for the specificity of the effect returns $SPEC_e(Y;\hat{X})^{\frac{2-0}{2}} = 1$ for (a), $SPEC_e(Y;\hat{X})^{\frac{1-0}{2}} = 0.5$ for (b), and $SPEC_e(Y;\hat{X})^{\frac{2-1}{1}} = 1$ for (c).

Recall from Section 2 that Woodward ([2010]) argued for specificity as closely related to causal influence, power, and control. Can one of our measures for specificity be related to these notions in a similar way? The basic idea here is that the more specific a cause is for its effect, the more causal influence, power, or control one can have over the effect by being able to manipulate the cause (and vice versa). We take it that intuitions like these motivated Hope

[9] This is, again, the basic idea also underlying Woodward's ([2010]) one-to-one specificity.

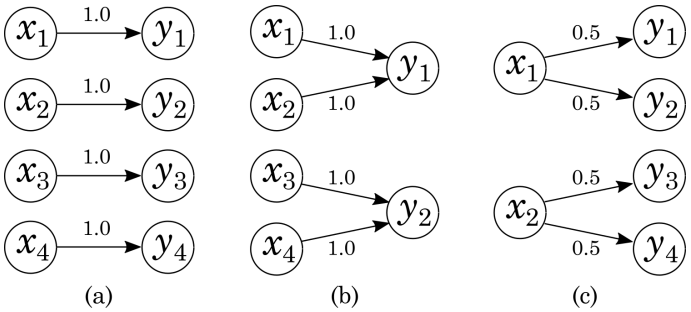

Figure 1. Dependence patterns between the cause X and the effect Y. Arrows stand for dependencies between values; numbers indicate conditional probabilities.

and Korb ([2005]) to propose that $I(Y;\hat{X})$ measures the strength of X's causal influence on Y or its causal power with respect to Y. As we saw above, however, one can cover different kinds of specificity depending on how $I(Y;\hat{X})$ is normalized. Since causal influence, power, and control are always directed from the cause to the effect, we propose that these magnitudes correspond to the specificity of the cause $SPEC_c(Y;\hat{X})$. This move can be motivated by another brief look at Figure 1: In (a) and (b) we can fully determine the effect via manipulating the cause, so the cause should have maximal causal influence, power, or control over the effect. In (c), however, we cannot fully control the effect via manipulating the cause. As our discussion of the three kinds of normalized mutual causal information above shows, $SPEC_c(Y;\hat{X})$ is the only candidate fitting these intuitions.

Let us briefly recapitulate and see where we stand. In Section 2 we saw that proportionality and, to a lesser extent, the related notion of specificity is often used to support the view that higher-level causes or explanations can be autonomous. The main goal of this article is to see what we can say about attempts to support higher-level causal or explanatory autonomy based on these notions from a causal modelling angle that implements information-theoretic concepts in order to make these notions precise. This gives us a tool for investigation that is more fine-grained than the arguments typically used in the philosophical literature. To this end, we introduced and discussed measures for different kinds of specificity in this section. It turned out that one (that is, *NVI* or its complement *PROP*) is capable of quantifying proportionality, while another one (that is, $SPEC_c$) nicely captures the specificity of the cause as well as causal power or influence. Next, we will use the tools we now have at hand to shed new light on the issue of higher-level causal and explanatory autonomy. In particular, we will use them to

answer the four main questions raised at the end of Section 1.

4 On the Limits of Higher-Level Causation and Explanation

In this section, we will apply the measures for the specificity of the cause (or causal power) and proportionality introduced in Section 3 to the philosophical debate about higher-level causation and explanation sketched in Section 2. To this end, we need a simple representative model reflecting the basic assumptions regarding higher-level causation.

First of all, supporters of both the causal and the explanatory autonomy view typically agree that higher-level properties supervene on lower-level properties and that higher-level properties are often multiply realizable by lower-level properties (Eronen [2011]).[10] A set of properties P (in this case, a set of higher-level properties) supervenes on a set of properties Q (in this case, a set of lower-level properties) if and only if there can be no changes in P-properties without there being some changes in Q-properties. Multiple realizability entails that a higher-level property can be realized by several distinct lower-level properties, and therefore is not identical to any lower-level property.

Let us assume that A and B are two variables that stand for distinct higher-level properties (for example, mental states) and that X and Y are two variables that stand for distinct lower-level properties (for example, neural states). In addition, we assume that A supervenes on X, that B supervenes on Y, and that both A and B are multiply realizable by their supervenience bases X and Y, respectively. The supervenience assumption implies that (i) each change in values of one of the higher-level variables corresponds to a change in the probability distribution over the corresponding lower-level variable and (ii) that for each value of one of the lower-level variables there will be a value of the corresponding higher-level variable such that the conditional probability of that value given the value of the lower-level variable will be one (cf. Sober [1999]; Gebharter [2017a]). Multiple realizability of the higher-level variables, on the other hand, implies that some values of each higher-level variable are such that if the higher-level variable takes one of these values, then no value of the corresponding lower-level variable gets a probability of one. More formally:

[10] Recall from Section 1 that we are not interested in whether higher-level causal explanations can be pragmatically superior in this article.

$$\forall x \forall x' \exists a : \text{If } x \neq x' \text{ then } P(a|x) \neq P(a|x') \tag{9}$$

$$\forall y \forall y' \exists b : \text{If } y \neq y' \text{ then } P(b|y) \neq P(b|y') \tag{10}$$

$$\forall x \exists a : P(a|x) = 1 \tag{11}$$

$$\forall y \exists b : P(b|y) = 1 \tag{12}$$

$$\exists a \forall x : P(x|a) < 1 \tag{13}$$

$$\exists b \forall y : P(y|b) < 1 \tag{14}$$

Since we want to use the model to evaluate different views about higher- and inter-level causation and explanation, we add all logical possibilities of how the variables from the set $\{A, X\}$ could be causes of the variables from the set $\{B, Y\}$. To this end, we assume that both A and X are directly causally relevant for both B and Y. If we then represent causal dependencies by solid arrows and supervenience/multiple realizability by dashed arrows, we end up with the structure in Figure 2.[11]

