Subjective Causality

Yotam Alexander*
Itzhak Gilboa**

When do people tend to believe that a phenomenon x was the cause of a subsequent phenomenon y? We suggest that the subjective sense of causality would emerge only if x explains y, for example, in the sense of reducing its Kolmogorov complexity. We also discuss the relationship of explanation to predictability as sources of the subjective sense of causality in more general set-ups.

CAUSALITÉ SUBJECTIVE

Dans quelles conditions les gens ont-ils tendance à croire qu'un phénomène x a été cause d'un phénomène ultérieur y ? Nous suggérons ici qu'un sens subjectif de causalité peut émerger si x explique y en tant qu'il correspond à une réduction par exemple de sa complexité au sens de Kolmogorov. Nous étudions également dans des configurations plus générales la relation existant entre l'explication et la prévisibilité entendues comme sources de ce sens subjectif de causalité.

Keywords: causality, Kolmogorov complexity

Mots clés : causalité, complexité de Kolmogorov

JEL Codes: D80.

INTRODUCTION

Scientists and laypeople alike reason about causes of past events and use such reasoning in making predictions about future eventualities. For example, there are various theories regarding the causes of the subprime crisis of 2007–2008; and the policy of quantitative easing is believed to be among the causes that the crisis

* Tel-Aviv University. *Correspondence:* Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 6997801, Israel. *Email:* yotam.alexander@gmail.com

** Tel-Aviv University and HEC, Paris. *Correspondence:* Berglas School of Economics, Tel Aviv University, Tel Aviv 6997801, Israel. *Email:* tzachigilboa@gmail.com

This paper is based on the first author's MA thesis at the Berglas School of Economics, Tel-Aviv University, under the supervision of the second author. We thank participants of the LMU-TAU-VIU Seminar on Deliberation and Prediction in Decision Theory and Rational Choice (Venice, 2019) for comments and references. Gilboa gratefully acknowledges ISF Grants 1077/17 and 1443/20, the Investissements d'Avenir ANR-11-IDEX-0003/Labex ECODEC No. ANR-11-LABX-0047, the AXA Chair for Decision Sciences at HEC and the Foerder Institute at Tel-Aviv University.

didn't develop into a depression. Similarly, the exact extent of global warming is controversial, as are its causes. Yet the overwhelming majority of climate scientists contend that human activities are among the causes of the phenomenon. And, to consider less controversial examples, people may agree that real estate prices rise because of an increase in demand, or that a social security system is in danger because of low rates of fertility. Causal reasoning seems natural and ubiquitous. Furthermore, it has direct implications for decision making, whether by individuals or institutions, firms or governments.

However, ever since Hume [1738] questioned the concept, causality has proven elusive to define and to establish. Is there causality in the objective world around us? If so, can we grasp it, or prove its existence? Or is causality just a habit that we conveniently maintain? Is it a relic of the past, erroneously supposed to do no harm, as suggested by Russell, likening causality to the monarchy (Russell [1918], 180)? Or can it be defined in a rigorous way, as suggested by Mackie [1965] and Lewis [1973]? The philosophical debate is as active as ever, regarding metaphysical, epistemic, as well as other notions of causality. (See Illari and Russo [2014] for a recent survey of approaches and questions.) The debate is particularly relevant to economics, as pointed out by Mongin [2002]. As exemplified by the financial crisis of 2007–2008, the ongoing debate on climate change, as well as the more recent discussions of the ways of coping with COVID-19, governments need to make crucial decisions in set-ups that do not allow experimentation. In such situations the definition of causality is not only a matter of philosophical interest; it is a profound philosophical question that can have a palpable impact on people's lives.

Causality is a problematic concept also in statistics and logic. Students in basic classes are warned not to assume that correlation is necessarily indicative of causation. Establishing causal linkages between variables is indeed a major challenge of empirical work. Wright [1921] pioneered path analysis, attempting to uncover causal relationships in the absence of randomized control trials (Fisher [1935]). More recent contributions include Rubin [1974], Robins [1989], and Pearl [2000], the latter suggesting a general theory of causality.[1] The formal theory of causal and counterfactual reasoning has been further developed by Halpern [2016], building on joint work with Pearl.

Our interest is in *subjective causality*, referring to the psychological phenomenon of people thinking in causal terms. We focus on the sense of causality that is shared by many and is not obviously erroneous. The model is descriptive in nature; it is not designed to help people correct mistaken reasoning, such as confounding cause and effect, or missing hidden causes. At the same time, we do not focus on causal reasoning that is clearly faulty.

