Correlation Detection with and without the Theory of Conditionals: A model update of Hattori & Oaksford (2007)

Tatsuji Takahashi[a,*], Kuratomo Oyo[b], Akihiro Tamatsukuri[a]

[a] *School of Science and Engineering, Tokyo Denki University*
[b] *School of Policy Studies, Kwansei Gakuin University*

Abstract

We view observational causal induction as a statistical independence test under an assumption of rarity. This paper complements the two-stage theory of causal induction proposed by Hattori and Oaksford (2007) with a computational analysis. We show that their dual-factor heuristic (DFH) model has a rational account as the square root of the index of (non-)independence under extreme rarity assumptions, contrary to the criticism that the DFH model is non-normative (e.g., Lu et al., 2008). We introduce a model that considers the proportion of assumed-to-be rare instances (pARIs), and can be seen as a simplified version of the DFH model, and the probability of biconditionals, based on the theory of compound conditionals. pARIs conservatively approximates the non-independence measure, the square of DFH. In reproducing the meta-analysis in Hattori and Oaksford (2007), we confirm that pARIs and DFH have the same level of descriptive adequacy, and that the two models have the highest fit among more than 40 models. Then, we critically examine their computer simulations run as a rational analysis in Hattori and Oaksford (2007). We point out two problems in their simluations: samples in some of the simulations being restricted to generative ones, and indefinite values of models because of the small samples. In the light of especially the latter problem of definability,

*Corresponding author
Email address: tatsuji.takahashi@gmail.com (Tatsuji Takahashi)
URL: https://sites.google.com/site/tatsujitakahashi (Tatsuji Takahashi)

pARIs shows higher applicability.

Keywords: causal induction; causal inference; dual stage theory; dual frame theory; biconditionals; similarity judgment; Jaccard index;

Contents

1 Introduction **3**

2 Causal induction **4**

2.1 The two stage theory of causal induction: Observation then intervention 4

2.2 Models of causal induction with one cause and one effect 5

3 Computational analysis **7**

3.1 Statistical independence under extreme rarity assumptions 8

3.2 pARIs and DFH under extreme rarity 8

4 Meta-analysis **9**

4.1 Result 11

4.2 Discussion 11

5 Critical examination of the simulations **12**

5.1 Problem 1: Samples being always generative 13

5.2 Problem 2: Definability of indices on the samples 14

5.3 Summary 16

6 General discussion **16**

6.1 DFH and pARIs 16

6.2 Causal induction models in terms of conditionals 18

6.3 Computational cost of causal induction models 20

6.4 Observational causal induction as similarity judgment 20

6.5 Future work 21

7 Conclusion **22**

8 Acknowledgments **22**

1. Introduction

Causality is with no doubt one of the bases of human cognition (see e.g., [1; 2]). Notwithstanding the elusive nature of causal relationships that cannot be directly observed but rather constructed through inductive inference by internal observers under uncertainty (e.g., [3]), animals (including humans) have been remarkably successful in adapting to new environments through predictions of the future, manipulation of the environment, diagnosis of causes, and explanation of the reasons for or attribution of rewards, all based on learning and inference of causal relationship [4].

Efficient dealing with causal relationships has a high importance as for development of intelligent, learnable and autonomous systems [2]. As the scope of the machine intelligence becomes broader, enabling the robots and agents to act in parallel with humans, the problems facing them become human-like and human-level as well. Although there is no necessity for agents to solve the problems just as humans do. There will be more and more problems facing the agents that are human-like and also similar in small amounts of input, instead of big data, where agents need to learn from small pieces of information given online through time. As humans' inferences actually often work well, it is worth learning from it [4]. Especially considering that the performance of human reasoning heavily depends on the existence of "mental mechanisms" that are causal models of the task to solve [5; 6], causal inference has high importance in human-like AI research as well [2].

In this study, inheriting the theoretical framework for causal induction by Hattori & Oaksford [7], we critically examinine their model (DFH: dual-factor heuristics) and the simulations, and propose a simpler model (pARIs: proportion of assumed-to-be rare instances) that solves the problems that were implicit in their study. DFH and pARIs have similar biconditional forms; the difference consists in whether the biconditional form was derived from the limiting op-

eration from the normative correlation coefficient (DFH) or derived from the theory of conditionals and compound conditionals (pARIs).