Before we apply the measures from Section 3 to the model, some observations are in order. Due to Equations 9 and 10, A and X as well as Y and B are probabilistically dependent, and due to Equations 11 and 12 both higher-level variables are fully determined by their corresponding lower-level variables. Because of this, probability flow between A and Y is only possible due to changes in X, probability flow between X and B is only possible due to changes in Y, and probability flow between A and B is only possible if both X and Y are allowed to change values. Finally, since the higher-level variables A and B are multiply realizable (Equations 13 and 14), probability flow between X and Y is—at least in principle—possible without changes in A- and B-values. These observations taken together imply that the black arrows in Figure 2 form a minimal Bayesian network; they mark the paths over which probabilistic influence spreads between variables. This also nicely fits the fact that lower-level causation is, contrary to higher- and inter-level causation, less controversial and more or less commonly accepted in the debate (see, for example, Kim [2005]).

[11] We assume that supervenience/multiple realizability arrows technically work like causal arrows in an ordinary causally interpreted Bayesian network. For an argument, see (Gebharter [2017a]).

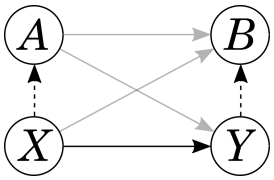

Figure 2

In order to apply the measures from Section 3 to the cause variables A and X, we must be able to intervene on A and B. How exactly interventions into systems featuring variables standing in other than causal relations work is still somewhat controversial (see, for example, Baumgartner [2013], Baumgartner and Gebharter [2016], Eronen and Brooks [2014], Gebharter [2017a]; [forthcoming], Woodward [2015]), but for the present purposes we can bracket the debate on how interventions work in such systems. It suffices to draw on what is commonly accepted: A supervening variable's value cannot be changed without changing the subvenient variable's value. Since X is an exogenous variable, interventions on X are unproblematic. Intervening on the higher-level cause A also seems straightforward. As no causes of A are captured by our model, we do not have to delete any causal arrows into A. And since higher-level variables can change their values only if the variables they supervene on do so as well, we can leave the arrow from X to A intact. This guarantees that intervening on A can lead to a change of the probability distribution over its supervenience base X as well as of the distributions over the effect variables B and Y. Interestingly, this treatment of interventions implies for this particular model that the post-intervention distributions used for computing the measures will be identical to the pre-intervention distributions.

Next, let us use our model in order to shed some new light on issues within the philosophical debate. With the measure for proportionality from Section 3 at hand, one question that comes quite naturally is whether higher-level causes can actually be more proportional with respect to their effects than their lower-level realizers, as supporters of the causal or explanatory autonomy view claim. Or does the presence of supervenience relationships already exclude this possibility? Does supervenience (conjoined with multiple realizability) perhaps render higher-level causes always more proportional with respect to their effects than their lower-level competitors? These questions will be answered in Section 4.1.

4.1 On proportionality as an empirical matter

The first interesting thing we can say about higher- versus lower-level causation and explanation is in line with List and Menzies ([2009]) (see also Menzies and List [2010]): Whether the higher-level cause or its lower-level supervenience base is more proportional with respect to the higher-level effect (or its lower-level supervenience base) cannot be decided *a priori*. Which candidate cause is more proportional is—even in non-pragmatic contexts—an empirical fact and each one of the causal arrows into one of the effect variables has the potential to trump the other causal arrow with which it competes. This can be verified by calculating the normalized variation of information for the exemplary probability distribution satisfying Equations 9–14 specified in Table 1. From (a), we get:

$$NVI(B;\hat{A}) = \frac{0+0}{1} = 0 \quad < \quad NVI(B;\hat{X}) = \frac{0+1}{2} = 0.5,$$
$$NVI(Y;\hat{A}) = \frac{1+0}{2} = 0.5 \quad > \quad NVI(Y;\hat{X}) = \frac{0+0}{2} = 0.$$

However, if we replace A's parameters in (a) by those in (b), we get[12]:

$$NVI(B;\hat{A}) = \frac{0.689+0.5}{1.5} = 0.793 \quad > \quad NVI(B;\hat{X}) = \frac{0+1}{2} = 0.5.$$

And if we replace Y's parameters in (a) by those in (c), we get:

$$NVI(Y;\hat{A}) = \frac{1+0}{2} = 0.5 \quad < \quad NVI(Y;\hat{X}) = \frac{1+1}{3} = 0.\dot{6}.$$

Thus, if (a) was the actual probability distribution over our system, the higher-level cause A would be more proportional than the lower-level cause X with respect to the higher-level effect B, while the lower-level cause X would be more proportional than the higher-level cause A with respect to the lower-level effect Y. If, however, we replace A's parameters in (a) by those in (b), then the lower-level cause X would trump the higher-level cause A with respect to the higher-level effect B. And if we replace Y's parameters in (a) by those in (c), then the higher-level cause A would trump the lower-level cause X with respect to the lower-level effect Y.

[12] Numbers are rounded to three digits.

$P(a_1\|x_1) = 1$	$P(b_1\|y_1) = 1$	$P(a_1\|x_1) = 1$	$P(y_1\|x_1) = 0.5$
$P(a_1\|x_2) = 1$	$P(b_1\|y_2) = 1$	$P(a_1\|x_2) = 1$	$P(y_2\|x_1) = 0.5$
$P(a_2\|x_3) = 1$	$P(b_2\|y_3) = 1$	$P(a_1\|x_3) = 1$	$P(y_1\|x_2) = 0.5$
$P(a_2\|x_4) = 1$	$P(b_2\|y_4) = 1$	$P(a_2\|x_4) = 1$	$P(y_2\|x_2) = 0.5$
			$P(y_3\|x_3) = 0.5$
$P(x_1) = 0.25$	$P(y_1\|x_1) = 1$		$P(y_4\|x_3) = 0.5$
$P(x_2) = 0.25$	$P(y_2\|x_2) = 1$		$P(y_3\|x_4) = 0.5$
$P(x_3) = 0.25$	$P(y_3\|x_3) = 1$		$P(y_4\|x_4) = 0.5$
$P(x_4) = 0.25$	$P(y_4\|x_4) = 1$		
(a)		(b)	(c)

Table 1. Exemplary probability distribution (a) and alternative parameters for A (b) and Y (c).