While our main motivation is to understand how people think about events in economic, political, and social life, it will be useful to consider also simpler examples that tend to be repeated under more or less the same conditions. Consider the following classical thought experiment, well known in the philosophical literature: a brick is thrown at a windowpane. If the window breaks after being hit by the brick, most people would tend to think that the brick was the cause of the window breaking. This has to do with the facts that (i) both events happened,

1. Pearl's theory has implications to formal decision theory. See, for instance, Zhang and Bareinboim [2017].

(ii) we tend to believe that, had the brick not been thrown, the window would not have been broken, and (iii) the presumed cause preceded the presumed effect. These three conditions can be traced back to Hume [1738] and Lewis [1973] and they seem necessary for subjective causality to emerge. However, they are not sufficient. For example, when we observe day being followed by night, and night by day, we do not think that the day is the cause of night, nor vice versa.[2] As in the case of the broken window, the three conditions seem to hold: we believe that, if it is daytime now, it will be nighttime 12 hours hence, and that, had it not been daytime now, it would not be nighttime 12 hours hence. It seems natural to ask, then, why don't we have a sense of causality in the case of the diurnal cycle as we do in the case of the broken window?

In this note we focus on an additional condition that we consider to be necessary for subjective causality: for x to be considered a cause of y, it has to explain it, in the sense of *reducing its complexity*. We assume that an agent has past observations, in each of which each of x and y did or did not occur. We suggest measuring the strength of the sense of causality that the agent would experience by the reduction in the Kolmogorov complexity (Kolmogorov [1963], [1965]) of y provided by x. This notion is in line with Janzing and Scholkopf [2010], Budhathoki and Vreeken [2016], [2017], and Marx and Vreeken [2019] who use similar Kolmogorov complexity measures for the definition of causality, mostly in a statistical context.

Our goal is to capture the degree to which people tend to think that x is the cause of y in this non-probabilistic context.[3] Specifically, in the case of the brick breaking the windowpane, it is generally hard to predict which windowpanes will break and when. Thus, the pattern of breaking windowpanes is rather complex; but when thrown bricks are introduced into the picture, it becomes much simpler. By contrast, in the case of the diurnal cycle, the pattern of days (or nights) is simple to begin with. If one needs to predict the pattern of nights, one can do it just as well with and without the occurrence of day as an explanatory variable. Thus, the variable x (day) does not help to reduce the complexity of y (night).

Observe that a variable x might fail to reduce the complexity of another variable, y, in two extreme cases. First, it might be the case that the two variables exhibit complex patterns that are unrelated to each other.[4] Second, as in the diurnal cycle example, y might be simple enough to begin with. In the former case, a person is likely to feel that y calls for an explanation, but that x fails to deliver it. In the latter, a reasonable person would feel that there is nothing to explain to begin with. Our focus on the *reduction* of complexity implies that, for a person to feel that x caused y, there should be a problem (namely, the complexity of y), and a solution (the reduction of complexity thanks to x). We argue that the sense of subjective causality can only emerge when there is this pattern of question-answer.

2. This is a well-known example (see, for instance, Illari and Russo [2014], 164). Notice, however, that the puzzle for us is the psychological phenomenon. We claim that an agent who would only observe days followed by nights and vice versa will not have a sense that one causes the other.

3. Janzing and Scholkopf [2010] also apply their concept to single events in a non-probabilistic context.

4. This case obviously doesn't satisfy the classical conditions (i)-(iii) above.

The notion of explanation is obviously related to prediction.[5] Typically, an explanation of past data suggests a way to predict future observations; and, vice versa, a theory that could have predicted the data in the past can be viewed as an explanation thereof. Yet, there are some distinctions between the two. First, an explanation, as measured by reduction in Kolmogorov complexity, deals only with the complexity of describing a program, and not of executing it. In case of high computational complexity one may have an explanation that is simple to describe but that does not allow for actual prediction. Second, when an agent considers her own choices, the sense of agency involves the feeling that her current decision cannot be predicted (even if past decisions are easily explained). We therefore believe that a more complete definition of subjective causality would involve a theory of predictability. We do not offer a formal model of predictability here for the sake of simplicity: such a model would be rather cumbersome, and involve many assumptions about the formulation of beliefs that may appear ad hoc.

The note is organized as follows. The first section conveys the main message. We first define our problem more precisely, distinguishing it from other problems related to causality that have received much attention in philosophy, statistics, and psychology. We then explain in more detail why the basic necessary conditions for causality are insufficient for our purposes, and informally describe the explanation condition. The second section provides a formal definition of subjective causality in a simplified model. We discuss predictability and possible formal models thereof in the third section. In particular, we discuss the effects of agency and computational complexity on the ability to predict and on the emergence of subjective causality.

THE QUESTION

What Is “Subjective Causality”?

This paper deals with causal reasoning as a rational psychological phenomenon. We ask, under which conditions will agents espouse causal theories, without being wrong in any obvious way. As this question is closely related to the discussions of causality in philosophy, statistics, and psychology, it will be helpful to focus the discussion in comparison with these fields.