2. Causal induction

There are at least two kinds of causal inference: inductive formation of causal relationships from observation of co-occurrence of events, and testing the supposed relationships through experimentation/intervention, The first, most inductive kind is the focus of this study.

2.1. The two stage theory of causal induction: Observation then intervention

We adopt the two stage theory of causal induction [7; 8; 9] as the theoretical framework in this study. According to the theory, in humans' causal induction in realstic environments abound with innumerable candidate causes and effects, and people first observe co-occurrences between events to determine the most relevant-seeming cause-effect pairs. From pure observation, especially only with co-occurrence, it is usually impossible to distinguish between correlation and causation. However, in many cases, causally related events are correlated. Hence, detecting correlation and filtering out irrelevant event pairs can be considered the goal of the first, observational stage of causal inference. The interventional stage comes next, in which experimentation can be used to test whether the distilled pairs are causal or spurious.

Because of the abundance of events, together with limits to mental computation, spatio-temporal observation scope and working memory, it is rational to use some simplified strategy in the first stage to detect correlated pairs of events that would be relatively rare among numerous irrelevant pairs. It is in this sense that the first, observational stage is called heuristic, whereas the second, interventional stage is called analytic. The most well-organised example of the second, interventional stage is modern scientific experiments. The two stages ordered this way is expected to reduce the cost of intervention and experimentation. The distinction between the two stages, especially in terms of

Table 1: The 2×2 contingency table for elemental causal induction. Ω denotes the universe.

	Effect (E)	No effect ($\neg E$)	Marginal frequency
Cause (C)	$N(C, E)$	$N(C, \neg E)$	$N(C)$
No cause ($\neg C$)	$N(\neg C, E)$	$N(\neg C, \neg E)$	$N(\neg C)$
Marginal frequency	$N(E)$	$N(\neg E)$	$N(\Omega)$

modelling, has gradually been established, both theoretically [8] and empirically [9].

2.2. Models of causal induction with one cause and one effect

The experimental and modelling framework of causal induction for this study is that agents form the causal strength from C to E based on the four frequencies in Table 1. E is the effect in focus and C is a candidate cause of E. We also assume that the models have the type of function with the domain of four natural numbers—$N(C, E)$, $N(C, \neg E)$, $N(\neg C, E)$ and $N(\neg C, \neg E)$—as in Table 1, to the co-domain of an interval of real numbers: $f : \mathcal{N}^4 \to V$. V is usually the unit interval $[0, 1]$ for probability or $[-1, 1]$ for indices like correlation coefficients. In the observational stage, the basic index on this table is the four-fold correlation coefficient ϕ:

$$\phi = \frac{P(C, E)P(\neg C, \neg E) - P(C, \neg E)P(\neg C, E)}{\sqrt{P(C)P(\neg C)P(E)P(\neg E)}}. \tag{1}$$

The most representative model of causal induction is ΔP [10]:

$$\Delta P = P(E|C) - P(E|\neg C) = \frac{N(C, E)}{N(C)} - \frac{N(\neg C, E)}{N(\neg C)}, \tag{2}$$

which is the increase in the probability of E caused by the presence of C relative to C's absence. This model formalises the basic idea of scientific experiments. For example, the first term, $P(E|C)$, represents the test group whereas the second term, $P(E|\neg C)$, represents the control group. ΔP is the regression slope from C to E, letting

$$\Delta P^{\dagger} = P(C|E) - P(\neg C|E), \tag{3}$$

$$\phi = \sqrt{\Delta P \cdot \Delta P^{\dagger}}, \tag{4}$$

neglecting the standard deviations. Despite its rational status, ΔP does not fit the responses of humans in causal induction experiments well, and it is easy to design an experiment in which there is no correlation between the data and the prediction by ΔP predictions (e.g., see HO07.2 and W03.2 in the meta-analysis below, in Table 2). This index is at the core of most Bayesian models of causal induction [11; 12; 13].

DFH was introduced as a descriptive model, which is the geometric mean of the 'predictability' of effect from cause $P(E|C)$ and its converse, the 'diagnosability' of cause from effect $P(C|E)$, predicting that humans feel the strong causality only when both are high:

$$\begin{aligned} \text{DFH} &= \sqrt{P(E|C)P(C|E)} && (5) \\ &= \frac{P(C,E)}{\sqrt{P(C)P(E)}} && (6) \\ &= \frac{N(C,E)}{\sqrt{N(C)N(E)}} && (7) \end{aligned}$$

DFH showed the highest correlation with the data among models in the meta-analysis in [7], compared with 32 other non-parametrised and 8 parametrised models. As grounds for the rationality of this model, Hattori and Oaksford showed that DFH is a limiting case of ϕ where C and E are both rare [7]:

$$\lim_{N(\neg C, \neg E) \to \infty} \phi = \text{DFH}. \tag{8}$$

The limit operation is called an extreme rarity assumption; it neglects the assumed-to-be very abundant $N(\neg C, \neg E)$-cell information.

