Andreas Lüchinger*

Searching Probabilistic Difference-Making within Specificity

<https://doi.org/10.1515/krt-2021-0034>

Published online January 4, 2022

Abstract: The idea that good explanations come with strong changes in probabilities has been very common. This criterion is called probabilistic difference-making. Since it is an intuitive criterion and has a long tradition in the literature on scientific explanation, it comes as a surprise that probabilistic difference-making is rarely discussed in the context of interventionist causal explanation. Specificity, proportionality, and stability are usually employed to measure explanatory power instead. This paper is a first step into the larger project of connecting difference-making to the interventionist debate, and it starts by investigating whether probabilistic difference-making is contained in the notion of specificity. The choice of specificity is motivated by the observation that both probabilistic difference-making and specificity build on similar underlying intuitions. When comparing measures for both specificity and probabilistic difference-making, it turns out that the measures are not strictly correlated, and so the thesis that probabilistic difference-making is encoded within specificity has to be rejected. Some consequences of this result are discussed as well.

Keywords: specificity, probabilistic difference-making, interventionism, causal models, explanatory strength

1 Introduction

Here is an intuitive property of good explanations: Given an explanandum E, a good explanation for E ought to render E expected given the explanation (Hempel and Oppenheim 1948). Let us consider an example. Suppose that you are receiving a vaccine shot for the COVID-19 virus and asking yourself why the vaccine protects you from getting COVID-19. Such an explanation would presumably refer to some properties of vaccines and physiological processes which make not getting COVID-19 probable. This is the idea that a good explanation of E increases the

***Corresponding author: Andreas Lüchinger**, Department of MCMP, LMU München Fakultät für Philosophie Wissenschaftstheorie und Religionswissenschaft, 80539 München, Germany, E-mail: andreas.luechinger@outlook.com

probability of E to a sufficiently high degree. Unfortunately, this property was shown to be lacking as a desideratum (Salmon 1971). As it turns out, probability increase is only a special case of a more general principle – probabilistic difference-making (Salmon 1971; Schupbach 2011). This is the idea that an explanation changes the degree to which we expect the explanandum (Schupbach and Sprenger 2011, 108).¹

While probabilistic difference-making certainly has a lot of intuitive appeal and has motivated more classic accounts of scientific explanation, it is usually not discussed in connection to causal explanations (Hitchcock and Woodward 2003; Woodward 2010). Given an interventionist understanding of causal explanation, three criteria of explanatory strength are usually considered: specificity, proportionality, and stability (henceforth: the three dimensions). It is thus surprising that a very general desideratum of good explanations is usually not discussed in relation to the three dimensions (although probabilistic difference-making has been applied to measure causal explanatory strength independently (Eva and Stern 2019)). Even worse, it has been suggested that the three dimensions of good causal explanations might conflict with the effect magnitude (Ofstedal 2020).

The main aim of this essay is to investigate how the desideratum of probabilistic difference-making relates to specificity. Specificity determines how much control of the effect one has with a cause. Intuitively, this seems to be closely related to the idea of difference-making, as a high amount of specificity indicates that the cause makes a big difference for the effect. I conclude that, if understood in a causal modelling framework, probabilistic difference-making is not connected to the criterion of specificity. This is achieved through comparing formal measures for both criteria and checking if the measures are equivalent or not, in the sense that the same changes in one measure always result in the same changes in the other measure.

Section 2 discusses probabilistic difference-making and presents the concept. Subsequently, Section 3 presents interventionism as a popular framework to model causation and causal explanation. Finally, Section 4 compares probabilistic difference-making to specificity, and Section 5 concludes.

2 An Abundant Desideratum

The idea that a scientific explanation ought to show how the explanandum was to be expected has been present since the deductive-nomological model of explanation (Hempel and Oppenheim 1948). If one considers a scientific type of argument, be it deductive or inductively strong, then the explanans ought to make

¹ The idea can of course also be challenged, although I do not engage with such criticisms in this paper.

the explanandum sufficiently likely, that is the conditional probability of the explanandum given the explanans is sufficiently high. Nevertheless, not all cases where one event is made very likely by another one are satisfactory explanations (Salmon 1971, 33). One problem with the condition of high conditional probability is that there are explanations where the explanans is irrelevant for the explanandum: Suppose Marc is a man, and he is not pregnant. The fact that Marc is not pregnant is not explained by the fact that he has been taking birth control pills since he is a man. The point here is, of course, that taking the birth control pill is irrelevant because men do not get pregnant anyway (but the probability of Marc not getting pregnant given that he takes the birth control pill is 1).

This is why another criterion was proposed – statistical relevance (Salmon 1971). Here we have to depart from the idea that scientific explanations are arguments. Rather, we employ probability theory, and the notion of a random variable specifically. In probability theory, possible worlds are the primitive objects, and those possible worlds exhaust all possibilities. A random variable is a partition of those possible worlds, such that all possible worlds are grouped into mutually exclusive and jointly exhaustive sets. The individual sets are the values of a variable, and those values can be assigned probabilities.