As opposed to much of the discussion in philosophy and in statistics, we do not ask whether causality truly exists, or when it can be established based on empirical data, but only when it will be adopted by humans in their reasoning. Thus, our discussion is more positive than normative in nature. On the other hand, we do not focus on erroneous causal reasoning. Certainly, many causal theories may be mistaken, and people undoubtedly make mistakes, such as confounding causation with correlation. (See, for example, Spiegler [2020a], [2020b] for economic implications of such errors.) Yet, we aim to better understand causal theories in those cases where they do make sense, and where experts would endorse them,

5. See Hempel and Oppenheim [1948], Helmer and Rescher [1959], and, more recently, Shmueli [2010].

as in the cases of the financial crisis or global warming cited above. Our question therefore has a greater overlap with those asked in philosophy and statistics than do studies in psychology or behavioral economics.[6]

To define our question more sharply, one can imagine two tests of causal claims. First, one may adopt a "strong artificial intelligence" motivation and ask, if we were to program a machine, and would like it to use causal statements in a way that humans do (and that other humans find appropriate), when would we program the machine to utter causal statements? (See Pearl and MacKenzie [2018] for a similar motivation.) Alternatively, one can use a social definition, asking, if people disagree on causal statements, when would one be able to convince another of such a statement? Under which conditions would the majority of listeners agree that x was indeed the cause of y? Importantly, our motivation is neither to engineer machines that pass Turing tests, nor to train debaters. We use both illustrations merely as metaphors in order to define the problem we wish to study: understanding the way people use the term "causality," when they do so in ways that they, and others, would not find fundamentally mistaken. (See Sloman [2005] and Sloman and Lagnado [2015] for a similar approach.)

An important distinction between our main question and the study of causality in statistics is that we wish the theory to apply to singular public events, such as financial crises, wars, etc. While it is desirable to have a theory that can also scale up to large numbers of repetitions, and merge with statistical theories of causation, this isn't our main focus. As mentioned in the introduction, we will use simple examples in which there are many observations of phenomena that are "repeated under the same conditions." These, however, are only benchmark examples that are used to test our definitions in simple cases.

Our problem differs from much of the literature on causality in statistics in two ways. On the one hand, our problem is conceptually simpler because we assume that events are known as soon as they occur, and thus temporal order is also observable. This is often not the case in statistical problems where variables may interact repeatedly over time. For example, observing a correlation between price and quantity, one cannot determine temporal precedence. Neither variable takes values at a given observable time, and each may causally affect the other. Hence, even if one assumes a direct causal relationship between them, its direction is hard to determine. By contrast, when asking whether the assassination in Sarajevo was the cause of WWI, one has to deal with many difficulties, but temporal precedence is not one of them. On the other hand, our problem may be conceptually more problematic than statistical causality due to paucity of data. As the same example illustrates, often probabilities cannot be assumed given. There is only one observation of a history in which the assassination in Sarajevo did take place, and none in which it didn't. One may formulate probabilistic beliefs about the occurrence of WWI conditional on the assassination occurring or not, but these would have to be subjective beliefs, and determining them might require causal reasoning again. Further, one may describe one's beliefs using more flexible models than probability theory. In any event, the language of conditional probabilities isn't the most natural way to describe many of our problems.

6. In the categorization of questions about causality of Illari and Russo ([2014], 237–240), ours is closest to the question of semantics.

Preliminary Conditions

When will we say that a person thinks in causal terms? Or, what is meant by a person who says that "x is the cause of y?" The answer is, in principle, an empirical one. If we wonder what people mean by a certain claim, we should find out what beliefs are associated with this claim, and the answer should be obtained by collecting data on usage of the term and the associated beliefs. In planning such data collection, one should be willing to accept answers that are somewhat messy. A person's use of the term "cause" might well vary with her education, culture, and so forth. There is also no reason to assume that a given person uses the term in a consistent way; indeed, a person may use the term in ways that she will reconsider and find inappropriate. We do not offer any such empirical test here. The following should therefore be viewed merely as a yet-untested conjecture about empirical facts, which, at this point, we submit to the reader's intuitive judgment in lieu of a rigorous scientific study.

As mentioned in the introduction, three conditions that appeared in the philosophical literature (Hume [1738]; Lewis [1973]) are obvious candidates for necessary conditions for subjective causality. Note that these conditions were presented as defining what causality actually means, which is not our problem here. However, we re-interpret them as conditions that, by and large, are needed for *people to endorse* causal statements. Specifically, we submit that most people would not endorse the claim "x was the cause of y" unless they think that (i) x occurred, and so did y; (ii) had x not occurred, y would not have occurred either; (iii) x preceded y.

Clearly, statement (ii) is a counterfactual, and its meaning, as such, is open to debate. What do people mean when they utter such counterfactuals? If there are many repetitions of the same circumstances, counterfactuals typically refer to past cases in which the antecedent did not hold. For example, consider the statement "The window broke because it was hit by a brick." It involves the counterfactual claim, "Had the brick not hit the window, the window would not have broken." This latter claim is considered valid because it refers to many cases in which stones did not hit windows and the latter were not shattered. However, causal statements, and the implied counterfactuals, are also used when such past cases do not readily suggest themselves. For example, consider again the statement, "WWI erupted because of the assassination in Sarajevo." Should this causal claim satisfy condition (ii), it should also mean that, "Had the Archduke Ferdinand not been assassinated, WWI would not have erupted." We argued above that empirical data do not equip us with "objective" conditional probabilities for such events. This raises the question, what does a person mean when they espouse such a counterfactual? How can one person convince another that the counterfactual is or is not valid?

