Title

Qualifying Causes as Pertinent

Permalink

https://escholarship.org/uc/item/66t1h4kr

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 40(0)

Authors

Sileno, Giovanni

Dessalles, Jean-Louis

Publication Date

2018

Qualifying Causes as Pertinent

Giovanni Sileno[1] **(giovanni.sileno@telecom-paristech.fr)**
Jean-Louis Dessalles[1] **(dessalles@telecom-paristech.fr)**
[1]LTCI, Télécom ParisTech, Université Paris-Saclay, 46 rue Barrault, 75013 Paris, France

Abstract

Several computational methods have been proposed to evaluate the relevance of an instantiated cause to an observed consequence. The paper reports on an experiment to investigate the adequacy of some of these methods as descriptors of human judgments about causal relevance.

Keywords: Actual Causation, Relevant Cause, Counterfactuals, Bayesian Inference, Relevance Theory, Simplicity Theory.

Introduction

Causes play a central role in the way we conceptualize the world. Seeking explanations of events, or, stated differently, attributing responsibility for their occurrence, is common in practically all human activities. Despite such widespread use, however, there is yet no established model about how people qualify a cause as *pertinent* (literally, *holding together*) to a specific event. As observed by (Glymour et al., 2010), most psychological literature focuses on judgments of *general causation* (about causal regularities), or, when investigating *actual causation* (about situational interpretation), it focuses on specific applications like the perception of causation and animacy in minimal "mechanical" settings—see e.g. the overview in (Rips, 2011)—or on higher-level tasks, like the attribution of moral responsibility. The logic-philosophical literature, for its part, focuses for the most on finding and incrementally resolving paradoxes on existing models, with little concerns about practical settings. A similar situation holds on the computational side, where several approaches compete—with mixed results—for producing human-like inferences about causation. In this context, this paper aims to assess the gap existing between theoretical models and empirical observations: it considers a short selection of methods that offer means to compute the relevance of causes and it investigates if these methods can produce outputs similar to people's responses collected in a dedicated experiment. For better decomposition, the study bypasses the natural language processing problem, and exploits a *domain model* for each task, expressed in forms that can be automatically processed by the methods to be evaluated.

The document consists of five sections: a brief overview of the frameworks used as a basis for the study, presentation of experiment, model, computation, and evaluation of results.

On causation and relevance

Counterfactuals The primary approach to actual causation, rooted in philosophy and logic, builds upon *counterfactuals*, a construct corresponding to the *but-for* test used in law: *"but-for" the event A, would the event B have occurred?* If not (or if it did), A did (or did not) cause B. In this form, it is well known that the method suffers from capturing all necessary elements for the generation to occur (without discriminating their relevance), and not the sufficient ones (cf. the *fire squad* case: if the sniper who killed hadn't shot, the victim would have died anyway). Despite this practical limitation, many formalizations of counterfactuals have been proposed in the last fifty years, attempting to find an unified account satisfying known and newly found paradoxes—see e.g. the famous work of (Lewis, 1973), based on *modal logic*, and the more recent account by (Hitchcock, 2001), based on *structural equations*. The interest of these methods resides in inferring, given a system of counterfactuals, other valid counterfactuals (except some paradoxical cases).

Bayesian inference Intuitively, part of the problem of the lack of sensitivity of counterfactuals may be due to determinism. Turning to probabilistic methods, and in particular Bayesian probability, a full research track investigates how to explain *why* certain variables are observed in certain states; see e.g. (Yuan, Lim, & Lu, 2011) for a reasoned overview.

Let us suppose to we are able to encode the domain model in a *Bayesian network* (BN). Amongst the proposed choices for computing the relevance of the occurrence of the event C to the occurrence of event E, the most commonly used are the likelihood $p(E|C)$, as in *maximum likelihood* (ML) estimation, or the product $p(E|C) \cdot p(C)$ as in *maximum a posteriori* (MAP) estimation. Alternatively, observing that the occurrence of a cause increases the possibility of occurrence of the effect, we could capture the *raise of probability* by computing differences as $p(E|C) - p(E)$ or $p(E|C) - p(E|\neg C)$, or ratios of these terms. By dealing with co-occurrences, these approaches can be seen as conflating causes to *evidential supports*. Interestingly, an experimental study on measures of evidential supports by (Tentori, Crupi, Bonini, & Osherson, 2007) finds empirical alignment of human responses with respect to two measures:[1]

$$\log \frac{p(E|C)}{p(E|\neg C)} \quad (1) \qquad\qquad \frac{p(E|C) - p(E|\neg C)}{p(E|C) + p(E|\neg C)} \quad (2)$$

Returning to the causal domain, to avoid undesired effects—like consequences that "cause" causes—additional machinery is required, e.g. introducing time-related constraints in the inferences—see e.g. (Williamson, 2009).