Now we know that whether a specific higher-level cause trumps its lower-level supervenience base is a purely empirical matter. The next natural question would be how common proportional higher-level causes are. Are almost all higher-level causes more proportional with respect to their higher-level effects than their lower-level realizers, as supporters of the causal autonomy view often assume? Is it even possible to come up with a philosophical argument about how frequently higher-level causes will be more proportional than their lower-level competitors? We turn to these questions in Section 4.2.

4.2 On how common proportional higher-level causes are

Supporters of the autonomy of higher-level causes such as Menzies and List ([2010]) often claim that higher-level causes of higher-level effects (for example, mental causes) will almost always trump their lower-level counterparts (for example, neural realizers) because most higher-level causal relations are realization-insensitive: Details about how the higher-level cause is realized make no difference for its particular effect. Though Menzies and List formulate realization-insensitivity in terms of possible worlds—we bracket the details here—there is a straightforward translation of the basic idea into our probabilistic framework:

> Realization-(In)sensitivity: $A \longrightarrow B$ is realization-sensitive if and only if there are some A-values a and X-values x realizing a such that $P(b|\hat{a}) \neq P(b|\hat{x})$ holds for some B-values b. $A \longrightarrow B$ is realization-insensitive if and only if it is not realization-sensitive.

More informally speaking, realization-insensitivity requires that bringing about any value a of

the higher-level cause A by intervention has the same probabilistic effect on the higher-level effect B as bringing about any of a's lower-level realizers x by intervention. This definition is also largely in line with what Woodward ([2008], p. 241) calls a realization independent dependency relationship, which is 'a relationship that both involves a dependency between the upper level variables (different values of $[A]$, produced by interventions, map into different values of $[B]$) and that is realization independent in the sense that it continues to stably hold for a range of different realizers of these values of $[A]$ and $[B]$'.

On the basis of realization-(in)sensitivity, we can now infer that in cases where $A \longrightarrow B$ is realization-insensitive, the higher-level cause A is at least as proportional with respect to B as its lower-level competitor X: Because we assumed realization-insensitivity, $H(B|\hat{A})$ equals $H(B|\hat{X})$, and because A's values are multiply realizable by X's values, $H(\hat{A}|B)$ is smaller than $H(\hat{X}|B)$. Finally, $H(B|\hat{A})$ and $H(B|\hat{X})$ cannot be greater than $H(B)$ by definition. From these observations

$$NVI(B;\hat{A}) = \frac{H(B|\hat{A}) + H(\hat{A}|B)}{H(\hat{A}|B) + H(B)} \leq NVI(B;\hat{X}) = \frac{H(B|\hat{X}) + H(\hat{X}|B)}{H(\hat{X}|B) + H(B)} \tag{15}$$

follows.

Though this result supports List and Menzies' ([2009]) claim that realization-insensitivity guarantees that higher-level causes are at least as proportional with respect to their higher-level effects as their supervenience bases are, it does not yet help their overall project to argue that most higher-level causes can be expected to be autonomous. The crucial question for this project is how much realization-insensitivity there actually is in the world. In our view, there is much less realization-insensitivity than Menzies and List ([2010]) suggest, as real-life causal structures are typically much messier than simple philosophical examples. More specifically, effects typically have a multitude of causes that interact with each other in complex ways, which can easily destroy realization-insensitivity. All that is required is that one of the values of the lower-level cause X realizing one of the values of the higher-level cause A has a slightly different probabilistic effect on one of the values of the higher-level effect B than another value of X realizing the same value of A has. For example, if mental state a_1 causes another mental state b_1 with a probability of 0.7 when it is realized by x_1, but with a probability of 0.69 when

$P(a_1\|x_1) = 1$	$P(b_1\|y_1) = 1$	$P(y_1\|x_1) = 0.4$	$P(y_1\|x_3) = 0.2$
$P(a_1\|x_2) = 1$	$P(b_1\|y_2) = 1$	$P(y_2\|x_1) = 0.3$	$P(y_2\|x_3) = 0.2$
$P(a_2\|x_3) = 1$	$P(b_2\|y_3) = 1$	$P(y_3\|x_1) = 0.2$	$P(y_3\|x_3) = 0.3$
$P(a_2\|x_4) = 1$	$P(b_2\|y_4) = 1$	$P(y_4\|x_1) = 0.1$	$P(y_4\|x_3) = 0.3$
		$P(y_1\|x_2) = 0.3$	$P(y_1\|x_4) = 0.2$
$P(x_1) = 0.25$		$P(y_2\|x_2) = 0.3$	$P(y_2\|x_4) = 0.2$
$P(x_2) = 0.25$		$P(y_3\|x_2) = 0.2$	$P(y_3\|x_4) = 0.3$
$P(x_3) = 0.25$		$P(y_4\|x_2) = 0.2$	$P(y_4\|x_4) = 0.3$
$P(x_4) = 0.25$			

Table 2

it is realized by x_2, realization-insensitivity is violated. And even if both $P(b_1|x_1)$ and $P(b_1|x_2)$ equal 0.7, there can be another cause C of B's supervenience base Y such that C's taking some of its values will change one of these conditional probabilities at least slightly. Considering the complexity of the brain and the neural realizers of mental states, such cases are likely to be very common.