To see what statistical relevance is, let us again consider the example above. The factor that somebody takes the birth control pill is not statistically relevant to the fact that somebody does not become pregnant if the person is a man. In this context, taking the pill is not statistically relevant to not being pregnant. On the other hand, taking the pill is statistically relevant to becoming pregnant if one is a woman. Statistical relevance of a property B concerning another property A and with respect to a reference class C is then given by $P(A|C, B) \neq P(A|C)$. Statistical relevance is therefore always relative to a reference class, which is here represented by C . However, for the remaining discussion, I employ a different formalism where reference classes are not used explicitly, namely random variables and their values (this formalism is also used in Eq. (1) below).

A short justification of statistical relevance is in place. I believe that the concept of probabilistic difference-making has a lot of intuitive appeal. Moreover, at least as a necessary condition for scientific explanation, it is desirable. To see this, consider whether a scientific explanation can be provided by a variable that is statistically independent of the explanandum variable (the variable encoding the information to be explained). I take this to be absurd, especially because explanation serves an epistemic function. The probabilities should give us some epistemic guidance. If two variables are probabilistically independent, there is no reason to adjust one's belief according to the explanation.

Moving away from statistical relevance, the idea has been taken up and has been refined, and this is where I arrive at a more recent understanding of probabilistic

difference-making. It can be seen that probabilistic difference-making has been influential in the literature on explanatory strength (Schupbach and Sprenger 2011). In this context, one is interested in determining how good explanations are. Measures should then determine explanatory power quantitatively. One possible yardstick for this purpose is the following expression (Schupbach and Sprenger 2011, 113):

$$\epsilon(h, e) = \frac{p(h|e) - p(h|\neg e)}{p(h|e) + p(h|\neg e)} \quad (1)$$

This measure satisfies several desiderata, among which is the desideratum of positive relevance: The greater the statistical relevance between the values e (the explanandum event) and h (the explanans event or hypothesis), the greater the value of the measure $\epsilon(h, e)$ (Schupbach 2011, 110). Thus, statistical relevance is embedded in such a measure, and statistical relevance increases the strength of an explanation. The measure $\epsilon(h, e)$ then contrasts the likelihood of h given e compared to h given $\neg e$, since the measure is constructed for propositional variables (with only two values). While ϵ is a measure for explanatory strength, I argue that it is also an adequate measure for probabilistic difference-making. The measure was developed to account for the degree to which a previously surprising phenomenon is rendered a matter of course by the explanation (Schupbach and Sprenger 2011, 108). Hence, ϵ is not only a measure of explanatory strength as a whole but also encapsulates the degree to which an explanans h makes a difference for e . The concept of difference-making is here represented as the degree to which a hypothesis makes the explanandum less surprising for an agent. For the rest of this essay, I explicate probabilistic difference-making in terms of $\epsilon(h, e)$.

This is where we can return to the main question of my essay, namely how probabilistic difference-making relates to the literature on the three dimensions. Although, as we have seen, probabilistic difference-making ought to be a typical condition of any explanation, it is even more surprising that it is not explicitly brought up in the interventionist criteria for good causal explanations. After I shortly present the interventionist account of scientific explanation, the main bulk of my essay is devoted to discussing how probabilistic difference-making is related to specificity.

3 Interventionism

Moving away from traditional accounts such as Hempel's and Salmon's, more recent accounts have focused on causal theories of scientific explanation. Moreover, one influential idea to uncover causal relationships has been interventionism. The bottom line of this framework is the following: Causes and causal relationships are revealed through interventions. Interventionism thus provides the tools necessary to

discover causal relationships. Moreover, it enables a pragmatic approach to causation, where manipulations instead of processes take centre stage. In interventionism, variables play a very important role. Variables themselves can be understood as explained in the last section in the context of probability theory, namely as partitions of possible worlds. Thus variables are equipped with different values. As an example, consider variables S for smoking and another variable C for cancer, which have the values $\{smoker, non-smoker\}$ and $\{cancer, no-cancer\}$.

Let us now take a look at an example of an explanation. Suppose Fred is diagnosed with lung cancer. I take it to be clear that smoking causes cancer, and that Fred's smoking together with some account of how smoking causes cancer qualifies as a scientific explanation. The sense in which variables are important in the framework is that value assignments need to be explained, and value assignments are at least part of an explanation. For instance, Fred having cancer can be expressed through the variable C and a corresponding value c representing having cancer. If one provides a functional statement relating C to S , for example, a causal law describing how the value of C depends on the value of S , then one has a complete explanation.

Not every such functional relation is eligible for good scientific explanations, however. The relationships need to be sufficiently invariant under interventions on the cause variable. For instance, a law providing the value of the variable C depending on the value of S ought to be invariant in this sense. To test whether any functional relationship obeys the intervention criterion, one manipulates the value of the variable S (makes people smoke or not) and observes whether this intervention results in changes of the variable C (people develop cancer or not) according to the relationship. The causal law together with a value of the cause variable then provides a legitimate explanation for Fred's unfortunate development of cancer.