Alternatively, (Pearl, 2000) proves that there is strict connection between structural equations and Bayesian networks,

[1] From a theoretical point of view, however, the authors argue that (2) is the only one to satisfy desirable mathematical properties.

and proposes an unifying notation—*causal Bayesian networks* (CBNs)—introducing an explicit *do* notation to distinguish *interventions* from standard probabilistic events. An acknowledged problem with this method is that variables relevant to the computation may be unobservable (Zhang, 2008).

At a more fundamental level, however, the problem of *model adequacy*, i.e. of which variables to include in the model, concerns *all* these computational methods. Intuitively, a cognitive basis might offer a more robust solution.

Relevance Theory Relevance Theory (RT) identifies general principles that are supposed to govern successful communication. A statement is said to be relevant if the addressee is able to draw inferences from it. Since inferences may always be produced from any statement, RT prioritizes those which can be produced "effortlessly" (Sperber & Wilson, 1986). This principle is supposed to guide listeners in determining causal relations, as in the two following examples (Wilson & Sperber, 1998):

(i) *John dropped the glass. It broke.*
(ii) *I got caught. My best friend betrayed me.*

According to RT, causality is inferred because it enriches the context. In (i), the first statement is understood as a cause of the second one, because such a material relation is easy to access from experience. In (ii), the friend's betrayal could be the consequence of the speaker's having been caught. However, the converse causality is preferred because it is more "accessible" (and the speaker would have expressed things differently to mean otherwise). It has been observed that notions such as "effort" or "accessibility" are crucial for RT to make predictions and yet remain external to the theory (Levinson, 1989): there is a risk that RT's principles be bent to justify any intuitively correct interpretation *ex post facto*.

Simplicity Theory Simplicity Theory (ST) has been introduced to account for *interestingness*, and offers an alternative definition of relevance (Dessalles, 2013). Relevant events must be *unexpected*, which means that they can be presented as *more complex to generate than to describe*. Cognitive complexities of *generation* and of *description* are measured as *minimum description lengths* (MDL) (Chater, 1999). In particular, generation or world complexity (denoted with C_W), generalizes the classical notion of (im)probability, as it computes odds without using set extensions. Considering two events e_1 and e_2, the complexity of their *sequential composition* (denoted by '$*$') is (*chain rule*):

$$C_W(e_1 * e_2) = C_W(e_1) + C_W(e_2|e_1) \tag{3}$$

where $C_W(e_2|e_1)$ is the conditional complexity of generating e_2 considering e_1 already realized.

Neglecting effects due to description complexity, a relevant (tentative) explanation, according to ST, is any piece of knowledge that diminishes generation complexity. The potential causal contribution of c on e is captured by $C_W(e) - C_W(e * c)$. When dealing with *actual causation*, c is realized, so $C_W(c) = 0$. Thus, c is a relevant *actual cause* for e if the generation complexity C_W of e is smaller conditionally to c, i.e. $C_W(e|c) \ll C_W(e)$. A measure of relevance could then be:

$$C_W(e) - C_W(e|c) \tag{4}$$

Consider again example (ii). "I got caught" is relevant as far as it can be perceived as unexpected: a complex set of circumstances was supposed necessary for this outcome to occur. The second statement "My best friend betrayed me" appears as a relevant cause as far as it makes the minimal causal path to being caught significantly shorter. Note that (4) is more constraining than RT's principles. It states both that $C_W(e)$ is large and that c provokes *complexity drop* once taken into account. When several causes are offered, the most relevant one should be the cause that provokes the largest drop.