Some of the authors have the pARIs model, which is similar in some ways to DFH, but simpler [14]:

$$\begin{aligned} \text{pARIs} = P(C \wedge E | C \vee E) &= \frac{P(C,E)}{P(C \vee E)} && (9) \\ &= \frac{P(C,E)}{P(C) + P(E) - P(C,E)} && (10) \\ &= \frac{N(C,E)}{N(C) + N(E) - N(C,E)} && (11) \end{aligned}$$

This model was named in reference to the proportion of confirmatory instances (pCI) model, proposed by White [15], which is defined as:

$$\mathrm{pCI} = \frac{N(C,E) + N(\neg C, \neg E)}{N(\Omega)} = \frac{N(C,E) + N(\neg C, \neg E)}{N(C,E) + N(C,\neg E) + N(\neg C, E) + N(\neg C, \neg E)}. \tag{12}$$

pARIs and pCI are conceptually different in that pARIs is a conditional probability and pCI is not (except in a trivial sense). Instead of the limit operation used in deriving DFH from ϕ, pARIs is based on the conditioning on the cases in which either C or E (or both) occurred. When events C and E are rare, among joint events between them, the three joint events, $C \wedge E$, $C \wedge \neg E$ and $\neg C \wedge E$ are rarer than the joint absence $\neg C \wedge \neg E$. The conditionalisation sets the focus on the rare joint events.

3. Computational analysis

Here, we give a computational analysis of observational causal induction and rationalise DFH and pARIs. According to the two-stage theory of causal induction—the distinctiveness of the stages has been empirically confirmed in [9] —in the first, observational stage, the goal is to weed out unrelated event pairs and to sift a small number of correlated ones. Whether the correlated relationship is causal, and the direction of the relationship, can be tested only in the second, interventional stage. Below, we show that DFH can be considered an index of dependence detection under an extreme rarity of causal events, and is rational in that sense, as far as it does relevance detection that is the principal goal of the first stage. The DFH model has been criticised as *non-normative* in that it is 'not derived from a well-specified computational analysis of the goals of causal learning' and is a groundless heuristic in [13]. Here, we give a counterargument.

3.1. Statistical independence under extreme rarity assumptions

If C and E are statistically independent and $P(C)$ and $P(E)$ are non-zero, the following three terms are equal:

$$P(C, E) = P(C)P(E) = P(C|E)P(E|C). \tag{13}$$

In terms of frequency, Eq. 13 is

$$\frac{N(C, E)}{N(\Omega)} = \frac{N(C)}{N(\Omega)} \cdot \frac{N(E)}{N(\Omega)} = \frac{N(C, E)}{N(E)} \cdot \frac{N(C, E)}{N(C)}. \tag{14}$$

The derivation of DFH is based on an extreme rarity assumption represented by the limit operation where the occurrence probabilities of C and E are both very small. The rarity assumption is often used in probabilistic modelling of higher cognition (e.g., [16; 17]). In conditions of extreme rarity, i.e., $N(\neg C, \neg E) \to \infty$, although the first and second terms go to 0 because of the divergence of the denominator $N(\Omega)$, the third term (the product of conditional probabilities) stays invariant under the limit operation. Here, $P(C|E)P(E|C)$ being close to 0 indicates the independence of C and E, whereas the term being close to 1 shows that C and E are dependent. The third term is the square of DFH, and

$$\lim_{N(\neg C, \neg E) \to \infty} \phi^2 = P(E|C)P(C|E) \tag{15}$$

holds, which means that $P(C|E)P(E|C)$ is the limiting case of the coefficient of determination on the 2×2 table, ϕ^2. On the other hand, the second term of ΔP, $P(E|\neg C)$, vanishes and only $P(E|C)$ remains: $\lim_{N(\neg C, \neg E) \to \infty} \Delta P = P(E|C)$. $P(E|C)$ fits the data quite poorly in the meta-analysis in [7] (Model 12).