In this paper we do not delve into the question of counterfactual reasoning. We assume that prediction—of y given the actual x, as well as given the counterfactual not-x—is performed in some agreed-upon way, and turn our attention to a different problem: why is it the case that some cases that satisfy (i)-(iii) do not seem appropriate for causal statements? Going back to the diurnal cycle example, it is true that it will be night at time $t + 12$ (y_t) if and only if it is day at time t (x_t), and that the latter precedes the former. Why would it sound inappropriate to attribute causality to these phenomena? What is missing in (i)-(iii)?

A MODEL OF SUBJECTIVE CAUSALITY

We present a formal model that can capture some main features of subjective causality. We seek a model that is as simple as possible for two reasons. First, a simple model will clarify the main message, and show its basic logic in a transparent way. Second, we wish to think of the model as describing the way people reason. As mentioned above, we do not focus on biases and errors of human reasoning; rather, we attempt to capture the type of causal attributions that would be considered reasonable by most people. However, given that constraint, the simpler the model, the more convincing it is as a description of human thinking. For the sake of simplicity we also assume away many issues that are important and interesting but not strictly necessary to clarify the message. In particular, we ignore the question of counterfactual theories, despite its centrality to causal thinking.[7]

Assume that at time $t = 0, 1, \ldots$ an agent observes realizations of two variables, first x_t and then y_t. Let us assume that both are binary: $x_t, y_t \in \{0, 1\}$. Given a history $((x_i, y_i))_{0 \le i \le t}$ when will the agent feel that x is the cause of y?

Kolmogorov complexity and minimum description length (MDL) have already been used to define causality by Janzing and Scholkopf [2010], Budhathoki and Vreeken [2016], [2017] and Marx and Vreeken [2019], typically in a statistical context.[8] We follow here a similar path.

Given a history $h_t = ((x_i, y_i))_{0 \le i \le t}$ with $t > 0$,[9] let $K(y)$ $(= K(y, h_t))$ be the length of the shortest program that can generate $(y_i)_{0 \le i \le t}$. One needs to specify a formal language in which the programs are written, and while this choice is immaterial for the main point, we can, for concreteness, choose PASCAL as such a formal language, where the length of a program is measured in the number of tokens it uses. We consider programs that, for every $i \le t$, when accepting i as an input, compute y_i in finite time.[10] We similarly define $K(x)$ $(= K(x, h_t))$. Next, we define the minimum description length of y given x, $K(y \mid x)$ $(= K(x, h_t))$, as above, only now the program that needs to compute y_i for a given i can also consult the sequence $(x_j)_{0 \le j \le t}$ (that is, the program can invoke a procedure that computes x_j for any $j \le i$. Importantly, the program that computes y given x need not recall or compute the x's and can use them at no cost (as far as K is concerned). We view these programs as models of the way people think. We do not assume that people are necessarily aware of Kolmogorov complexity arguments, and certainly not that they can compute the minimal complexity program for any input.[11] We only

7. The model presented here can be extended to deal with counterfactuals, for example along the lines of Di Tillio, Gilboa and Samuelson [2013]. They model counterfactual reasoning employing the "unified model of induction" of Gilboa, Samuelson, and Schmeidler [2013]. While a more complete theory of subjective causality would have to rely on a theory of (subjective) counterfactuals, we do not explicitly model these here.

8. Solomonoff [1964a], [1964b] already used Kolmogorov's complexity measure as a model of human reasoning in the context of philosophy of science. Gilboa [1994] applied the concept for modeling everyday reasoning, including rather mundane examples such as "understanding a movie." The current use of MDL's is in line with these applications.

9. It is convenient to rule out the empty history so that all MDL's involved are strictly positive.

10. This definition depends on the exact notion of "tokens" and the way their complexity is measured. For example, is a natural number n considered to be a single token of length 1? Or should it be described by $\log(n)$ bits? The main points we wish to make do not depend on these details.

11. In general, this task is not computable.

suggest that, in many cases of interest, the complexity of programs in a language such as PASCAL can be a reasonable model of explanations that people can come up with, and of the way they rank some explanations as more convincing than others.

We mention without proof:

OBSERVATION 1. *There exists a c >0 such that, for any* $((x_i, y_i))_{0 \le i \le t}$,

$$K(x) + K(y \mid x) + c \ge K(y) \ge K(y \mid x).$$

We define the *measure of subjective causality* (for x being the cause of y given history h_t) to be

$$C(x, y) = 1 - \frac{K(y \mid x)}{K(y)}.$$

Given Observation 1, $0 \le C(x, y) \le 1$.[12] We argue that it captures the intuitive notion of causality. Consider the following three cases:

(i) Suppose that x denotes the occurrence of rain, and y, the lawn being wet. Assume also that the lawn is only wet when there is rain, so that there is a simple functional relationship by which $y_i = x_i$. Thus, the complexity of y given x is low. However, the patterns of both x and y are rather complex. Let us assume that both seem " random" in the sense that they have high Kolmogorov complexity. Hence,

$$K(x), K(y) \approx t$$

$$K(y \mid x) = c,$$

for some (low) constant c. It follows that $\frac{K(y \mid x)}{K(y)} \approx 0$ and $C(x, y) \approx 1$. Indeed, in this case an agent would tend to think that rain is the cause of the lawn being wet.