Experimental test

Participants are asked to read five passages of a short story and to rank the pertinence of answers to simple *why* questions according to the scale NR: irrelevant, 1: low relevance, 5: high relevance (same ranking allowed).[2] They are instructed to approach each answer as if it were the only answer produced by another locutor.

Passage 1 *Johnny is 7 years old. In recent months his mother has been worried because he developed a craving for sweet things. She bought some pots of strawberry jam and put them into the larder (a small room near the kitchen). Then one afternoon she finds that Johnny has gone into the larder and has eaten half a pot of strawberry jam.*

Q1. *Why is "half a pot of jam gone"?*
Q2. *Why did "Johnny eat the jam"?*
Q3. *Why did "Johnny go into the larder"?*

Each question had the same candidate answers:

a. because of Johnny's gluttony
b. because Johnny ate it
c. because mother has put the pot in the larder

save the second answer of Q3:

b'. because johnny wanted to eat a pot of jam being there

Passage 2 *The mother says: "That's naughty. In the future you are never to enter the larder without my permission." Several incidents then follow. First, Johnny gets a broom, hooks the pot of jam from above the shelf without entering into the larder and helps himself.*

Q4. *Why did "Johnny use the broom to hook the pot"?*

a. because Johnny wanted to take the pot
b. because of Johnny's gluttony
c. because mother has put the pot in the larder
d. because mother forbade Johnny to enter the larder
e. because the broom was in the house
f. because of gravity

[2] The story (the "legalistic child" case) is adapted from (Rissland & Skalak, 1989), in turn revisiting (Twining & Miers, 1982) .

Passage 3 *Mother finds Johnny eating the jam, but he says to her: "I didn't enter the larder". Then another day, the cat enters the larder and attacks the salmon which mother has bought for a special occasion.*

Q5. *Why did "the cat enter the larder and eat the salmon"?*

a. because the cat was hungry
b. because mother has put the salmon in the larder
c. because the salmon was in the larder

Passage 4 *Mother, upstairs, hears Johnny laughing. She comes down to see Johnny standing outside the larder door watching the cat eating the fish. 'I may not go into the larder' he says.*

Q6. *Why did "the cat enter the larder and eat the salmon"?*

a. because the cat was hungry
b. because mother has put the salmon in the larder
c. because the salmon was in the larder
d. because mother forbade Johnny to enter the larder
e. because Johnny made fun of mother

Passage 5 *Finally, Johnny's parents were out and Johnny was watched by his usually iron-willed babysitter, Maggie. Johnny's parents forgot to tell Maggie anything about dinner. Supper was late and Johnny was hungry. Johnny asked permission from the babysitter to enter the larder. She said OK. Johnny feasted on jam.* The questions were:

Q7. *Why did "Johnny enter the larder"?*
Q8. *Why did "Johnny feast on jam"?*

The candidate answers were for both:

a. because he was hungry
b. because of Johnny's gluttony
c. because Maggie granted him permission
d. because his parents forgot to tell Maggie about dinner
e. because supper was late

Domain model

Referring to the traditional terminology, each question specifies an *explanandum* (something which has to be explained), each answer proposes an *explanans* or explanation for the *explanandum*. The test implicitly builds upon three roles: X, the (virtual) person who asks the *why* question; Y, the (virtual) person who gives the answer; and Z, the (actual) respondent, who evaluates the response given by Y. In order to produce an answer or to evaluate its relevance, Z requires a model of the world in which the story has occurred. Such representation is not required to be isomorphic with the input in all details, but just to be an adequate synthesis.

General action-scheme All questions are about events—for the most, actions performed by an agent. Analyzing verbal reports of legal cases, (Pennington & Hastie, 1993) found that explanations of human behaviour in legal decision-making converge to the following *action-scheme*—here in the version proposed by (Bex & Verheij, 2011):

$$Motive \Rightarrow Intent \Rightarrow Action \Rightarrow Consequences$$

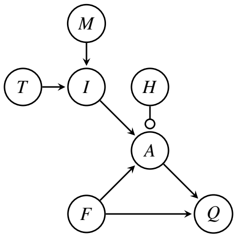

Figure 1: Action-scheme as general model of action

To cover further cases, we considered additional elements (Sileno, 2016, Ch. 7); *motive* is interpreted as a situation perceived by the agent, triggering a preexisting *motivation* and producing an *intent* (here in the sense of specific desire); the intent develops into a certain *action* (or a course of actions) if coupled with perceiving the associated *affordance* and no *inhibition* is put in place by other motivational components; the action brings about a certain *consequence* depending on the actual environmental *disposition*. The following model components are then considered:

M	motivation	T	motive
I	intent	F	affordance of I via A
A	action	H	inhibition of A
Q	consequences		

For instance, for a boy craving for sweets (M), the fact that there is a jam pot (T) "generates" a desire to eat that jam (I). If he is already able to eat it (F), he just does it (A). As a side effect, there will be less jam in the pot (Q). The relative dependencies of the model components are illustrated in the graph in Fig.1 (for simplicity perceptual with actual affordance are aligned). Inhibition, specified using an empty-circle arrow, is a *negative dependence*: the absence of the parent element enables the child element to occur.

Actions usually have consequences relevant to other actions. For instance the agent might need an additional action A' to bring about F (e.g. to go near the pot); in this case the new action should be added in in A's action model as a new element $\bullet F$, parent of F. Note however that the generation of the affordance might be independent from the agent; when the action is intentionally *preparatory* (i.e. part of a plan to perform A), $\bullet F$ depends on I as well. Small case letters will be used in case of ambiguities.

Passage 1 (Fig. 2) The central event for passage 1 is "Johnny eating the jam". Reading the propositional content provided in questions and answers through an action-scheme centered around this action, the following associations hold:

Q1 *"half a pot of jam is gone"* is a consequence (Q)
Q2, b *"Johnny eats the jam"* is the core action (A)
Q3 *"Johnny goes into the larder"* is a preparatory action to perform the core action ($\bullet F$)
a *"Johnny is gluttonous"* identifies the motivation (M)
c *"mother has put the pot in the larder"* is an event generating the motive starting the course of action ($\bullet T$)
b' *"Johnny wanted to eat the jam"* is the intent (I)

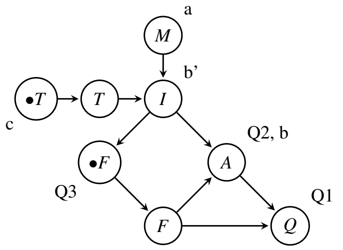

Figure 2: Model of passage 1

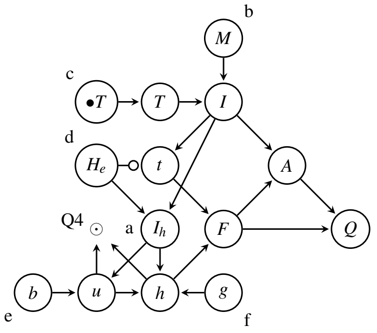

Figure 3: Model of passage 2

Passage 2 (Fig. 3) The easiest way to eat the jam would be to take it directly (t), however, the prohibition of the mother to enter (H_e) is inhibiting this easier action. Hooking the pot (h) is an alternative plan. Then, using the broom (u) to hook the pot (to eat the jam) is a nested preparatory action. The presence of broom (b) is a necessary condition to perform it, while gravity (g) is a necessary condition to hook the pot. Q4 ("use the broom to hook the pot") cannot be modeled simply as u, a possible option is the composite action $u * h$.

Passages 3 and 4 (Fig. 4) The core action is the cat eating the salmon. Passage 4 brings to the foreground the non-intervention of Johnny. The boy is normally (n) expected to stop the cat (s), overriding the prohibition issued by his mother ($\bullet H_e$). Making fun of her (f) is a possible motivation behind the anomaly. Q5 and Q6 are modeled as $\bullet F * A$.

Passage 5 (Fig. 5) The action is centered again around Johnny eating jam. A new motivational state is added: hunger (h)—caused by the supper being late (l), in turn a consequence of the lack of instructions by the parents (f). Hunger does not enter directly in the action scheme, but only behind the scenes, as the reason why Maggie gives permission (p) to enter the larder (e), thus overriding the prohibition.

Computing relevance

The models in the previous section serve as a common ground for a direct operationalization of the methods presented in the introduction (except for RT, as it does not specify the notion of "effort"; its comparative evaluation is then left as an open question).