Interestingly, it also turns out that realization-insensitivity is not necessary for a higher-level cause to be more proportional than its lower-level supervenience base. The distribution specified in Table 2, for example, violates realization-insensitivity and still renders the higher-level cause A slightly more proportional with respect to its higher-level effect B than the lower-level cause X since

$$NVI(B;\hat{A}) = \frac{0.953 + 0.954}{1.953} = 0.977 \quad < \quad NVI(B;\hat{X}) = \frac{0.949 + 1.950}{2.949} = 0.983 \qquad (16)$$

holds.[13]

Beyond examples, is it possible to determine more generally the conditions under which the higher-level cause trumps the lower-level cause? Let us have another look at Equation 15. One of the assumptions used to derive Equation **??** was that $H(B|\hat{A})$ equals $H(B|\hat{X})$, which is the case when realization-insensitivity is assumed, but in the presence of realization-insensitivity violations, we cannot rely on this equality. However, we know from the fact that the black arrows in Figure 2 form a Bayesian network that $H(B|\hat{A})$ has to be greater than (or equal to) $H(B|\hat{X})$. Thus, for Equation 15 to hold, the difference between $H(B|\hat{A})$ and $H(B|\hat{X})$ must be

[13] Numbers are rounded to three digits.

compensated by a greater difference between $H(\hat{A}|B)$ and $H(\hat{X}|B)$.[14] Whether this is the case hinges, in the end, on the numerical details of the specific case and, thus, on the actual difference between $H(B|\hat{A})$ and $H(B|\hat{X})$ and the value of $H(B)$. For this reason, it seems that one cannot provide a philosophical argument for how often higher-level causes will actually trump their lower-level supervenience bases; the question must be answered by empirical investigations.

Let us briefly halt to see where we stand. Supporters of the causal autonomy view typically claim that higher-level causes are often or even almost always more proportional than their lower-level competitors. In Section 4.1 we could verify the claim that higher-level causes can be more proportional, but that the same also goes for their lower-level realizers. We then saw in Section 4.2 that we can say nothing about how common more proportional higher-level causes are. In fact the kind of realization-insensitivity required to guarantee higher-level proportionality is highly fragile and can be easily destroyed in real-world causal settings. This renders philosophical arguments for the view that most higher-level causes are proportional at least dubious. However, supporters of the causal autonomy view might still see our results as a kind of evidence in favour of their view: Since it is an empirical matter whether higher-level causes are more proportional, it is still conceptually possible that they are. Thus, higher-level causal autonomy is conceptually possible and cannot be ruled out by philosophical arguments (such as the exclusion argument) either. In Section 4.3 we will put this line of reasoning to the test.

4.3 On the autonomy of higher-level causes

Supporters of the causal autonomy view, such as Menzies and List ([2010]), take the result that proportionality is an empirical matter to support the view that higher-level causation cannot be excluded *a priori*. Menzies and List (see also List and Menzies [2009]) argue that causation should be constrained by proportionality—a strategy also many other philosophers (for example, Yablo [1992]) endorse. Constraining causation by proportionality means that only the causal variable most proportional for a specific effect variable is accepted as a cause of the latter, while all competing causal variables are denied the status of cause for this particular effect

[14] Recall that $H(\hat{A}|B) < H(\hat{X}|B)$ holds due to multiple realizability.

variable.[15] They then go on and argue that since higher-level variables that are more proportional with respect to their purported higher-level effects than their lower-level supervenience bases are at least conceptually possible, higher-level causation is possible as well. As lower-level variables would not satisfy the proportionality constraint in such cases, the higher-level variables would have causal powers their lower-level competitors lack. They would, hence, be causally autonomous.

This argument stands and falls with the proportionality constraint, that is the question of whether for C to be a cause of E it is necessary that C is more proportional with respect to E than any other variable C' competing with C. In our view (and in line with Woodward [2010];8), constraining causation in such a way is problematic for several reasons. One reason is that most effects have several same-level causes working together to bring them about. A proportionality constraint on causation would select only the most proportional of these factors as a cause and deny the causal status of all the other factors. Another reason is that when proportionality is analysed in a causal modelling framework, it seems unavoidable that it comes in degrees. As the example in Equation 16 shows, higher-level variables might turn out to be only slightly more proportional than lower-level variables, where this difference in proportionality can be made arbitrarily small. In such cases the higher-level variable and the lower-level variable are almost equally proportional and it seems arbitrary to grant the status of a cause to the former but not to the latter.

In contrast to Menzies and List ([2010]), we also believe that proportionality is a bad indicator for the causal powers of a variable. A brief look at our measure for proportionality (normalized variation of causal information) shows that proportionality is symmetric. It measures how close the pattern of dependencies of the effect's values on the cause's values is to a bijection. It not only reflects what can be learned about the value of the effect by learning the value of the cause, but also what can be learned about the value of the cause by learning the value of the effect. This stands in contrast to what one expects from a measure of causal power:

[15] Note that this is our reconstruction of the proportionality constraint. Menzies and List ([2010]) use a simplified difference-making theory, according to which causes are necessarily proportional. In their simplified theory, proportionality is a yes-or-no matter. To represent their results about causal upward and downward exclusion (see also List and Menzies [2009]) within a richer account of causation like the one we use in this article, proportionality must be understood as a matter of degree and our interpretation of the proportionality constraint seems unavoidable.

It should be asymmetric, and reflect the extent to which a variable has influence over another (effect) variable.

This is exactly the kind of information the measure $SPEC_c$ is designed to provide. We can therefore rephrase the crucial question of whether higher-level causation is possible: Whether a higher-level variable has causal powers over and above its lower-level competitor does not depend on the question of which variable is more proportional with respect to the purported effect, but on the question of whether the higher-level cause variable is more specific. And here the answer is clear: Although a higher-level variable A can be more proportional than its lower-level competitor X with respect to the higher-level effect B (as we have seen in Section 4.1), A can never trump X when it comes to the specificity of the cause. This follows from the fact that probability flow between A and B is only possible over the black arrows in Figure 2, and that with each arrow on the path there is the possibility of information loss. Hence, $SPEC_c(B;\hat{A}) \leq SPEC_c(B;\hat{X})$. In words: A cannot have any causal powers with respect to B that X does not have—control over the higher-level variable A cannot give one more causal influence on B than control over the lower-level variable X.[16]

One can also put this as follows: for the higher-level variable A to have causal powers with respect to B in addition to the ones possessed by its supervenience base X, A has to have a positive specificity value with respect to B conditional on X. A measure for conditional specificity of the cause (or causal power) can be introduced similarly to the unconditional measure (conditional mutual information):

Conditional Mutual Information:

$$I(Y;X|Z) = H(Y|Z) - H(Y|X,Z) \tag{17}$$

Conditional Mutual Causal Information:

$$I(Y;\hat{X}|\hat{Z}) = H(Y|\hat{Z}) - H(Y|\hat{X},\hat{Z}) \tag{18}$$

[16] Note that the same argument holds if B is replaced by Y and, thus, that A also cannot have any causal powers with respect to Y that X does not have.