(ii) Suppose next that x is the occurrence of rain in one location, and y, the lawn being wet in a remote location. In this case both phenomena are complex, but x does not help much in predicting y. We could expect

$$K(x), K(y), K(y \mid x) \approx t$$

and thus $\frac{K(y \mid x)}{K(y)} \approx 1$ and $C(x, y) \approx 0$. Since x does not contribute much to explaining y, it is natural to assume that x isn't the cause of y.

(iii) Finally, assume that x denotes daytime, and y, daytime 12 hours later. In this case there is a simple functional relationship between x and y, with $y_i = 1 - x_i$. But each of x and y is also simple on its own, so that

$$K(x), K(y), K(y \mid x) \le c$$

12. The expression is similar to the definition of $d_s(x, y)$ in Janzing and Scholkopf [2010] (see p. 10). There are a number of differences in technical details and in motivation. In particular, our definition does not refer to the minimal description of the strings, which are not generally computable.

for a low c. Because $K(y) \leq c$ we find that $\frac{K(y \mid x)}{K(y)}$ cannot be too low. For concreteness, if $K(y) = K(y \mid x) = c$ we obtain $C(x, y) = 0$.

In case (ii) x isn't perceived to be the cause of y as it leaves y as complex as it was before knowing x: there is a lot to explain, but x doesn't explain much. By contrast, in case (iii) x isn't perceived to be the cause of y simply because (all puns intended) there was nothing to explain to begin with. Only in case (i) does subjective causality emerge, because there is a lot to explain a priori, but much less so after x is taken into account.

Our definition of $C(x, y)$ might bring to mind the "coefficient of determination," in linear regression analysis, defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

where SST stands for the overall variation to be explained (the variation of y irrespective of x), the SSR for the explained variation, and the SSE for the unexplained variation. Thus, the unexplained variation, SSE, plays a similar role to $K(y \mid x)$ in our definition, as it measures the amount of explanation needed after having taken x into account. Yet, the definitions vary in many ways, and, in particular, the function C need not be symmetric (see below), while R^2 is.

A Few Properties

We briefly mention a few properties of the definition above.

Causes Need to Be Complex

Observe that the definition of $C(x, y)$ does not make explicit reference to the complexity of x. However, the measures are related by Observation 1. Specifically,

> Observation 2. *There exists a* $c \geq 0$ *such that, for any* $((x_i, y_i))_{0 \leq i \leq t}$,
>
> $$C(x,y) \leq \frac{K(x)}{K(x) + K(y \mid x) + c} \leq \frac{1}{1 + \frac{K(y \mid x)}{K(x)}}.$$

Thus, the measure of subjective causality has an upper bound that depends on the complexity of x (and of y). If x is a very simple phenomenon, it will not generate a strong sense of causality.

Relative Frequency of Subjective Causality

Let us restrict attention to the event on which y is a function of x, say $y_i = x_i$ for all $i \leq t$. While we do not make any assumptions about a data generating process that assigns probabilities to the realizations of x and y, it is worthwhile to note that, over this event (that is, if $y = x$), most sequences $((x_i, y_i))_{0 \leq i \leq t}$ would result in a high measure of subjective causality, because most sequences are complex. If we let t tend to ∞, it is straightforward that in most states in this event, $C(x, y)$ will converge to 1.

Symmetry

Our main interest is in phenomena where temporal precedence is observable, and thus there is no room to ask whether x is the cause of y or vice versa. Yet, given a sequence $((x_i, y_i))_{0 \le i \le t}$, the question of symmetry is a natural and mathematically well defined one.

It is easy to see, however, that the measure of subjective causality, C, is not symmetric. For example, let t be large (say, tend to ∞) and consider a sequence $(y_i)_{i \le t}$ that is complex, with Kolmogorov complexity approaching t. Next define $(x_i)_{i \le t}$ to be 0 on odd periods i and $x_i = y_i$ for even i. Then $K(y) \approx t$ and $K(x) \approx t/2$. However, $K(y \mid x) \approx t/2$ while $K(x \mid y) \approx 0$. Thus,

$$C(x,y) = 1 - \frac{K(y \mid x)}{K(y)} \approx 1 - \frac{t/2}{t} = \frac{1}{2},$$

but

$$C(y,x) = 1 - \frac{K(x \mid y)}{K(x)} \approx 1 - \frac{0}{t/2} = 1.$$

Intuitively, y contains enough information to explain all of the complexity x, while x can only explain half of the complexity of y.