Counterfactuals Applying informally the *but-for* test on the model of passage 1, all answers qualify as causes. Following the formalization given by (Hitchcock, 2001), each element of our example is modeled as a binary variable associated to the occurrence or the non-occurrence of the event. The associated set of *deterministic structural equations* would be:

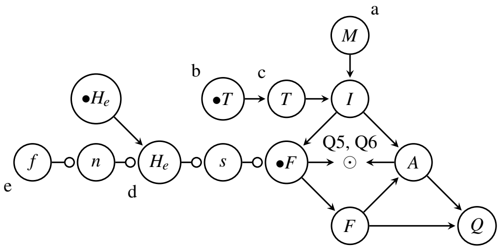

Figure 4: Model of passage 4 (without the inhibitors, 3)

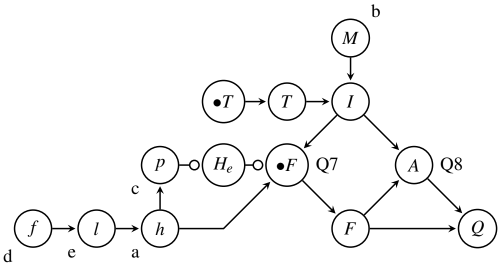

Figure 5: Model of passage 5

$$I := M \wedge T \qquad A := I \wedge F \qquad Q := A \wedge F$$
$$F := \bullet F \qquad T := \bullet T \qquad \bullet F := I$$

Each equation can be seen as encoding counterfactual information related to a causal dependence. Also in this case, applying Hitchcock's definition of *active route* for the determination of actual causes, we qualify positively all answers. Same results are obtained with the other passages.

Bayesian inference The application of Bayesian inference requires the domain model to be encoded in a Bayesian network. In principle, these graphs should be diagrammatically very similar to e.g. Fig. 2. A practical problem arises for deciding the parameters of the conditional probability tables. Even acknowledging *subjective probability* (i.e. capturing *degree* or *strength of belief*), it is not evident to provide solid backup for these numbers. Nevertheless, this is a required step to proceed with this method. With subjective estimations as parameters, we have extracted the relevance measures defined in the introduction, obtaining the results on Table 1.

Simplicity theory The *chain rule* formula (3) enables us to run through the models in search of the *shortest path from cause to effect*. Path lengths are measured by the sum of conditional complexities associated to the transitions and the complexity of nodes with no parents required to proceed in the path. Let us assume that all the dependencies belonging to the general action-scheme (Fig. 1) carry similar complexity of transition C_0. The graph implies:

$$C_W(I) = C_W(M) + C_W(T) + C_0$$
$$C_W(A) = C_W(F) + C_W(I) + C_W(\neg H) + C_0$$
$$C_W(Q) = C_W(F) + C_W(A|F) + C_0 = C_W(A) + C_0$$

	a	b	c	d	e	f	a	b	c	d	e	f
	$p(E\mid C)$						$p(E\mid C)\cdot p(C)$					
Q1	0.15	**0.53**	0.11				0.03	**0.05**	**0.05**			
Q2	0.16	**1**	0.11				0.03	**0.1**	0.05			
Q3	0.18	**0.9**	0.12				0.04	**0.06**	**0.06**			
Q4	**0.67**	0.13	0.11	0.12	0.11	0.11	0.06	0.03	0.05	0.07	0.1	**0.11**
Q5	**0.33**	0.18	0.19				0.07	**0.09**	**0.09**			
Q6	**0.94**	0.4	0.47	0.16	0.66		**0.31**	0.14	0.16	0.06	0.22	
Q7	0.14	**0.15**	0.14	0.12	0.13		**0.04**	0.03	**0.04**	**0.04**	**0.04**	
Q8	0.13	**0.16**	0.12	0.11	0.12		**0.04**	0.03	**0.04**	0.03	**0.04**	
	(1)						(2)					
Q1	0.83	**3.42**	0.23				0.28	**0.83**	0.08			
Q2	0.9	∞	0.26				0.3	**1.0**	0.09			
Q3	0.98	**4.17**	0.28				0.33	**0.89**	0.1			
Q4	**3.69**	0.42	0.11	0.54	1.01	nan	**0.86**	0.15	0.04	0.18	0.34	nan
Q5	**2.01**	1.01	1.18				**0.6**	0.34	0.39			
Q6	**0.11**	0.08	0.08	0.07	0.1		0.02	0.04	0.04	**0.06**	0.01	
Q7	**1.49**	1.28	1.47	1.08	1.27		**0.48**	0.42	0.47	0.36	0.41	
Q8	1.06	**1.34**	1.04	0.78	0.91		0.35	**0.43**	0.35	0.26	0.3	