DFH is the square root of $P(C|E)P(E|C)$, the measure for dependence under extreme rarity, and pARIs is its approximation within the range of a single conditional probability.

3.2. pARIs and DFH under extreme rarity

Under extreme rarity, an index should take a certain value when C and E are independent, and its antipodal value when C and E are not independent,

as discussed concerning Eq. 14. In this regard, when C and E are independent,

$$\begin{aligned} \text{pARIs}^{\text{indep}} &= \frac{P(C)P(E)}{P(C)+P(E)-P(C)P(E)} && (16) \\ &= \frac{1}{\frac{1}{P(E)}+\frac{1}{P(C)}-1}, && (17) \end{aligned}$$

which implies that if $P(C)$ and $P(E)$ are both small (large), it goes to 0 (1). Also,

$$\text{DFH}^{\text{indep}} = \frac{P(C)P(E)}{\sqrt{P(C)P(E)}} = \sqrt{P(C)P(E)}, \tag{18}$$

under independence and shows a similar property to $\text{pARIs}^{\text{indep}}$. We can also see that pARIs and DFH are 0 and 1 when $P(E|C) = P(C|E) = 0$ and $P(E|C) = P(C|E) = 1$, respectively. The difference in the behaviour of pARIs and that of DFH is how they connect the two extreme values, 0 and 1. It holds that pARIs $\leq$ DFH, because letting $U = \max\{(P(C), P(E)\}$, $P(C)P(E) \leq U^2$ and hence, $\sqrt{P(C)P(E)} \leq U$. From Eqs. 6 and 10:

$$\begin{aligned} \frac{\text{DFH}}{\text{pARIs}} &= \frac{P(C)+P(E)-P(C,E)}{\sqrt{P(C)P(E)}} && (19) \\ &\geq \frac{P(C)+P(E)-P(C,E)}{U} \geq 1. && (20) \end{aligned}$$

This means that pARIs is more prudent or conservative than DFH in that it takes a high value only when the independence is very high [18], which could be useful in weeding out unrelated pairs in the first stage.

4. Meta-analysis

To compare the fit of the models introduced in the previous section, we reproduced the meta-analysis by Hattori and Oaksford [7] [1], with which they showed

[1] There was a small mistake in the meta-analysis in [7]. The mean of the responses by the participants for the stimulus $(N(C,E), N(C,\neg E), N(\neg C,E), N(\neg C,\neg E)) = (28, 0, 21, 7)$ in the experiment abbreviated as LS00 (in Experiment 2 in [19]). It was listed as 34 in the analysis, but is actually 65. This correction does not affect the result very much and therefore their argument is not affected.

that DFH best fits the data among 41 existing indices [7]. The experiments included in the meta-analysis were on causality, correlation or contigency, and selected according to the following five conditions. (1) The participants were humans, (2) there was only one effect and one candidate cause, (3) the answers were nearly continuous values, (4) the samples were given sequentially according to the predetermined frequencies, and (5) the causes were generative (a cause makes the effect occur), not preventive (a cause prevents the effect from occurring). In terms of the stimulus given to the participants, the last condition is defined by $N(C, E)N(\neg C, \neg E) \geq N(C, \neg E)N(\neg C, E)$, which is equivalent to $\phi \geq 0$ and $\Delta P \geq 0$.

Eight experiments are included in the meta-analysis: Experiment 1 of Experiment 1 of [20] (AS95); Experiments 1 and 3 of [21] (BCC03.1 and BCC03.3) [2]; Experiments 1 and 2 of [7] (HO07.1 and HO07.2); Experiments 1–3 of [19] (LS00); and Experiments 2 and 6 of [15] (W03.2 and W03.6). To measure each index's fit to the data, we calculated the coefficient of correlation (r) between each index and participants' mean ratings of causal strength.