Transitivity

Assume that the agent is aware of three (binary) variables, x, y, z, and observes their values $((x_i, y_i, z_i))_{0 \le i \le t}$, for a large t. If x generates a sense of causing y, and y, of causing z, will the agent tend to feel that x is a cause of z?

A qualified answer is the affirmative follows from simple calculations. Observe that, for a constant c,

$$K(z \mid x) \le K(y \mid x) + K(z \mid y) + c.$$

Indeed, one way in which z_i can be computed, having access to $(x_j)_{j \le t}$, is by computing z_i using past values of y, $(y_i)_{j \le i}$, and, whenever y_i is needed, invoking $(x_l)_{l \le j}$ to compute it. The total length of such a program is bounded by $K(y \mid x) + K(z \mid y)$ and some additive constant (independent of t). We thus obtain

$$\frac{K(z \mid x)}{K(z)} \le \frac{K(y \mid x)}{K(z)} + \frac{K(z \mid y)}{K(z)} + \frac{c}{K(z)}.$$

If z is complex enough, say, $K(z) \ge K(y)$, $\frac{K(y \mid x)}{K(z)} \le \frac{K(y \mid x)}{K(y)}$, we also obtain

$$\frac{K(z \mid x)}{K(z)} \le \frac{K(y \mid x)}{K(y)} + \frac{K(z \mid y)}{K(z)} + \frac{c}{K(z)},$$

and then, if $\frac{K(y|x)}{K(y)}, \frac{K(z|y)}{K(z)} \approx 0$, so will be $\frac{K(z|x)}{K(z)}$. That is, a strong sense of causality between x and y and between y and z will imply a strong sense of causality between x and z. However, this need not follow if z is not complex. In particular, some of the classical examples of intransitive causation (Hitchcock [2001]) would be, in our set-up, examples where z is simple.[13]

PREDICTABILITY

Uncertainty about the Antecedent

We argued that statements (i)-(iii) in subsection "Preliminary Conditions" are not sufficient for the subjective sense of causality to emerge, and suggested to add a condition about the degree to which x explains y, as captured by the reduction of the latter's Kolmogorov complexity. Observation 2 showed that this reduction in complexity is related to the complexity of x itself: if x were simple, it could not generate a strong sense of causality.

This conclusion could be used as another way to supplement statements (i)-(iii) in the quest for a set of necessary conditions for the sense of causality to emerge: rather than discussing the complexity of sequences of observations, one could say that x could not be predicted. When a person says "x was the cause of y", she could be understood as saying, "I know that x did actually happen, but I don't think it *had* to happen. Before observing x and y, I didn't know x. Or, to be precise, my confidence in the conditional statement, that y would occur if and only if x would, by far exceeded my confidence in the unconditional statement that x would occur."[14]

Let us first see that this additional condition allows us to distinguish between the broken window and the diurnal cycle examples. In the former, x stands for "A brick is thrown at the window", and y, "The window breaks." It stands to reason that an agent would say, "I am rather sure that the window will break if a brick is thrown, but who am I to tell whether a brick would be thrown at it?" By contrast, in the diurnal cycle example x stands for "It is daytime at time t" and y for "It is nighttime at time $t + 12$" (where t is measured in hours). Again, the agent has very high confidence in predicting y given x. However, she also has

13. For example, if an assassin-in-training may not shoot (see Hitchcock [2001], 276), but he is backed up by a Supervisor who shoots in case he doesn't, Victim dies in any event, and there is little complexity to explain in z.

14. More formally, consider a state space $S = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ where, for each $s = (s_X, s_Y) \in S$, s_X denotes the truth value of X and s_Y, of Y. Then, the pair (s, F) with $s = (1, 1)$ and $F = \{(0, 0), (1, 1)\}$, can be read as "I know that S happened to have occurred. However, S was not the only possible state. What was necessarily true is only F, that is, Y would have assumed the same value as X whether it were 1 (as actually happened) or 0." Going back in time, before either X or Y were observed, the statement "X is the cause of Y" could be read as saying "I know F, but not more than that." Thus, we argue that the sense of subjective causality is strongest when the event F is believed to be true, while the value of X is unknown. That is, uncertainty about the antecedent is essential for the phenomenon to occur.

very high confidence in her prediction of x itself. Thus, if the sense of causality is defined as some measure that compares the two, it will not be as high in the diurnal cycle example as in the broken window example.

The sense of predictability can capture subjective causality, or absence thereof, also in cases that do not involve sequences of observations. For example, consider reasoning about mathematics. A somewhat naive model of mathematics would involve no uncertainty: mathematical statements are true or false. Before solving a mathematical problem, an agent might have subjective uncertainty about it, and entertain more than one state of the world in her mind. Yet, if she finds that x is true and so is y, she would typically not have a sense that x is the cause of y. They are simply both true. One way to capture this absence of subjective causality would be by the notion of predictability: even though the agent did not know that x was true, she has a sense that this could be known, that is, that x could have been predicted. As a result, no sense of causality emerges when her subjective state of knowledge changes.