Table 1: Relevance measures computed using Bayesian networks, with parameters: for passages 1, 2, 5, $p(\bullet T) = 0.5$, $p(M) = 0.2$, $p(\bullet|PT) = 0.2$; for passage 2, $p(H_e) = 0.6$, $p(b) = 0.9$, $p(g) = 1$; for passages 3 and 4, $p(M) = 0.2$, $p(\bullet T) = 0.5$; for passage 4, $p(f) = 0.1$, $p(H_e) = 0.8$; for passage 5, $p(f) = 0.3$.

	a	b	c
Q1	$C_W(M)$	$C_W(M)+C_W(\bullet T * T)+C_W(F \mid \bullet F)+3C_0$	$C_W(\bullet T)$
Q2	$C_W(M)$	0	$C_W(\bullet T)$
Q3	$C_W(M)$	$C_W(M)+C_W(\bullet T * T)+C_0 \equiv C_W(I)$	$C_W(\bullet T)$

	a	b	c	d	e	f
Q4	$C_W(I)+C_W(H_e)+C_0$	$C_W(M)$	$C_W(\bullet T)$	$C_W(H_e)$	$C_W(b)$	$C_W(g)$

	a	b	c	d	e
Q5	$C_W(M)$	$C_W(\bullet T)$	$C_W(\bullet T * T)$		
Q6	$C_W(M)$	$C_W(\bullet T)$	$C_W(\bullet T * T)$	$C_W(f)+C_W(\bullet H_e)+5C_0$	$C_W(f)$

	a	b	c	d	e
Q7	$C_W(f*l*h)$	$C_W(M)$	$C_W(f*l*h*p)$	$C_W(f)$	$C_W(f*l)$
Q8	$C_W(f*l*h)$	$C_W(M)$	$C_W(f*l*h*p)$	$C_W(f)$	$C_W(f*l)$

Table 2: Relevance strengths computed as drops of generation complexity (C_0 is the transition complexity of action-scheme dependencies).

and then constraints as: $C_W(I) > C_W(M)$ and $C_W(I) > C_W(T)$, etc. Turning upon the passage models[3] , by applying (4) we obtain the expressions reported on Table 2, from which we can extract similar constraints. Note how this analytical form does not require to decide the parameters upfront.

Results

Empirical results The participants to our experiment were 102 individuals (54% female, 71% age 31-50), mostly European researchers, recruited via social networks. The test was conducted online. Analyzing the responses, we initially quantified the ranking of answers from 0 (irrelevant, NR) up to 5 (highly relevant), as in the test. The average and standard deviation of rankings are reported on Table 3. These measures are however not necessarily the most illustrative for our study, as the resulting histograms have various shapes. We then considered ordering decreasingly the participants' rankings and reading the (minimal) ranking value attained by 51% or 75% of the population. This method enables us to associate to the selected value a *majority* (simple or qualified) for which that answer has *at least* that ranking of pertinence. Table 4 illustrates that the relative ordering of such minimal rankings is consistent passing from 51% to 75% of the population.

	a	b	c	d	e	f
Q1	3.5 ± 1.3	**4.5** ± 1.1	0.7 ± 1.1			
Q2	**4.5** ± 1.0	1.1 ± 1.8	1.1 ± 1.3			
Q3	3.1 ± 1.6	**3.8** ± 1.7	3.7 ± 1.5			
Q4	**4.4** ± 1.1	3.1 ± 1.6	2.1 ± 1.7	**4.4** ± 1.2	0.9 ± 1.2	0.7 ± 1.0
Q5	**4.1** ± 1.4	2.4 ± 1.8	3.5 ± 1.8			
Q6	**4.0** ± 1.5	2.8 ± 1.7	3.6 ± 1.7	0.6 ± 1.1	0.5 ± 1.0	
Q7	3.8 ± 1.6	2.9 ± 1.8	**4.2** ± 1.3	2.3 ± 1.5	2.5 ± 1.6	
Q8	3.6 ± 1.7	**4.2** ± 1.3	1.8 ± 1.7	1.9 ± 1.6	2.1 ± 1.7	