When intervening upon a variable, one has to be careful about how the notion of intervention is spelled out. One way to achieve a satisfying intervention is by introducing another variable, called an intervention variable (denoted by ' T '), which is used to express the manipulative force upon the cause variable in question (Woodward and Hitchcock 2003, 12–13). However, intervention variables are only one method to actualise a more general idea of intervention: directly targeting only one variable in the causal modal, and manipulating only this variable. Imagine the following example:

Let P be a probability distribution over the variables. Now we consider an intervention on the variable X_2 , for which I write $\hat{X}_2$. The hat operator denotes that the arrows going into X_2 are broken (I adopt the convention that the post-intervention probability distribution adopts the prior probabilities of the values of X_2 in the pre-intervention distribution. The prior probabilities over the values of $\hat{X}_2$ are identical to the prior probabilities over the values of X_2 , but the dependence on causes of X_2 is removed by breaking the arrows.):

The post-intervention graph illustrates how the direct causal influence of X_1 has been taken away. This ensures that, if the causal influence of X_2 on X_3 is considered, the influence due to the common cause X_1 is eliminated. This is interesting since it is then possible to consider the net effect of X_2 on X_3 while excluding correlations resulting from common causes (the variable X_1). In the example, $\hat{X}_2$ does not exert any causal influence on X_3 at all.

After this brief introduction to interventionism, I move on to the three dimensions of explanatory strength. In this framework, there are a few indicators that separate good explanations from bad ones. The three most commonly discussed dimensions are specificity, proportionality, and stability. After shortly discussing all three dimensions, it will become clear that probabilistic difference-making is not explicit in those three dimensions. This is surprising because the three dimensions are sometimes regarded to be sufficient for evaluating explanatory strength.

Specificity expresses the amount of control the cause variable has on the effect variable. Another way to put it is that by changing the values of the cause variable, specificity captures how well the different values of the effect variable can be enforced. This is important for explanations since the more precisely the effect values are connected to cause values, the more relevant the cause is for making the effect happen. The more specificity an explanation offers, the more values of the cause variable determine values of the effect variable, and in the case of maximal specificity the values between the variables are connected through a surjective function from the cause variable to the effect variable (Woodward 2010, 305). Phrasing it probabilistically: The more specificity an explanation offers, the closer conditional probabilities $p(e|c)$ (effect values given cause values) are to the values 0 or 1.

Let us consider an example. Imagine a switch with two possible settings, up and down, and the switch is connected to a light bulb. When the switch setting is ‘up’, the light bulb is on, and when the switch setting is ‘down’, the light bulb is off. Then, the switch is maximally specific for the effect, since one has maximal control over the possible settings of the light bulb. If we now imagine another switch with three settings – up, down, and middle – and the light bulb can still be only on or off, then one can still have maximal control over the light bulb with the switch, for example in the following manner: When the switch is up, then the light bulb is on, when the switch is down, the light bulb is off, and when the switch is in the middle, the light bulb is also on. One still has maximal control over the light bulb with the switch, although the switch has an unnecessary setting. A maximally unspecific switch for the light bulb is one where, given the setting of the switch, the light bulb can be either on or off with a probability of 0.5. This means that the switch effectively grants no control over the light bulb.

Proportionality encodes how much essential information is encoded in an explanation, while unnecessary information should be excluded as much as possible. An example is usually very helpful here: Consider a pigeon that has been conditioned to peck if and only if it is presented with a red coloured cloth. Now consider two explanations for the bird’s pecking (Yablo 1992):

- The bird pecked because it was presented with a red cloth.
- The bird pecked because it was presented with a scarlet cloth.

It should be apparent that the first explanation is better than the second one. The reason why the second explanation is inferior is that it only provides information for cases with scarlet cloths, whereas the first explanation also covers cases where cloths with other shades of red are presented. Thus, the first explanation correctly provides all necessary and only the necessary information. This exemplifies proportionality: The first explanation is more proportional than the second. In other words, it provides the necessary information while excluding irrelevant information.

Now I turn to stability (Woodward 2010, 291–96). The key intuition here is the following: the more stable an explanation under changes in the background conditions, the better. When a causal explanation is given, this is always relative to certain background assumptions. If an explanation is exceptionally stable, it remains valid even if those background assumptions are altered. Stability itself differs somewhat from specificity and proportionality, as stability does not only depend on the cause and effect variables, but also all of the background variables. Stability ought to account for the fact that we tend to prefer (*ceteris paribus*) explanations which have a more general validity than explanations that are limited to a smaller area of application (compare explanations in chemistry to explanations in physics, for instance).

As is apparent from the descriptions of the three dimensions of causal explanatory strength, none of them explicitly refers to probabilistic difference-making. This leads to my main argument of this essay: I investigate whether probabilistic difference-making is contained within the notion of specificity after shortly motivating why this might be the case. For this aim, I consider a formal measure of specificity. Moreover, probabilistic difference-making has to be generalised to relate variables instead of values. Only once such a generalisation is obtained, can the comparison between the concepts be conducted.