Free Will

Our discussion suggests that subjective causality would emerge only when the antecedent x cannot be predicted, due to the absence of information, or to its complexity. There is another reason for which x might be unpredictable: the sense of agency and a consistent account of rational decision making. As argued in Gilboa [1999], [2009], rationality demands that, while making a decision, the decision maker put on hold whatever she knows or believes about that very decision. In order to consider all possible choices, a decision maker should be able to imagine, for each possible choice a, a coherent world in which she indeed chose a, and reason her way to the possible outcomes of that choice. It might be the case that an outside observer can predict that the choice be a, based on data about past choices. A rational decision maker should also be able to make this prediction. But we would probably not regard her as rational if she were incapable of imagining a world in which she chooses $b \neq a$. And to imagine such a world, she needs to suspend her belief that her choice is going to be a.

For example, a decision maker may be asked to contemplate jumping out of the window from a high floor. Knowing the decision maker's past behavior, we can probably predict that she is unlikely to jump. Indeed, so can she. However, in order to make a reasoned choice, the decision maker should be able to imagine the gruesome outcome of jumping, adding the statement "I jump out of the window at time t" to her knowledge base. If the contrary of that statement is included in this knowledge base, the latter will be inconsistent.[15] Thus, a rational account of the process of choice has to involve a step in which whatever the decision maker knows about the specific choice she is about to make is suspended. She may and probably should maintain whatever information she has about her past choices and about her future agents' choices, but not about the current one. This operation

15. It is claimed that the basic problem of understanding free will exists in such an example, and need not rely on determinism of any type: it suffices that an agent has a subjective sense of free will regarding *one specific instance of choice*, as well as practical knowledge of her own choice in that instance, to raise the problem.

Revue économique – vol. 74, n° 4, juillet 2023, p. 1-15

of "suspending" one's knowledge about oneself is reminiscent of Pearl's [2000] "do" operator and his model of counterfactual reasoning (see also Zhang and Bareinboim [2017]).

We therefore find that there are two main reasons for which an agent, who can predict y given x with a high degree of confidence, may feel that x is the cause of y: either she has uncertainty about x, as a result of absence of information or of computational complexity, or because x describes her own choice, in which case rational choice demands that she pretend that she does not know the value of x even if she practically does (see also Sloman [2005]).

When Are Variables Determined?

There are many problems in which one believes that x is the cause of y while observing x later than y, or not observing x at all. Indeed, when one infers causes from effects one believes that the value of x has been determined at a given time point, even though one does not (yet) know it.

We can think of an agent believing that a variable has been "determined" if the agent knows its value, or thinks that she *could possibly have known* this value. For example, an agent might know that a sports match is over, not know its outcome, but believe she could have known it had she been watching the match. Or, an agent might not have known a mathematical result at a given time, but feel that she could have known it, had she worked out the proof earlier.

It follows that the subjective sense of causality involves counterfactuals at two levels: the *substantive* level, which has to do with what would have happened in the external world (that is, what would have been the value of y had x assumed another value) and the *epistemic*, dealing with what the agent could have known.[16] We suggest that both types of counterfactuals are generated and justified in similar ways. A model of counterfactual reasoning (such as Di Tillio, Gilboa and Samuelson [2013]) can be applied to both types of counterfactual statements, in one case reflecting beliefs about the way the world works, and in the other, also about the way the agent herself reasons.

Observe also that the use of causal statements in everyday parlance can sometimes confound epistemic with substantive counterfactuals. Consider mathematical proofs again. We have argued that they admit no room for causation, at least in the naive model of mathematics. Yet, causal statements are not infrequent in mathematical proofs. But, we submit, these are statements about one's own state of knowledge, not about mathematics itself. If we prove that n is odd and write "therefore $(n + 1)$ is even," the causation refers to an epistemic counterfactual, not to a substantive one.[17]

16. This distinction is similar to, and in some cases precisely overlaps the distinction between indicative and subjunctive counterfactuals (see Stalnaker [1975]).

17. Substantive counterfactuals would emerge when one considers which axiomatic system to adopt. When one realizes that, say, adopting the "axiom of choice" is a matter of choice, one does entertain more than one state of the world also when thinking about mathematics.

REFERENCES

Budhathoki, K. and Vreeken, J. [2016]. “Causal Inference by Compression.” In Bonchi, F., Domingo-Ferrer, J., Baeza-Yates, R., Zhou, Z.-H. and Wu, X. (eds). *ICDM & ICDMW 2016: IEEE International Conference on Data Mining and Workshops, 12–15 December 2016, Barcelona, Catalonia, Spain*. Washington (D.C.): IEEE Computer Society, p. 41–50.

Budhathoki, K. and Vreeken, J. [2017]. “MDL for Causal Inference on Discrete Data.” In Gottumukkala, R., Ning, X., Dong, G., Raghavan, V., Aluru, S., Karypis, G., Miele, L. and Wu, X. (eds). *ICDM 2017 New Orleans: IEEE International Conference on Data Mining, 18–21 November 2017, New Orleans, Louisiana*. Washington (D.C.): IEEE Computer Society, p. 751–756.