Table 3: Ranking attributed to answers: mean ± st.dev.

	a	b	c	d	e	f
Q1	4 (3)	**5** (**5**)	0 (0)			
Q2	**5** (**4**)	0 (0)	0 (0)			
Q3	3 (2)	**5** (**3**)	4 (**3**)			
Q4	**5** (**4**)	3 (2)	2 (1)	**5** (**4**)	0 (0)	0 (0)
Q5	**5** (**4**)	2 (1)	4 (2)			
Q6	**5** (**3**)	3 (1)	4 (**3**)	0 (0)	0 (0)	
Q7	**5** (3)	3 (1)	**5** (**4**)	2 (1)	3 (1)	
Q8	4 (3)	**5** (**4**)	1 (0)	2 (0)	2 (0)	

Table 4: Minimal rankings for simple (qualified) majorities.

We have also extracted the relative ordering of the rankings given by individual respondents for each question, in case there were cross-relations between answers that were lost by the previous analysis. Even if respondents were instructed to consider options as independent, one can indeed reasonably expect some repositioning effects due to the available choices. These relative orderings, reported on Table 5 (1 means ranked as most relevant by participants), are consistent with the previous results, although they lose information about the relative gap of pertinence between answers.

Comparative evaluation As we can see on the tables, no measure of Bayesian inference is fully consistent with our experiments; likelihood and (1) are more aligned, followed by (2). In all cases, we observe a pathological response for Q2b (tautological answer). In many cases, even if the most pertinent cause is correctly identified, the relative order between the answers does not follow the empirical results. This may be due to a wrong choice of parameters or even to wrong dependencies in the model. Unfortunately, the framework is quite opaque to model correction tasks.

In contrast, the relevance measure computed via ST are quite aligned to the empirical results, with fewer (constraints on) parameters and at inferior computational cost. The *irrelevance* of Q2b is correctly captured. The relative ranking of Q1 and Q2, in the plausible hypothesis that $C_W(M) \gg C_W(\bullet T)$ (the child being gluttonous vs putting jam in the larder), is correct. Q4 is also aligned: *a* is necessarily the most pertinent cause; *d* the second one, at the condition that $C_W(H_e)$ is sufficiently high (e.g. by considering a plausible dependency of H_e w.r.t. M); *b* and *c* are consistent with the previous ranking; *e* and *f* have low pertinence, because their complexity is very low. For Q4 and Q5, consistency holds if

[3] In a context in which A would occur if not inhibited by H: $C_W(A|\neg H) = C_0$, the double inhibition $B \multimap H \multimap A$ translates into $C_W(\neg H|B) = C_0 + C_W(B)$ and so $C_W(A) = 2C_0 + C_W(B)$.

	a	b	c	d	e	f
Q1	1.7 ± 0.5	**1.2** ± 0.5	2.8 ± 0.5			
Q2	**1.1** ± 0.4	2.1 ± 0.7	2.2 ± 0.5			
Q3	2.0 ± 0.8	**1.4** ± 0.7	1.6 ± 0.8			
Q4	**1.4** ± 0.6	2.6 ± 1.2	3.3 ± 1.3	1.5 ± 1.0	4.4 ± 1.1	4.6 ± 1.1
Q5	**1.3** ± 0.6	2.1 ± 0.8	1.5 ± 0.6			
Q6	**1.4** ± 0.7	2.1 ± 0.9	1.6 ± 0.8	3.5 ± 0.9	3.6 ± 1.1	
Q7	1.9 ± 1.2	2.7 ± 1.5	**1.5** ± 0.9	3.3 ± 1.3	3.1 ± 1.3	
Q8	1.8 ± 1.0	**1.5** ± 1.1	3.1 ± 1.2	3.1 ± 1.0	3.0 ± 1.1	

Table 5: Relative ordering per question, mean ± st.dev.