4 Comparing Specificity and Probabilistic Difference-Making

4.1 Specificity and Information Theory

Let me first motivate why specificity, and not proportionality or stability, may show interesting relations to probabilistic difference-making. If we consider specificity as the degree to which the effect values are the image of a surjective function with the cause values as arguments, then this seems to indicate that the effect value should be expected given a corresponding cause value. Another reason for this suspicion is due to control being tantamount to maximising the conditional probabilities of single values of the effect variable given some values of the cause variables. If we then relate $p(h|e)$ to $p(e|h)$ through Bayes' theorem, we obtain: $p(h|e) = \frac{p(e|h)p(h)}{p(e)}$, and so one might suspect that through increasing specificity, we also affect probabilistic difference-making, which contains $p(h|e)$.

I have presented specificity in a manner that makes use of functional relationships between values, where specificity should extract how closely the values of the effect variable are the image of some function with the values of the cause variables as arguments. Such a function ought to be surjective since this property grants maximal control of the effect through the cause. A first notable feature is that, while the ideal is a specific type of a function, specificity comes in degrees. Hence, we are interested in approximations to a (surjective) function. But how can one determine approximations to a function?

It has been suggested that information theory is very helpful in this context (Pocheville, Griffiths, and Stotz 2017). Since specificity determines the amount of control, the information that variables contain or share with other variables is very helpful. Importantly, information theory makes use of probability theory, and thus the connection to probabilistic difference-making can be examined much more

easily. For this purpose, the notion of entropy has to be introduced. Entropy then renders the mutual information of variables precise. The entropy of a variable X is:

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i) \quad (2)$$

and the conditional entropy of a variable X given a variable Y is:

$$H(X|Y) = -\sum_j p(y_j) \sum_i p(x_i|y_j) \log_2 p(x_i|y_j) \quad (3)$$

Entropy measures the uncertainty contained in a variable. Uncertainty ought to be understood as the indeterminacy of values, where uncertainty is maximal if the probability distribution over a variable is uniform and minimal if the distribution is extreme for one value. Thus, if entropy is very low, then the amount of information in a variable is very high. Information can not only be determined for single variables, but also as shared information between variables.

Mutual information is then defined as

$$I(X; Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y) \quad (4)$$

Put into words, mutual information measures how much knowing Y reduces the uncertainty in a variable X . To obtain a clearer picture of how mutual information works, consider the formulation $I(X; Y) = H(X) - H(X|Y)$. If Y does not have anything to say about X , then $H(X|Y) = H(X)$ and therefore their mutual information is minimal. If Y completely determines X , then $H(X|Y) = 0$, and the mutual information of the variables is maximal and equal to the entropy of X . Any other amount of mutual information can be found between those two extreme cases.

In order to render mutual information suitable for measuring causal impact, one has to introduce interventions into mutual causation. For interventions on a variable, say C , I use the hat symbol: $\hat{C}$ (remember that an intervention does not change prior probabilities, but only breaks arrows and deletes all dependence relations of C upon causes of C). One possible measure for causal specificity is the following (Pocheville, Griffiths, and Stotz 2017):²

$$I(E; \hat{C}) = H(E) - H(E|\hat{C}) \quad (5)$$

This conception of specificity allows us to investigate how probabilistic difference-making might be encoded in specificity.³ Let us first observe the right-hand side of Eq.

² The discussion in this essay presupposes that a probability distribution is given for a situation of interest. Of course, how such a distribution can be obtained is an interesting question, but not one I attempt to answer here. The presented discussion itself also presupposes that such a probability distribution is given and builds on this assumption.

³ While mutual causal information has been regarded as the groundwork for specificity, it has been argued that mutual causal information needs to be compared to different quantities to obtain

(5) above. Specificity is then composed of the entropy of the effect variable, and the conditional entropy of E given $\hat{C}$ is subtracted. If we assume that C is very specific for E , then the second term ought to be comparably small. Likewise, if C is not specific at all, then the second term must be comparably large (assuming that the entropy of E is just given, it does change with respect to differences in other variables).

This further suggests that specificity might contain the notion of difference-making, as the more mutual information the variables have, the bigger the difference-making effect (at least intuitively). There is one remaining hurdle before the two concepts can be compared: Probabilistic difference-making relates a value of the cause variable to two values of the effect variable. This is a hurdle because specificity is a measure relating variables to each other, and not values. For this reason, we need to make the measure of probabilistic difference-making fit for variables. Once this aim is accomplished, specificity and probabilistic difference-making can be compared to each other. This reveals how the concepts are related. But why should we be able to extrapolate probabilistic difference-making to a function of variables instead of individual values at all?

4.2 Why Probabilistic Difference-Making for Variables?

As we have observed, specificity and probabilistic difference-making need to be brought on the same level in order to be compared. This leaves us with two options: The first is formulating probabilistic difference-making such that it connects variables, and the second amending specificity to connect individual values instead of entire variables. I argue that the second option is inferior to the first one. This is mainly due to the attraction of interventionism which guides us to consider patterns of counterfactual dependence when evaluating explanatory strength. Moreover, this essay is concerned with causal explanations, which is the reason why counterfactual considerations have to influence explanatory strength. Hence, not only single values determine explanatory strength, but entire variables.