Among the eight experiments included, AS95 differs from others in some ways. One is that the number of stimuli is very large (80 out of 143, the total of all eight experiments: more than the half of the total), which gives a much higher weight to AS95 in calculating the overall result than all other experiments. In this experiment, 40 participants responded to all the 80 stimuli, taking one hour, which is much longer than for other experiments, and the stimuli varied differentially. Although the participants should independently respond to each stimulus, totally forgetting the past stimuli presented before, it is possible that they kept learning co-occurrence information; hence, their responses may have

[2] In the BCC03.1 experiment, there are 15 stimuli, of which only 13 are handled in [7]. We excluded two stimuli where $P(E|C) = P(E|\neg C) = 0$ and $P(E|C) = P(E|\neg C) = 1$ and roughly replicated the original result. For the former, DFH and the Power PC model [11] cannot be computed because their denominators are zero (i.e. Case 3 in Table 3). We guess that the latter was excluded because the stimulus was not strictly generative.

become increasingly confounded with past stimuli as the experiment proceeds. Although this effect may have happened in other experiments, it may be the most prominent in AS95.

Even without the possible problems mentioned above, it is not ideal that a meta-analysis result depends so much on a single experiment that necessarily has a specific set of conditions. Therefore, we calculated the overall data fit with and without AS95, as shown in Table 2.

4.1. Result

In replicating the meta-analysis, we completely reproduced the results in [7], except for the small mistake in the original data as mentioned in Footnote 1. As we confirmed that DFH shows the best fit among the 33 non-parametrised models, we focused on comparing DFH and pARIs.

Table 2 shows the result of the meta-analysis of DFH and pARIs. Despite its simple form, pARIs has the same high correlation to the data in comparison with DFH, with the lowest fit of $r^2 = .78$ to W03.2, and $r^2 = .69$ for DFH. On the data from the AS95 experiment, which has the largest number of responses (80), DFH fit better than pARIs. We show the overall fit and error for the seven experiments (excluding AS95) in Table 2 as $\bar{r}^2{}'$ and RMSE$'$, as well. Without AS95, pARIs fits the data better than DFH, and vice versa with AS95. The root mean squared error for pARIs is overall smaller than DFH, with or without AS95.

4.2. Discussion

We conclude that pARIs has the same level of descriptive performance as DFH in many cases, despite its simpler definition. pARIs fits the data as well as DFH, which means that the data fit of pARIs is also better than the other 32 non-parametrised and 8 parametrised models compared in [7].

From the result of the meta-analysis, we can state that the causal intuition of human beings roughly follows the pARIs rule, as well as that of DFH. Then, what are the adaptive merits of having a pARIs-like causal intuition? One way to answer the question is computer simulations as carried out in [7].

Table 2: Replication of the meta-analysis in [7], showing the determination coefficient r^2 between the model predictions and the mean of the participants' responses. $\bar{r}^2$ is the weighted average with Fisher's z-transformation (the weights are the number of stimuli minus three), and RMSE is the root mean square error. $r^2{}'$ and RMSE$'$ are the results excluding the AS95 experiment. The last rows show the number of stimuli.

	AS95	BCC03.1	BCC03.3	HO07.1	HO07.2	LS00	W03.2	W03.6	$\bar{r}^2$	RMSE
pARIs	.89	.96	.94	.93	.99	.80	.78	.88	.89	9.44
DFH	.91	.95	.91	.93	.96	.80	.69	.80	.91	12.16
Stimuli	80	13	6	12	9	11	8	4	(143)	143
									$r^2{}'$	RMSE$'$
pARIs									.93	8.30
DFH									.91	14.85
Stimuli	—	13	6	12	9	11	8	4	(63)	63

5. Critical examination of the simulations

Hattori and Oaksford ran simulations that showed the adaptive merit of DFH-like causal intuition which is that it approximates the sample and population correlation coefficients ϕ and ϕ_0 well from contingency tables under some realistic conditions, such as limitations to working memory and the rarity and equiprobability of cause and effect [7]. It is considered the main part of their rational analysis [22] which is to test the optimality of a part of cognition in terms of the cognitive limitations of the agent and the structure of the environment in which the cognitive agent acts and adapts.

In their paper, Simulation 1 tests effectiveness and parsimony (Fig. 3a, 3b, and 4 in [7]) of some models and Simulation 2 treats their efficiency (Fig. 5 in [7]). For convenience, we rearrange the simulations into three groups rather than two. Simulation A (Fig. 3a in [7]) tests descriptive performance of DFH, that is, how similarly it behaves to ϕ (the correlation coefficient of the sample). Simulation B (Fig. 3b and 4 in [7]) tests inferential performance, or how well DFH and ϕ can be used to infer the value of ϕ_0, the correlation coefficient of the population. Simulation C (Fig. 5 in [7]) tests the speed and accuracy of the inference process as DFH sequentially approximates the value of ϕ_0 and shows the time development of the index values.