$C_W(M) > C_W(\bullet T)$ (the cat being hungry vs holding salmon in the larder). The relative gap between Q5b and Q5c is confirmed in all cases. In Q6, respondents do not consider relevant the conditions related to the prohibition (d, e). Seeing the graph, because of the conjunction in the question, we need to generate all the rest to obtain the target, while the inhibition branch is independent and therefore less complex. Similar considerations hold for Q7 and Q8: the decreasing order of c, a, e, d is respected (but for Q8c); Q7b is aligned with the experiment if $C_W(M) < C_W(f * l * h)$, Q8b for the opposite condition.

Perspective In perspective, ST offers two additional advantages. First, taking into account *description complexity*, neglected in this study, ST can provide explanations for the observed misalignments. Informally, by framing the vocabulary of the question around the larder (Q3), a jam pot has an higher associative strength (and then less description complexity) than gluttony. The effect is inverted when questions are framed around eating jam. Q5a, rather than generation complexity, might be influenced by description complexity: "eating" strongly associates with "hungry"; consider for instance the alternative question: "why did the cat enter the larder?". Similar considerations apply for the different empirical results of Q7 and Q8, identical w.r.t. C_W.

Second, in our modeling exercise, we haven't specified a method for choosing the core action (e.g. "eating the jam") around which the action model given by the story may be constructed. ST considers relevance to be a matter not only of unexpectedness, but also of emotional interest. For instance, in passage 1, the "worrying" of the mother presents an "ought" that, if contradicted, would raise emotional interest. This is what occurs with "eating the jam".

Conclusion

The paper presents an early assessment of the gap between theoretical models and empirical observations with respect to the task of qualifying relevant causes. Our experiment suggests that *simplicity theory* (ST) might offer a better operational framework for the computation of pertinence of causes. Probabilistic methods, like Bayesian inference or causal Bayesian networks, implicitly assume a set-extensional semantics (classes of events), but such holistic approach to modeling implies a closure which is not cognitively plausible, and difficult to be maintained, even in simple stories like the ones studied here. Furthermore, these methods put aside the fundamental problem of contextualizing interpretation by deciding the set of variables under study upfront. This study was necessarily limited to models made by hand to test the adequacy of methods; however, the positive confirmation of good judgment prediction of ST theory is a strong motivation towards automatizing the model construction process, by inverting the problem: an action-scheme is nothing more than pertinent answers to a sequence of *why* questions.

References

Bex, F., & Verheij, B. (2011). Solving a Murder Case by Asking Critical Questions. *Argumentation*, *26*(3), 325–353.

Chater, N. (1999). The search for Simplicity: a fundamental Cognitive Principle? *The Quarterly Journal of Experimental Psychology*, *52A*(2), 273–302.

Dessalles, J. L. (2013). Algorithmic simplicity and relevance. *Algorithmic probability and friends*, *7070 LNAI*, 119–130.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., … Zhang, J. (2010). Actual causation: A stone soup essay. *Synthese*, *175*(2), 169–192.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy*, *98*(6), 273–299.

Levinson, S. C. (1989). A review of relevance. *Journal of Linguistics*, *25*(2), 455–472.

Lewis, D. K. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision making. *Cognition*, *49*, 123–163.

Rips, L. J. (2011). Causation From Perception. *Perspectives on Psychological Science*, *6*, 77–97.

Rissland, E. L., & Skalak, D. B. (1989). Combining case-based and rule-based reasoning: A heuristic approach. In *Proceedings of the 11th IJCAI* (pp. 524–530).

Sileno, G. (2016). *Aligning Law and Action*. Doctoral dissertation, University of Amsterdam.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Wiley.

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*(1), 107–119.

Twining, W., & Miers, D. (1982). *How to Do Things with Rules* (2nd ed.). Cambridge University Press.

Williamson, J. (2009). Probabilistic Theories of Causality. In *The oxford handbook of causation.* OUP.

Wilson, D., & Sperber, D. (1998). Pragmatics and time. In *Relevance theory: Applications and implications* (pp. 1–22). John Benjamins Publishing.

Yuan, C., Lim, H., & Lu, T. C. (2011). Most relevant explanation in bayesian networks. *Journal of Artificial Intelligence Research*, *42*, 309–352.

Zhang, J. (2008). Causal Reasoning with Ancestral Graphs. *Journal of Machine Learning Research*, *9*, 1437–1474.