This approach presupposes that counterfactual dependence is crucial for evaluating explanatory strength. This is both the position I endorse and a philosophical conception that has been widely influential since its proposal (Hitchcock and Woodward 2003; Pearl 2000; Spirtes and Glymour 2000; Woodward and Hitchcock 2003). Since I am concerned with causal explanatory

different senses of specificity (Gebharder and Eronen 2021). One such sense is the specificity of the cause for the effect: $\frac{I(C;E)}{H(E)}$. As it turns out, the same argument presented in this paper also holds when specificity of the cause for the effect is considered, so I will continue using mutual causal information for simplicity.

strength, an interventionist framework is one of the most attractive choices for explicating causation and causal explanation. Once we adopt interventionism, the focus on entire variables becomes very natural. If we move the spotlight onto probabilistic difference-making, this suggests that we ought to look for it on a variable level (if it can be found within interventionism at all). Consequently, we ought to concentrate on variables rather than individual values.

The other side of the coin is that finding a value level complement of specificity is not attractive. Of course, we could technically propose that a single term for single values could be extracted from mutual information, where only the values of interest would be considered instead of sums over all values. My point is rather that such a move would be at odds with the interventionist spirit. Only taking individual values would completely neglect patterns of counterfactual dependence. Omitting other values does not consider invariance of a causal relationship under certain testing interventions, hence the tension with interventionism. It is therefore not a technical impossibility to formulate specificity on a value level, but an undesirable strategy given interventionism as a theory of causation.

Given interventionism, it makes sense to look for difference-making on the level of entire variables. If probabilistic difference-making can be made sense of in interventionism, it is most likely a property connecting variables and not individual values. Those considerations show that we should lift probabilistic difference-making to the variable level rather than bringing specificity down to the value level. Interestingly, the presented considerations are not entirely limited to the framework of interventionism. In fact, any analysis of causation that makes use of such counterfactual notions would motivate the same strategy. Following my argument, the remaining part of this section constructs a variable level analogue of probabilistic difference-making and compares it to specificity.

4.3 Constructing a Measure and Comparison

A first important step is to generalise the measure $\epsilon(c, e)$ from propositional variables (variables with only two values) to arbitrarily many values. This can be achieved in the following manner:

$$\epsilon(c_i, e_j) = \frac{p(c_i|e_j) - p(c_i|\neg e_j)}{p(c_i|e_j) + p(c_i|\neg e_j)} \quad (6)$$

The measure is both generalised and the notation makes explicit that causes are identified with hypotheses. Probabilistic difference-making is generalised in the sense that it compares the conditional probability of a value c_i given some value e_j in comparison to the conditional probability of c_i given the union of all remaining

values, that is $\neg e_j := \cup_{k \neq j} e_k$. Since values are sets of possible worlds, $\neg e_j$ then includes all possible worlds which are not members of e_j . This allows a comparison of all values c_i of the cause variable and the difference different values e_j make in comparison to the union of the other values e_k , and applying this measure on all possible combinations yields the probabilistic difference-making on the level of variables instead of values.

An initial idea to measure probabilistic difference-making on a variable level might simply be to average over all value combinations of the variables. Moreover, given an interventionist framework of causal explanation, we highlight the fact that an intervention operation has been executed on C and so we write $\hat{C}$ for the intervention on the cause variable and $\hat{c}_i$ for the post-intervention values, respectively:

$$\epsilon'(\hat{C}, E) = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \frac{p(\hat{c}_i|e_j) - p(\hat{c}_i|\neg e_j)}{p(\hat{c}_i|e_j) + p(\hat{c}_i|\neg e_j)} \quad (7)$$

The measure contains a factor $\frac{1}{mn}$ to account for the fact that different variables may have different amounts of values, yet then a direct comparison with the measure without the factor would be misleading (the score of a variable with many values may be exaggerated). The sums ensure that all difference-making factors of the variables are taken into account, and so a weighted average is taken over all individual pairs of values. However, there is one major problem with this measure: Consider two propositional variables C, E with values c_1, c_2 and e_1, e_2 . Then, $\epsilon'(\hat{C}, E) = 0$ since

$$\begin{aligned} \epsilon'(\hat{C}, E) &= \frac{1}{4} \left(\frac{p(\hat{c}_1|e_1) - p(\hat{c}_1|\neg e_1)}{p(\hat{c}_1|e_1) + p(\hat{c}_1|\neg e_1)} + \frac{p(\hat{c}_1|e_2) - p(\hat{c}_1|\neg e_2)}{p(\hat{c}_1|e_2) + p(\hat{c}_1|\neg e_2)} \right. \\ &\quad \left. + \frac{p(\hat{c}_2|e_1) - p(\hat{c}_2|\neg e_1)}{p(\hat{c}_2|e_1) + p(\hat{c}_2|\neg e_1)} + \frac{p(\hat{c}_2|e_2) - p(\hat{c}_2|\neg e_2)}{p(\hat{c}_2|e_2) + p(\hat{c}_2|\neg e_2)} \right) \\ &= \frac{1}{4} \left(\frac{p(\hat{c}_1|e_1) - p(\hat{c}_1|e_2)}{p(\hat{c}_1|e_1) + p(\hat{c}_1|e_2)} + \frac{p(\hat{c}_1|e_2) - p(\hat{c}_1|e_1)}{p(\hat{c}_1|e_2) + p(\hat{c}_1|e_1)} + \frac{p(\hat{c}_2|e_1) - p(\hat{c}_2|e_2)}{p(\hat{c}_2|e_1) + p(\hat{c}_2|e_2)} \right. \\ &\quad \left. + \frac{p(\hat{c}_2|e_2) - p(\hat{c}_2|e_1)}{p(\hat{c}_2|e_2) + p(\hat{c}_2|e_1)} \right) = 0 \end{aligned}$$