Di Tillio, A., Gilboa, I. and Samuelson, L. [2013]. “The Predictive Role of Counterfactuals,” *Theory and Decision*, 74: 167–182.

Fisher, R. A. [1935]. *The Design of Experiments*. London: Oliver & Boyd.

Gilboa, I. [1994]. “Philosophical Applications of Kolmogorov’s Complexity Measure.” In Prawitz, D. and Westerstahl, D. (eds). *Logic and Philosophy of Science in Uppsala: Papers from the 9th International Congress of Logic, Methodology and Philosophy of Science*. Berlin: Kluwer Academic Press, p. 205-230.

Gilboa, I. [1999]. “Can Free Choice Be Known?” In Bicchieri, C., Jeffrey, R. and Skyrms, B. (eds.). *The Logic of Strategy*. Oxford: Oxford University Press, p. 163–174.

Gilboa, I. [2009]. *Theory of Decision under Uncertainty*. Cambridge: Cambridge University Press.

Gilboa, I., Samuelson, L. and Schmeidler, D. [2013]. “The Dynamics of Induction in a Unified Model,” *Journal of Economic Theory*, 148: 1399-1432.

Halpern, J. [2016]. *Actual Causality*. Cambridge (Mass.): MIT press.

Helmer, O. and Rescher, N. [1959]. “On the Epistemology of the Inexact Sciences,” *Management Science*, 5: 25–52.

Hempel, C. and Oppenheim, P. [1948]. “Studies in the Logic of Explanation,” *Philosophical Science*, 15: 135–175.

Hitchcock, C. [2001]. “The Intransitivity of Causation Revealed in Equations and Graphs,” *Journal of Philosophy*, 98: 273–299.

Hume, D. [1738]. *A Treatise of Human Nature*. Cambridge: Cambridge University Press, 2012.

Janzing, D. and Scholkopf, B. [2010]. “Causal Inference Using the Algorithmic Markov Condition,” *IEEE Transactions on Information Theory*, 56: 5168–5194.

Illari, P. and Russo, F. [2014]. *Causality: Philosophical Theory Meets Scientific Practice.* Oxford: Oxford University Press.

Kolmogorov, A. N. [1963]. “On Tables of Random Numbers,” *Sankhyā: The Indian Journal of Statistics, Series A*, 369–376.

Kolmogorov, A. N. [1965]. “Three Approaches to the Quantitative Definition of Information,” *Probability and Information Transmission*, 1 (1), 4–7.

Lewis, D. K. [1973]. “Causation,” *Journal of Philosophy*, 70 (17): 556–567.

Mackie, J. L. [1965]. “Causes and Conditions,” *American Philosophical Quarterly*, 12: 245–265.

Marx, A. and Vreeken, J. [2019]. “Telling Cause from Effect Using MDL-Based Local and Global Regression,” *Knowledge and Information Systems*, 60: 1277–1305.

Mongin, P. [2002]. “La conception deductive de l’explication scientifique et l’economie,” *Information sur les Science Sociales*, 41: 139–165.

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

Pearl, J. and MacKenzie, D. [2018]. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Robins, J. M. [1989]. “The Control of Confounding by Intermediate Variables,” *Statistics in Medicine*, 8: 679–701.

Rubin, D. B. [1974]. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 56: 688–701.

Russell, B. [1918]. "On the Notion of Cause." In *Mysticism and Logic*. New York: Dover Publications, p. 180–208.

Shmueli, G. [2010]. "To Explain or to Predict?" *Statistical Science*, 25: 289–310.

Sloman, S. A. [2005]. *Causal Models: How People Think about the World and Its Alternatives.* Oxford: Oxford University Press.

Sloman, S. A. and Lagnado, D. [2015]. "Causality in Thought," *Annual Review of Psychology*, 66: 223–247.

Solomonoff, R. [1964a]. "A Formal Theory of Inductive Inference I," *Information Control*, 7 (1): 1–22.

Solomonoff, R. [1964b]. "A Formal Theory of Inductive Inference II," *Information Control*, 7 (2): 224–254

Spiegler, R. [2020a]. "Behavioral Implications of Causal Misperceptions," *Annual Review of Economics*, 12: 81–106.

Spiegler, R. [2020b]. "Can Agents with Causal Misperceptions be Systematically Fooled?" *Journal of the European Economic Association*, 18 (2): 583–617.

Stalnaker, R. C. [1975]. "Indicative Conditionals," *Philosophia*, 5: 269–286.

Wright, S. [1921]. "Correlation and Causation," *Journal of Agricultural Research*, 7 (7): 557–585.

Zhang, J. and Bareinboim, E. [2017]. "Transfer Learning in Multi-Armed Bandits: A Causal Approach." In Sierra, C. (ed.). *Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia 19-25 August 2017*. International Joint Conferences on Artificial Intelligence, p. 1340–1346.