Throughout the simulations, there are two kinds of sampling. One is the usual sampling ('N-sampling') where the number of instances is $N = N(\omega) = N(C, E) + N(C, \neg E) + N(\neg C, E) + N(\neg C, \neg E)$. The other is called '$N_W$-sampling' where $N_W = N(C \vee E) = N(C, E) + N(C, \neg E) + N(\neg C, E)$. In the latter sampling, inspired by consideration of limitation in working memory capacity [3], only the assumed-to-be rare instances are counted, and the number of assumed-to-be abundant instances, $N(\neg C, \neg E)$, is neglected. For example, it can be that $N_W = 7$ while $N = 10$ when $N(\neg C, \neg E) = 3$. In Simulation C, the time development of the simluation ticks with the increments of N or N_W.

See [18] for the preliminary results, including those pARIs. Here, we just point out two implicit problems in the original simulations.

5.1. Problem 1: Samples being always generative

The first problem is the implicit filtering of the contingency tables generated from the probability distribution. The probability distributions in the simulations were fixed to be *generative* (C and E are positively correlated, i.e. $P(C, E) \geq P(C)P(E)$) rather than *preventive* (C and E are negatively correlated, i.e. $P(C, E) < P(C)P(E)$), because of the scope of their theory and model. The problem is that the contingency tables used for calcu-

[3] In [7], the limitation in working memory capacity is argued in terms of the magic number 7 ± 2, which has recently been considered to be around 4 [23]. However, especially considering the small samples treated in their and our studies, where the numbers on all four cells tend to be small (all usually one digit in the decimal system), it could be more appropriate to consider the number of joint event types, instead of the total frequencies, N or N_W, to be within the limitation. One prediction that can be derived from this is that people would quite accurately remember the total frequency of four types such as on the 2×2 contingency tables, while people would tend to ignore some of the joint event types on the 3×3 or larger contingency tables. The 3×3 tables are called for in cases in which some of the occurrences of C or E may be uncertain, which often happens in scientific data collected by sensors [24; 25], and are similar to 3×3 de Finetti truth tables [26]. Still, sample sizes as small as 7 ± 2 is a realistic condition that people can deal with only a smaller number of frequencies because of a temporal rather than memory limit, where we usually make an action as soon as possible, in addition to the memory limitation.

lating and evaluating the indices are limited to generative ones, i.e., where $N(C,E)N(\neg C,\neg E) > N(C,\neg E)N(\neg C,E)$ holds. We have confirmed this problem with an exhaustive test of possible simulation settings. The probability distribution being generative does not mean that its sample is always generative. This implicit condition of limiting the samples to be generative was imposed on Simulation A and B, but not C.

We think that preventive tables sampled from generative populations should be included, because it is unnatural that agents can discard contingently preventive cases and keep sampling until the sample is generative with the same, fixed generative population. Preventive samples should be judged at a lower value of the indices, rather than arbitrarily discarded. In preliminary simulations, we found that the correlation between DFH and ϕ is generally lower when preventive tables are included, when $P(C)$ and $P(E)$ are not small. Seemingly, it does not affect the interpretations of the original result that focus on the cases in which $P(C)$ and $P(E)$ are both small, but further investigation is needed.

5.2. Problem 2: Definability of indices on the samples

The other problem is in the treatment of contingency tables on which some indices are not defined. For example, DFH takes the undefined value 0/0 because of the zero denominator, for example when $(N(C,E), N(C,\neg E), N(\neg C,E), N(\neg C,\neg E)) = (0,0,1,10)$, whereas pARIs takes the value 0. The cases in which the indices are defined (i.e. have a definite value) have an inclusion order, and when ϕ is defined, DFH is always defined, as we see below. The possible undefinedness affects the significance of all three simulations, especially the early stage of Simulation C.

We analyse the 'definability' of the models, especially because we deal with very small samples that model the quite limited capacity of our working memory, and we often do not have sufficient time to gather data, nor to wait for more occurrences. The models have different definability, as summarised in Table 3, where the combinations of the values of three cells—excluding $N(\neg C,\neg E)$, zero or positive—are considered. In the eight cases, min $N \leq 4$ and min $N_W \leq 3$.