Normalized Conditional Mutual Causal Information:

$$SPEC_c(Y; \hat{X}|\hat{Z}) = \frac{I(Y; \hat{X}|\hat{Z})}{H(Y|\hat{Z})} \tag{19}$$

It follows from the fact that the black arrows in Figure 2 form a Bayesian network that A and B are screened off by X. This implies that $I(B; \hat{A}|\hat{X}) = 0$ (cf. Ay and Polani [2008]), which implies that $SPEC_c(B; \hat{A}|\hat{X}) = 0$. $SPEC_c(B; \hat{X}|\hat{A}) > 0$, on the other hand, is possible. This means that the higher-level cause (for example, a mental state) can give no new causal information with respect to the higher-level effect (for example, another mental state) that goes beyond the information that the lower-level cause (for example, the neural realizer) provides, but not the other way round.

To summarize, if the interest is in comparing the causal powers or causal influence of higher- versus lower-level properties, proportionality is not a good measure. Rather, the specificity of the cause seems to be more appropriate for this purpose, and based on this measure, higher-level causes cannot have any causal power or influence in addition to the causal power or influence their lower-level realizers have. Where do we stand now? While the results from Section 4.1 and Section 4.2 took a good portion of wind out of the supporters of the causal autonomy view's sails, the results from Section 4.3 finally sinks the ship. However, one might quite naturally wonder whether this result also has some bearing on the logically weaker (see Section 2) explanatory autonomy view. In Section 4.4 we consider whether the explanatory autonomy view is equally threatened by the information-theoretic approach to proportionality and specificity.

4.4 On the autonomy of higher-level causal explanations

The question of whether higher-level explanations bear some kind of autonomy with respect to corresponding lower-level explanations is not uniquely determined by the results we produced so far. How higher-level explanations fare versus lower-level explanations depends, in the end, on what exactly one wants from a good explanation.[17] As mentioned in the introduction, here we focus on one widely held view, according to which explanatory power is a matter of

[17] Recall that we are only interested in the non-pragmatic superiority of higher-level causal explanations.

$P(a_1\|x_1) = 1$	$P(b_1\|y_1) = 1$	$P(y_1\|x_1) = 0.4$	$P(y_1\|x_3) = 0.1$
$P(a_1\|x_2) = 1$	$P(b_1\|y_2) = 1$	$P(y_2\|x_1) = 0.1$	$P(y_2\|x_3) = 0.1$
$P(a_2\|x_3) = 1$	$P(b_2\|y_3) = 1$	$P(y_3\|x_1) = 0.1$	$P(y_3\|x_3) = 0.4$
$P(a_2\|x_4) = 1$	$P(b_2\|y_4) = 1$	$P(y_4\|x_1) = 0.4$	$P(y_4\|x_3) = 0.4$
		$P(y_1\|x_2) = 0.4$	$P(y_1\|x_4) = 0.1$
$P(x_1) = 0.25$		$P(y_2\|x_2) = 0.4$	$P(y_2\|x_4) = 0.1$
$P(x_2) = 0.25$		$P(y_3\|x_2) = 0.1$	$P(y_3\|x_4) = 0.4$
$P(x_3) = 0.25$		$P(y_4\|x_2) = 0.1$	$P(y_4\|x_4) = 0.4$
$P(x_4) = 0.25$			

Table 3

providing answers to what-if-things-had-been-different questions or w-questions (Hitchcock and Woodward [2003]; Woodward [2003]; Woodward and Hitchcock [2003]). According to this view, an explanation is better the more it captures the dependencies of different values of the effect variable on different values of the cause variable.[18] In other words, an explanation is the better the more information the cited cause variable provides about the effect variable. Since this is exactly what $SPEC_c$ measures, an account of explanation aiming at answering w-questions is committed to the view that citing the cause with the highest $SPEC_c$ provides the best explanation for a particular effect.

But what about proportionality? Can it help in establishing explanatory autonomy of higher-level causal relations? From the perspective of an account of explanation whose goal consists in answering w-questions, proportionality would establish the autonomy of higher-level explanations if the most proportional cause for a particular effect would provide the most answers to w-questions involving this effect. However, in effect, this amounts to being the most specific cause for this effect. Hence, the question of whether proportionality is sufficient for higher-level explanatory autonomy reduces to the question of whether the most proportional cause is, at the same time, necessarily the most specific cause. That this is not the case can be verified by having a brief look at the exemplary probability distribution specified in Table 3. From this

[18] Strictly speaking, this is only one dimension of Hitchcock and Woodward's account. Another dimension consists in providing information about what would happen if background circumstances were different. Since the general results we present in this subsection are not sensitive to different background circumstances, we will ignore that dimension.

distribution we get

$$NVI(B;\hat{A}) = \frac{0.828 + 0.844}{1.828} = 0.915 \quad < \quad NVI(B;\hat{X}) = \frac{0.791 + 1.808}{2.791} = 0.931,$$

and

$$SPEC_c(B;\hat{A}) = \frac{0.984 - 0.828}{0.984} = 0.158 \quad < \quad SPEC_c(B;\hat{X}) = \frac{0.984 - 0.791}{0.984} = 0.195,$$

showing that even if the higher-level cause A is more proportional with respect to B than its lower-level competitor X, A can still be less specific. Thus, a cause's proportionality seems, in the end, to be irrelevant for whether this cause provides the best explanation in the sense of providing the most informative answers to w-questions.