It is evident that the first two and the last two terms of the sum cancel each other out, and the total sum is equal to 0. This should not be the case. All terms indicate a potential amount of difference-making, no matter if those terms are negative or positive, and thus a solution is required. One way to fix the problem is by squaring the term over which the sum is taken, such that:

$$\epsilon(\hat{C}, E) = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \left(\frac{p(\hat{c}_i|e_j) - p(\hat{c}_i|\neg e_j)}{p(\hat{c}_i|e_j) + p(\hat{c}_i|\neg e_j)} \right)^2 \quad (8)$$

The squaring of the term ensures that the terms do not cancel each other out in the binary case, since they are both positive. This is desirable since $\epsilon(\hat{C}, E)$ ought to measure the amount of difference which C makes for E . For this purpose, it is not important whether the difference is positive or negative, what matters is the magnitude of the difference which values of C make for values of E . Whether the difference-making term is positive or negative does matter when explanatory strength is concerned, since then a negative value indicates that the cause value would have been more likely if the explanandum had been absent. However, it does not matter if we are merely concerned with absolute difference-making, which does not differentiate between such positive and negative contributions to overall explanatory strength, but simply the degree of difference-making independent of such a positive or negative contribution. Squaring the term guarantees this desired property. Finally, $\epsilon(\hat{C}, E)$ averages over all possible combinations of cause and effect values.

One might still think that the double-counting of values is odd because for each pair e_j and its negation the respective two terms are identical when squared (in the case of propositional variables), but this feature should not affect the comparison with specificity, as specificity also compares all different value combinations. Another interesting remark stems from the fact that if E only has one value, then the measure cannot be applied, but this case is not interesting anyway. The entire discussion has not addressed the issue of the denominator being equal to zero, in which case one could simply set the term equal to zero as in this case there is no sense in which such a term expresses any difference-making contribution.

Now specificity and the measure $\epsilon(\hat{C}, E)$ can be compared side by side. For this purpose, the two measures can be applied to different probability distributions. Once the values are obtained, their changes can be observed. More precisely, I am interested in observing whether both measures always increase or decrease from one probability distribution to another. For this purpose, I have restricted the variables to propositional variables, that is C and E only have two values (c_1, c_2, e_1, e_2). Consequently, only three parameters need to be specified for a complete probability distribution (under the assumption that C causes E , such that $C \rightarrow E$): the prior probability of c_1 , and the conditional probabilities $p(e_1|c_1)$ and $p(e_1|c_2)$. The remaining probabilities are determined by the axioms of probability theory ($p(c_2) = 1 - p(c_1)$, $p(e_2|c_1) = 1 - p(e_1|c_1)$, ...). The three parameters were given one of the probabilities 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The extreme values 0 and 1 were avoided due to complications when dividing by 0 or taking the logarithm of 0. All possible combinations of probabilities resulted in $9^3 = 729$ different probability distributions, for which specificity and probabilistic difference-making can be compared. From this table, I extracted the values of the measure $I(E; \hat{C})$ for

Figure 1: Values (vertical axis) of specificity (blue) and probabilistic difference-making (orange) over the probability distributions (horizontal axis), where the distributions are ordered such that specificity is gradually decreasing from left to right. While $I(E; \hat{C})$ decreases monotonously, $\epsilon(\hat{C}, E)$ fluctuates between increases and decreases.

specificity and $\epsilon(\hat{C}, E)$ for probabilistic difference-making, sorted them with decreasing values of specificity, and plotted the result in Figure 1.

An interesting question concerning specificity and probabilistic difference-making is whether an increase in one implies an increase/decrease in the other and vice versa. This would indicate that the information of probabilistic difference-making is indeed contained in the notion of specificity. However, in general, this is not the case, at least when the entire probability distribution is allowed to differ. Let us formulate the thesis that an increase of specificity implies an increase/decrease of probabilistic difference-making and vice versa. Then, Figure 1 shows that while specificity is gradually decreasing from left to right, probabilistic difference-making fluctuates. One counterexample is extracted in Table 1, but the graph shows that there are many others.

Table 1: Three probability distributions that show that increases/decreases of the measure $I(E; \hat{C})$ for specificity are not strictly correlated with increases/decreases of $\epsilon(\hat{C}, E)$ for probabilistic difference-making in general.