Table 3: The definability (D: defined, –: not defined) of the four indices ϕ, ΔP, DFH and pARIs, on the $2^3 = 8$ cases where three frequencies $N(C, E)$, $N(C, \neg E)$ and $N(\neg C, E)$ are zero or positive (+). We fixed $N(\neg C, \neg E)$ to be positive in $N_W \geq 1$ cases for simplicity.

Case	1	2	3	4	5	6	7	8
$N(C, E)$	0	0	0	0	+	+	+	+
$N(C, \neg E)$	0	0	+	+	0	0	+	+
$N(\neg C, E)$	0	+	0	+	0	+	0	+
$N(\neg C, \neg E)$	0	+	+	+	+	+	+	+
min N	0	2	2	3	2	3	3	4
min N_W	0	1	1	2	1	2	2	3
ϕ	–	–	–	D	D	D	D	D
ΔP	–	–	D	D	D	D	D	D
DFH	–	–	–	D	D	D	D	D
pARIs	–	D	D	D	D	D	D	D

As for the four indices ϕ, ΔP, DFH and pARIs, the more complicated the index is, the fewer cases it is defined for. ϕ, ΔP, DFH and pARIs have 5, 6, 5 and 7 definable cases, respectively. pARIs is defined simply whenever $N_W \geq 1$. DFH is defined when $\min N_W \geq 2$ or $N(C, E) \geq 1$. A sufficient condition for ϕ and ΔP to be defined is $\min N \geq 3$.

The original simulations in [7] did not consider definability. Definability can affect the validity of the results if indices with different definability are compared, and this implicitly happened in their study. A valid comparison would be to say one model is better than another only when the former is defined on strictly more cases than the latter and still performs better.

In Simulation A (Fig. 3a of [7]), only contingency tables in which ϕ is defined are included in the result. When the sample size is small (around $N = 7$), a large part of the contingency tables with small $P(C)$ and $P(E)$ do not define ϕ: only Cases 4–8 in Table 3 do. It discounts the meaningfulness of Simulation A in terms of small samples, and this is a heavy condition to impose on the tables when N or N_W is so small such as $N \sim \mathcal{N}(7, 1^2)$, especially when we assume

that $P(C)$ and $P(E)$ are both small. We calculated the proportion of tables generated in Simulation A on which DFH is defined, which were around 35% and 65% of the sampled tables when $P(C) = P(E) = .1$ and $P(C) = P(E) = .2$, respectively. pARIs is defined on around 62% and 86% of the tables, respectively, which means that pARIs faces this problem strictly less than DFH.

This condition of ϕ being definable is also imposed on Simulation B, although it is not necessary in this simulation setting because ϕ_0 is the measure to be inferred. pARIs gives qualitatively similar but significantly better inferences of the population correlation coefficient ϕ_0 in Simulation B than DFH, even though pARIs is computed on a wider variety of contingency tables.

In Fig. 5 in [7] from Simulation C, in which the time developments of the mean index value of the indices from sample sizes 1 to 30 are shown, the mean value of DFH always decreases from 1.0 at sample size 1, because DFH is definable only for $(N(C, E), N(C, \neg E), N(\neg C, E)) = (1, 0, 0)$ when $N = 1$ or $N_W = 1$. Such samples are less than 20% at $N = 1$, and 18%–90% at $N_W = 1$, depending on the population correlation coefficient $\phi_0 \in \{.1, .2, ..., .9\}$. The percentages monotonically increase in N or N_W increments. The percentages of the samples on which pARIs can be defined are 20%–40% at $N = 1$ and 100% at $N_W = 1$, and a strict order relation holds with DFH, as we can see in Table 3.

5.3. Summary

Although DFH is the limiting case of the correlation coefficient, pARIs better infers the population correlation, and in much wider cases. This property makes a larger difference when considering multiple causes and effects, making the simulation settings more realistic.

6. General discussion

6.1. DFH and pARIs

We begin our discussion by comparing pARIs and DFH, the two similar models. Both are related to causal biconditionals, (i.e. 'if cause then effect,

and if effect then cause'). DFH represents both conditionals as the geometric mean of $P(E|C)$ and $P(C|E)$, and pARIs directly uses the probability of the biconditional:

$$\text{pARIs} = P((E|C) \wedge (C|E)) = P(C \wedge E|C \vee E), \tag{21}$$

according to the theories of compound conditionals [27; 28; 29; 30], or in a conditional event algebra initiated in [31] and then developed in, e.g., [32]. In the formal system, $E|C$ and $C|E$ are themselves considered to be events, and hence can be logically connected.