Instead, only the specificity of the cause seems to be relevant for how good a cause is in answering w-questions about a particular effect. As we already know from Section 4.3 that lower-level causes are always at least as specific as their higher-level counterparts (that is, $SPEC_c(B;\hat{A}) \leq SPEC_c(B;\hat{X})$), it follows that higher-level causal explanations cannot be autonomous in the sense that they can provide information about the effect that the corresponding lower-level causal explanation could not provide.[19] However, they can still be explanatory autonomous in the weaker sense that they are not worse in answering w-questions about the effect than their lower-level competitors, a view recently defended by Woodward. According to Woodward ([2018], Section 5), this will be the case if (†) 'changes in [A] are causally relevant to [B ...] and conditional on the values taken by [A], further variations in [X] make no difference to [B]'. Woodward ([2018], Section 3) suggests the following updated characterization of proportionality, which is supposed to license the inference from (†) to the autonomy of the higher-level explanation:

(P*) Suppose we are considering several different causal claims/explanations formulated

[19] Recall from Section 4.3 that this holds in general, meaning that replacing B by its lower-level counterpart Y in the example above would make no difference. Also note that the higher-level cause A might still be more specific for the higher-level effect B than it is for Y. We are indebted to an anonymous reviewer for this point. However, the question of interest in this section is whether it is possible that the higher-level cause provides more information than its lower-level competitor with respect to the particular effect to be explained, that is a specific variable. Once we have decided on which effect we want to explain (either B or Y), the lower-level cause X is always at least as informative as A.

> in terms of different variables and representing different claims about patterns of dependency relations involving some target effect or explanandum E and where all of these satisfy some minimal interventionist condition. Then, other things being equal, we should prefer those causal claims/explanations that more fully represent or exhibit those patterns of dependence that hold with respect to E.

The idea is that if (†) holds, then both explanations, the one citing A as well as the one citing X, provide the same patterns of dependence that hold with respect to the higher-level effect B. Formulated within our framework, (†) requires that A has some causal influence on B (that is, $SPEC_c(B;\hat{A}) > 0$) and that learning X's value does not provide any additional causal information about B's value if we already know A's value (that is, $SPEC_c(B;\hat{X}|\hat{A}) = 0$). Since $SPEC_c(B;\hat{A})$ cannot be greater than $SPEC_c(B;\hat{X})$ and X does not provide any information in addition to A, it follows that $SPEC_c(B;\hat{A}) = SPEC_c(B;\hat{X})$. Hence, it turns out that higher-level causal explanations actually can be autonomous in Woodward's ([2018]) weak sense. However, note that (in contrast to what Woodward seems to assume) no new version of the proportionality constraint is needed to license this autonomy. All the work is already done by the specificity of the cause. As our analysis shows, replacing **(P)** by **(P*)** amounts to abandoning proportionality and relying on specificity of the cause in order to argue for the possibility of higher-level explanatory autonomy. The condition (b) requiring that the cause variable should have no irrelevant values present in **(P)** is gone in **(P*)**; **(P*)** only requires the cause to be maximally informative with respect to the effect, which is exactly what $SPEC_c$ measures.

Summarizing, the explanatory autonomy view fares a bit better than the causal autonomy view. Though it turns out that higher-level causal explanations (for example, psychological explanations) cannot be better than explanations based on their lower-level realizers (for example, neuroscientific explanations), they can, in accordance with Woodward's ([2018]) latest work, sometimes be equally good. However, it is interesting to see that in the end proportionality plays no role in arriving at this consequence. Like in the case of the causal autonomy view discussed in Section 4.3, it is rather the specificity of the cause that does all the work required to arrive at this result.

5 Conclusion

In this article, we argued that normalized mutual causal information of type one ($SPEC_c$) and normalized variation of information (NVI or its complement $PROP$) are suitable measures for the specificity of the cause and proportionality, respectively. We then applied these measures to the debate on the autonomy of higher-level causes and causal explanation. We showed that even though many philosophers have placed high hopes on proportionality as key to higher-level autonomy, it turns out to be more or less irrelevant. Even when higher-level causes (for example, mental states) are more proportional than their lower-level realizers (for example, neural states), they cannot have causal powers or influence that goes beyond the causal powers or influence of their realizers. They also cannot provide better explanations, at least if explanation is understood in terms of the ability to answer w-questions. This is because the specificity of the cause provides a better measure than proportionality for the extent to which a causal explanation provides answers to w-questions, and causes cited in higher-level explanations cannot be more specific than the causes cited in corresponding lower-level explanations. At best, higher-level explanations can be no worse than lower-level explanations.

This still leaves room for a modest explanatory autonomy. Although higher-level explanations cannot be better than lower-level explanations with respect to answering w-questions, they can be equally good, meaning that in those cases turning to lower-level explanations does not provide any additional answers to w-questions (Woodward [2018]). Importantly, there are also many pragmatic reasons for why higher-level explanations can be preferable to lower-level ones. For example, they can be more cognitively salient, more generalizable, or more computationally tractable (cf. Ylikoski and Kuorikoski [2010]; Woodward [2018]). In this article, however, we focused exclusively on non-pragmatic considerations.

More generally, we hope to have shown that the combination of causal Bayes nets and information theory provides a useful framework for analysing more philosophical issues concerning higher- versus lower-level causation. In future research, it can be applied to shed light on other topics such as downward causation, mechanistic explanation, and constitutive relevance as well.

Appendix

A.1 Causal modelling

The causal interpretation of Bayesian networks Pearl ([1988]) was developed by Spirtes *et al.* ([1993]) and later by Pearl ([2000]).[20] Causal models can be used for formulating and testing complex causal hypotheses, for prediction, as devices for computing the effects of hypothetical interventions, and as a basis for algorithmic procedures for uncovering causal structure.

A causal model is a triple $\langle \mathbf{V}, \mathbf{E}, P \rangle$, where $\mathbf{V}$ is a set of random variables $X_1, ..., X_n$, $\mathbf{E}$ is a set of directed edges connecting variables in $\mathbf{V}$, and P is a probability distribution over $\mathbf{V}$. The variables describe properties or event types and the directed edges indicate direct causal dependencies. Direct causes are also called causal parents. The set of causal parents of a variable X_i is $\mathbf{Par}(X_i)$. A chain of edges is called a causal path. Finally, P is intended to provide information about the strengths of the causal influences propagated over causal paths.