Distribution	$p(c_1)$	$p(e_1 c_1)$	$p(e_1 c_2)$	$I(E; \hat{C})$	$\epsilon(\hat{C}, E)$
1	0.5000	0.1000	0.9000	0.5310	0.6400
2	0.4000	0.1000	0.8000	0.3781	0.4959
3	0.2000	0.9000	0.1000	0.3578	0.5626

Consider distributions 1 and 2 in Table 1. It can be observed that both $I(E; \widehat{C})$ and $\epsilon(\widehat{C}, E)$ decrease from 1 to 2. While $I(E; \widehat{C})$ also decreases from 2 to 3, $\epsilon(\widehat{C}, E)$ in fact increases. So there are cases in which specificity decreases and probabilistic difference-making also decreases, and there are cases in which specificity decreases, yet probabilistic difference-making increases. If a measure for specificity also were a measure for probabilistic difference-making, then this scenario would be impossible. So, it has been shown that probabilistic difference-making is not contained within specificity.

However, there may still be two constraints worth investigating, which might salvage the claim that probabilistic difference-making is contained within specificity. Let us consider a first constraint. When specificity is employed, one is usually interested in comparing different amounts of specificity relative to one specific effect. This is the case when multiple explanations for an observed phenomenon are compared. The observation fixes the probability distribution over E , and comparing different explanations amounts to changing the probability distributions over the cause variables, but not the effect variable. Consequently, the prior probabilities over the effect variable do not change. Thus, one might add the first constraint to the claim: *Holding the prior probabilities over the values of the effect variable constant*, an increase of specificity implies an increase/decrease of probabilistic difference-making and vice versa.

This claim is consistent with the probability distributions from Table 1 since the prior probabilities over the variable E are not constant. Nevertheless, even the thesis equipped with the first constraint turns out to be incorrect, as the counterexample in Table 2 shows. The prior probabilities over E are kept constant, while the other parameters are allowed to change. From the distribution 4 to 5, both $I(E; \widehat{C})$ and $\epsilon(\widehat{C}, E)$ increase. Yet, from 5 to 6, $I(E; \widehat{C})$ increases, while $\epsilon(\widehat{C}, E)$ decreases. Thus, even the first constrained thesis fails.

A second possible constraint is to fix the probability distribution over the cause variable instead of the effect variable. Rather than comparing different

Table 2: Three probability distributions that show that increases/decreases of the measure $I(E; \widehat{C})$ for specificity are not strictly correlated with increases/decreases of $\epsilon(\widehat{C}, E)$ for probabilistic difference-making, even if the prior probabilities over the effect variable are kept constant.

Distribution	$p(c_1)$	$p(e_1 c_1)$	$p(e_1 c_2)$	$p(e_1)$	$I(E; \widehat{C})$	$\epsilon(\widehat{C}, E)$
4	0.2000	0.5000	0.4000	0.4200	0.0047	0.0137
5	0.1000	0.6000	0.4000	0.4200	0.0105	0.0617
6	0.4000	0.3000	0.5000	0.4200	0.0289	0.0457

explanations for a given observation, one would compare different explanations for a given cause variable, that is how much explanatory strength one cause offers for different effects. One might think of cases where one is interested in comparing the potential of a cause to provide explanations for different phenomena in one causal model. Although I believe that the first constraint treated above is more relevant for pragmatic reasons, the second constraint might still offer a last possible way in which specificity and probabilistic difference-making are related. Therefore, the second constrained thesis is the following: *Holding the prior probabilities of the values of the cause variable constant*, an increase of specificity implies an increase/decrease of probabilistic difference-making and vice versa.

Alas, the thesis with the second constraint is also incorrect, and a counterexample is shown in Table 3. While specificity decreases both from 7 to 8 and from 8 to 9, probabilistic difference-making decreases from 7 to 8, but increases from 8 to 9. Consequently, the second constrained thesis also turns out to be incorrect.

Therefore, it has to be concluded that probabilistic difference-making cannot be found within the notion of specificity. Since I have argued that there might be some intuitive reasons why the two concepts might be closely related, this circumstance asks for an explanation. Interestingly, those intuitions might arise due to the fact, while specificity and probabilistic difference-making are not strictly correlated, they are loosely correlated (as Figure 1 indicates). Neglecting fluctuations in probabilistic difference-making, the overall trend of both measures is a decrease from left to right, and this might explain the intuition that the two concepts are closely related. However, the counterexamples show that control over the effect variable with the cause variable can increase, while the difference-making between the variables in fact decreases. I argue that an intuitive reason why specificity and probabilistic difference-making come apart is that difference-making is contrastive, while specificity is not. The precise mathematical measures reveal that there are structural aspects about causation and explanation which allow specificity and probabilistic difference-making to part ways, and those differences are engrained within the counterexamples in the tables.

Table 3: Three probability distributions that show that increases/decreases of the measure $I(E; \hat{C})$ for specificity are not strictly correlated with increases/decreases of $\epsilon(\hat{C}, E)$ for probabilistic difference-making, even if the prior probabilities over the cause variable are kept constant.