Causal models are assumed to satisfy the Markov condition (Spirtes *et al.* [1993], p. 16):

> Markov Condition: $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$ and P satisfy the Markov condition if and only if

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i | \mathbf{par}(X_i)). \tag{20}$$

The probabilities $P(x_i|\mathbf{par}(X_i))$ are called X_i's parameters. If interpreted causally, the Markov condition guarantees that every dependence between variables in $\mathbf{V}$ is due to some causal connection.

Causal models allow for drawing a distinction between predicting the values of variables if the values of other variables are observed and the effects of interventions. The probability distribution of a variable X_j after observing X_i's value can be computed on the basis of the Markov condition and the model's original graph $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, while X_j's distribution after intervening on X_i can be computed by applying the Markov condition to the graph resulting from $\mathbf{G}$ after deleting all the arrows pointing at X_i. While observing X_i's value may provide information about X_i's causes as well as its effects, intervening on X_i breaks the influences other causes might have on X_i. An intervention on X_i can, hence, only influence effects of

[20] Many of the relevant ideas were independently developed by Spohn ([1980]).

X_i. '$\hat{x}_i$' stands short for 'x_i is induced by an intervention on X_i'. If figuring in probability statements '$P(... \hat{x}_i ...)$' it indicates that the truncated graph resulting from deleting the arrows into X_i should be used for computing these probabilities. Typically it is assumed that $\hat{x}_i$ gets a probability of one while all other X_i-values get a probability of zero. In this article, however, we assume that $P(\hat{x}_i) = P(x_i)$.

A.2 Information theory

Information theory dates back to Shannon ([1948]). According to one interpretation it links the concept of information to the degree of uncertainty one might have about possible outcomes of interest: the less uncertainty, the more information. The following measures are relevant for this article:

Entropy:

$$H(Y) = -\sum_{y \in Y} P(y) \cdot log_2 P(y) \tag{21}$$

'$y \in Y$' stands short for 'y is a value of Y'. The entropy of Y can be interpreted as the degree of uncertainty about Y's value. It will be minimal if Y's probability distribution is extreme and maximal if all $y \in Y$ are equally likely (for example, in the case of a fair coin toss).

Conditional Entropy:

$$H(Y|X) = -\sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \cdot log_2 P(y|x), \tag{22}$$

where $log_2 0$ and $P(y|x)$ if $x = 0$ are treated as 0.

The conditional entropy $H(Y|X)$ can be interpreted as the average uncertainty about Y's value if one would learn some X-value. It is minimal if Y is fully determined by X and maximal if Y and X are probabilistically independent.

Mutual Information:

$$I(Y;X) = H(Y) - H(Y|X) \tag{23}$$

$I(Y;X)$ measures the degree of information one of the two variables bears about the other. It can be interpreted as a measure for how much learning the value of one of the variables would

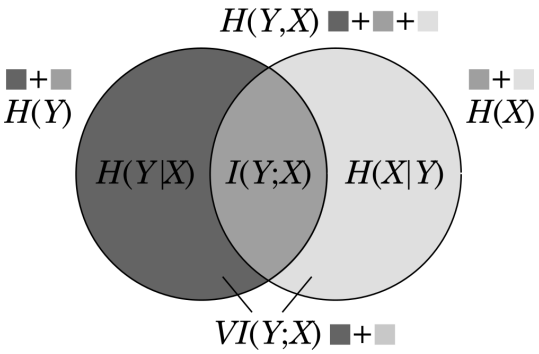

Figure 3

reduce the uncertainty about the value of the other variable. Mutual information of Y and X is minimal if Y and X are independent and maximal if Y and X fully determine each other. In contrast to conditional entropy, mutual information is symmetric.

Variation of Information:

$$VI(Y;X) = H(Y|X) + H(X|Y) \tag{24}$$

$VI(Y;X)$ is minimal if Y's value is fully determined by X's value (and vice versa) and maximal if no Y-value has a probabilistic impact on any X-value (and vice versa). Variation of information is closely related to mutual information:

$$VI(Y;X) = H(Y) + H(X) - 2I(Y;X) = H(Y,X) - I(Y;X). \tag{25}$$

How entropy, mutual information, and variation of information relate to each other is further illustrated by the diagram in Figure 3.

Acknowledgements

We would like to thank audiences at the workshop Compositional Explanation in Biology and the Neurosciences at Rutgers and Seton Hall University, at the PCCP meeting at the University of Groningen, and the Research Colloquium at the University of Düsseldorf for their input. Thanks also to the anonymous referees for helpful comments on an earlier version of this article.

Alexander Gebharter

Department of Theoretical Philosophy

University of Groningen

Groningen, Netherlands

alexander.gebharter@gmail.com

Markus I. Eronen

Department of Theoretical Philosophy

University of Groningen

Groningen, Netherlands

m.i.eronen@rug.nl

References

Ay, N. and Polani, D. [2008]: 'Information Flows in Causal Networks', *Advances in Complex Systems*, **11**, pp. 17–41.

Baumgartner, M. [2013]: 'Rendering Interventionism and Non-reductive Physicalism Compatible', *Dialectica*, **67**, pp. 1–27.

Baumgartner, M. and Gebharter, A. [2016]: 'Constitutive Relevance, Mutual Manipulability, and Fat-Handedness', *British Journal for the Philosophy of Science*, **67**, pp. 731–56.

Bontly, T. D. [2005]: 'Proportionality, Causation, and Exclusion', *Philosophia*, **32**, pp. 331–48.

Bourrat, P. [2019]: 'Variation of Information as a Measure of One-To-One Causal Specificity', *European Journal for Philosophy of Science*, **9**, pp. 1–18.

Davidson, E. H. [2001]: *Genomic Regulatory Systems: Development and Evolution*, London: Academic Press.

Eronen, M. I. [2011]: *Reduction in Philosophy of Mind*, Heusenstamm: De Gruyter.

Eronen, M. I. [2015]: 'Levels of Organization: A Deflationary Account', *Biology and Philosophy*, **30**, pp. 39–58.

Eronen, M. I. and Brooks, D. S. [2014]: 'Interventionism and Supervenience: A New Problem and Provisional Solution', *International Studies in the Philosophy of Science*, **28**, pp. 185–202.

Eronen, M. I. and Brooks, D. S. [2018]: 'Levels of Organization in Biology', in E. N. Zalta (*ed.*), *Stanford Encyclopedia of Philosophy*, available at <plato.stanford.edu/archives/spr2018/ entries/levels-org-biology/>.

Franklin-Hall, L. [2016]: 'High-Level Explanation and the Interventionist's Variables Problem', *British Journal for the Philosophy of Science*, **67**, pp. 553–77.

Gebharter, A. [2017a]: 'Causal Exclusion and Causal Bayes Nets', *Philosophy and Phenomenological Research*, **95**, pp. 353–75.

Gebharter, A. [2017b]: *Causal Nets, Interventionism, and Mechanisms: Philosophical Foundations and Applications*, Cham: Springer.

Gebharter, A. [2019]: 'A Causal Bayes Net Analysis of Glennan's Mechanistic Account of Higher-Level Causation (and Some Consequences)', *British Journal for the Philosophy of Science*, available at <doi.org/10.1093/bjps/axz034>.

Griffiths, P. E., Pocheville, A., Calcott, B., Stolz, K., Kim, H. and Knight, R. [2015]: 'Measuring Causal Specificity', *Philosophy of Science*, **82**, pp. 529–55.

Hitchcock, C. and Woodward, J. [2003]: 'Explanatory Generalizations, Part II: Plumbing Explanatory Depth', *Noûs*, **37**, pp. 181–99.

Hope, L. R. and Korb, K. [2005]: 'An Information-theoretic Causal Power Theory', in S. Zhang and R. Jarvis (eds), *AI 2005: Advances in Artificial Intelligence*, Berlin: Springer, pp. 805–11.

Kim, J. [2005]: *Physicalism, or Something Near Enough*, Princeton, NJ: Princeton University Press.

Lewis, D. [2004]: 'Causation As Influence', *Journal of Philosophy*, **97**, pp. 182–97.

List, C. and Menzies, P. [2009]: 'Nonreductive Physicalism and the Limits of the Exclusion Principle', *Journal of Philosophy*, **106**, pp. 475–502.

McDonnell, N. [2017]: 'Causal Exclusion and the Limits of Proportionality', *Philosophical Studies*, **174**, pp. 1459–74.

McLaughlin, B. P. [2007]: 'Mental Causation and Shoemaker-Realization', *Erkenntnis*, **62**, pp. 149–72.

Menzies. P. and List, C. [2010]: 'The Causal Autonomy of the Special Sciences', in P. Menzies and C. List (*eds*), *Emergence in Mind*, Oxford: Oxford University Press, pp. 108–29.

Pearl, J. [1988]: *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, San Mateo, CA: Morgan Kaufmann.

Pearl, J. [2000]: *Causality*, Cambridge: Cambridge University Press.

Pocheville, A., Griffiths, P. E. and Stolz, K. [2017]: 'Comparing Causes: An Information-theoretic Approach to Specificity, Proportionality and Stability', in H. Leitgeb, I. Niiniluoto, P. Seppälä and E. Sober (*eds*), *Logic, Methodology and Philosophy of Science: Proceedings of the Fifteenth International Congress of Logic, Methodology and Philosophy of Science*, London: College Publications, pp. 250–75.

Shoemaker, S. [2000]: 'Realization and Mental Causation', in J. Hintikka, R. Neville, E. Sosa and A. Olson (*eds*), *Proceedings of the Twentieth World Congress of Philosophy*, Philosophy Documentation Center, pp. 23–33.

Shannon, C. E. [1948]: 'A Mathematical Theory of Communication', *Bell System Technical Journal*, **27**, pp. 379–423.

Shapiro, L. A. [2010]: 'Lessons from Causal Exclusion', *Philosophy and Phenomenological Research*, **81**, pp. 594–604.

Sober, E. [1999]: 'Physicalism from a Probabilistic Point of View', *Philosophical Studies*, **95**, pp. 135–74.

Spirtes, P., Glymour, C. and Scheines, R. [1993]: *Causation, Prediction, and Search*, Dordrecht: Springer.

Spohn, W. [1980]: 'Stochastic independence, causal independence, and shieldability', *Journal of Philosophical Logic*, **9**, pp. 73–99.

Waters, C. K. [2007]: 'Causes That Make a Difference', *Journal of Philosophy*, **104**, pp. 551–79.

Weslake, B. [2013]: 'Proportionality, Contrast and Explanation', *Australasian Journal of Philosophy*, **91**, pp. 785–97.

Woodward, J. [2003]: *Making Things Happen*, Oxford: Oxford University Press.

Woodward, J. [2008]: 'Mental Causation and Neural Mechanisms', in J. Hohwy and J. Kallestrup (*eds*), *Being Reduced*, Oxford: Oxford University Press, pp. 218–62.

Woodward, J. [2010]: 'Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation', *Biology and Philosophy*, **25**, pp. 287–318.

Woodward, J. [2015]: 'Interventionism and Causal Exclusion', *Philosophy and Phenomenological Research*, **91**, pp. 303–47.

Woodward, J. [2018]: 'Explanatory Autonomy: The Role of Proportionality, Stability, and Conditional Irrelevance', *Synthese*, **61**, pp. 1–29.

Woodward, J. and Hitchcock, C. [2003]: 'Explanatory Generalizations, Part I: A Counterfactual Account', *Noûs*, **37**, pp. 1–24.

Yablo, S. [1992]: 'Mental Causation', *Philosophical Review*, **101**, pp. 245–80.

Ylikoski, P. and Kuorikoski, J. [2010]: 'Dissecting Explanatory Power', *Philosophical Studies*, **148**, pp. 201–19.

Zhong, L. [2014]: 'Sophisticated Exclusion and Sophisticated Causation', *Journal of Philosophy*, **11**, pp. 341–60.