Distribution	$p(c_1)$	$p(e_1 c_1)$	$p(e_1 c_2)$	$I(E; \hat{C})$	$\epsilon(\hat{C}, E)$
7	0.1000	0.9000	0.1000	0.2111	0.5073
8	0.1000	0.7000	0.2000	0.0734	0.2915
9	0.1000	0.1000	0.6000	0.0720	0.3524

A possible objection to my argument is that I have employed a specific measure of specificity. While it is true that a formal representation of specificity is employed, I believe that it does not weaken my argument. A representation of specificity enables a precise and concise comparison between concepts, and especially logical relations can be explored in great detail. Moreover, the application of information theory to an interventionist account of causal explanation has been explored before (Bourrat 2019; Pocheville, Griffiths, and Stotz 2017). In case one prefers another measure to mutual causal information, then the same analysis can be applied to the different measure. It is worth noting that the same result can also be obtained if specificity is conceived of as specificity of the cause for the effect, so the choice of the measure is not vital for my argument.

One may also reject the measure $\epsilon(\hat{C}, E)$. In this case, my argument is impaired significantly. On the other hand, such a rejection would show why the measure is mistaken. The investigation of probabilistic difference-making on the level of variables is worthwhile in its own right, and even if the measure $\epsilon(\hat{C}, E)$ turns out to be incorrect, the reasons for such a rejection are valuable on their own terms. Hence, I hope that any refusal of the ideas I have put forward in this paper leads to a fruitful discussion.

5 Conclusions

This essay has posed a riddle for interventionist accounts of causal explanation, namely why probabilistic difference-making is not explicitly included as a marker for good explanations despite being a commonly accepted criterion. In the course of my argument, I have presented an information-theoretic measure for specificity. Moreover, probabilistic difference-making has been interpreted on the level of variables which enabled a comparison with specificity. I have shown that in general probabilistic difference-making is not contained within specificity, and even two constrained theses have to be rejected due to counterexamples.

This leaves the three dimensions of interventionism somewhat incomplete. If probabilistic difference-making is not covered by the three dimensions, then it needs to be added, since otherwise, a vital feature of explanatory strength is missing. Hence, the three dimensions may lack crucial information for good explanations. If this turns out to be the case, then interventionist criteria for evaluating causal explanatory strength are not sufficient for evaluating explanatory power. Interventionism would then have neglected an important aspect of explanatory power. This raises the question of how probabilistic difference-making can adequately be added to the three dimensions to evaluate explanatory

power. It is thus evident that filling the gap between interventionism and probabilistic difference-making helps uncover some important aspects of explanations.

That being said, I have not investigated possible connections to proportionality and stability. It might still be the case that any other dimension, or sum of dimensions, provides the required information about difference-making. This points the way to future research. One idea is that proportionality, stability, and specificity may jointly provide the information contained within probabilistic difference-making. This is because proportionality ensures that irrelevant information is excluded, and stability provides background information about the causal relationship. Therefore, while specificity does not bear the desired connection to probabilistic difference-making, the three dimensions may jointly achieve this aim.

References

- Bourrat, P. 2019. "Variation of Information as a Measure of One-to-One Causal Specificity." *European Journal for Philosophy of Science* 9 (1): 1–18.
- Eva, B., and R. Stern. 2019. "Causal Explanatory Power." *The British Journal for the Philosophy of Science* 70 (4): 1029–50.
- Gebharder, A., and M. Eronen. 2021. "Quantifying Proportionality and the Limits of Higher-Level Causation and Explanation." *The British Journal for the Philosophy of Science*: 1–44, <https://doi.org/10.1086/714818>.
- Hempel, C., and P. Oppenheim. 1948. "Studies in the Logic of Explanation." *Philosophy of Science* 15 (2): 135–75.
- Hitchcock, C., and J. Woodward. 2003. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." *Noûs* 37 (2): 181–99.
- Oftedal, G. 2020. "Problems with Using Stability, Specificity, and Proportionality as Criteria for Evaluating Strength of Scientific Causal Explanations: Commentary on Lynch et al. (2019)." *Biology and Philosophy* 35 (1): 1–5.
- Pearl, J. 2000. *Causality*. Cambridge: Cambridge University Press.
- Pocheville, A., P. E. Griffiths, and K. Stotz. 2017. "Comparing Causes: An Information-Theoretic Approach to Specificity, Proportionality and Stability." In *Logic, Methodology and Philosophy of Science: Proceedings of the Fifteenth International Congress*, edited by H. Leitgeb, I. Niiniluoto, P. Seppälä, and E. Sober, 250–75. London: College Publications.
- Salmon, W. C. 1971. "Statistical Explanation." In *Statistical Explanation and Statistical Relevance*, 29–88. Pittsburgh, PA: University of Pittsburgh Press.
- Schupbach, J. N. 2011. "Comparing Probabilistic Measures of Explanatory Power." *Philosophy of Science* 78 (5): 813–29.
- Schupbach, J. N., and J. Sprenger. 2011. "The Logic of Explanatory Power." *Philosophy of Science* 78 (1): 105–27.
- Spiroes, P., and C. Glymour. 2000. *Causation, Prediction, and Search*, 2nd ed. Cambridge: MIT Press.

- Woodward, J. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology and Philosophy* 25 (3): 287–318.
- Woodward, J., and C. Hitchcock. 2003. "Explanatory Generalizations, Part I: A Counterfactual Account." *Noûs* 37 (1): 1–24.
- Yablo, S. 1992. "Mental Causation." *Philosophical Review* 101 (2): 245–80.