Another difference between pARIs and DFH is the process by which they ignore the absence-absence $\neg C \wedge \neg E$ cases. pARIs conditions on $C \vee E$ cases, whereas DFH starts from ϕ and assumes that $N(\neg C, \neg E)$ diverges to positive infinity. Let us consider the nature of $\neg C \wedge \neg E$ cases to see which is more appropriate for dealing with them.

The prerequisite for frequency information is counting occurrences. However, how do we count the absence, or non-occurrence, of events? Often, counting non-occurrences is indefinite or ill-defined. When we consider the causal relationship between the effect of an engine starting (E) and a candidate cause of turning the key (C), there are four cases of co-occurrence. We have no difficulty in counting the other three cases, but the $\neg C \wedge \neg E$ cases are problematic. How many times have we observed 'no key turning ($\neg C$) and no engine starting ($\neg E$)'? We could say we have observed this case infinite times ($N(\neg C, \neg E) \to \infty$). We can also say that we have observed it as many as the intervals between key turning and/or engine starting ($N(\neg C, \neg E) := N(C \vee E) - 1$), or we might want to say we have never observed such an event.

The count of $N(\neg C, \neg E)$ depends on the spatio-temporal framing of events. A more definite framing than the examples above is the division of a day into 144 ten-minute intervals and the identification of occurrences of C and E. Or we can count the occurrence of times when a driver enters a car, turns the key, and then exits the car after we observe the effect. In this way, $\neg C \wedge \neg E$ cases can be made definite, although there are many ways to frame it. In our

everyday causal judgment, it would be quite rare for framing to be rigorous and unambiguous. Defining the unit for $\neg C \wedge \neg E$ cases is a necessary condition for determining the *probabilistic model* in the sense of assignment of probabilities to atomic events. In other words, the probabilistic models that we use in everyday life are often not well-defined. In contrast, modern scientific experiments with a control group have a well-defined framing. The division of test and control groups itself is the effect of intervention, and causal induction is no longer purely observational.

Considering the indefiniteness of framing in everyday causal induction, in which we can count $\neg C \wedge \neg E$ cases arbitrarily, neglecting $N(\neg C, \neg E)$ by conditioning may be more natural and robust than always expecting countless $N(\neg C, \neg E)$. pARIs can be considered conditioning on $C \vee E$ of the conjunctive case $C \wedge E$ or of the probability of logical equivalence $P(C \wedge E) + P(\neg C \wedge \neg E)$ (equivalent to the pCI rule). Considering the often indefinite $N(\neg C, \neg E)$, ignoring them via conditioning and judgment about the statistical relationship between C and E from the three other, more definite cases would give a stable estimate. pARIs performs precisely this function.

6.2. Causal induction models in terms of conditionals

We compare the three models pARIs, DFH and ΔP, in terms of their use of conditionals, based on the fact that the subjective probability of a causal conditional 'if C, then E' is well described by the conditional probability $P(E|C)$ (e.g., see [33]).

$\Delta P = P(E|C) - P(E|\neg C) = P(E|C) + P(\neg E|\neg C) - 1$ is the adjusted sum of the probabilities of the conditional 'if C, then E' and its inverse, 'if not C, then not E', which correspond to sufficient and necessary conditions, respectively [34]. However, the latter probability vanishes under extreme rarity. In pARIs and DFH, the converse ('if effect then cause', $P(C|E)$) is used as a kind of necessary condition, instead of the inverse, which stays valid under the rarity assumption. Although the contrapositive conditionals, the inverse and the converse, are equivalent in classical logic, they are not equivalent in probabilistic

logic. Still, $P(C|E) = N(C,E)/N(E)$ and $P(\neg E|\neg C) = N(\neg C, \neg E)/N(\neg C)$ have the common element $N(\neg C, E)$ in their denominators, and therefore when $N(C,E)$ and $N(\neg C, \neg E)$ are close., or if the value of $N(C,E)$ is used in place of $N(\neg C, \neg E)$, especially when the latter is indefinite or uncertain, $P(C|E)$could well work on behalf of $P(\neg E|\neg C)$.

Several authors have theoretically shown that biconditionals of the form 'if p, then q, and if q, then p' has the form similar to pARIs as in Eq. 21 [27; 28; 29; 30]. One of the authors has confirmed that the abstract (non-causal) biconditionals has a well corresponding truth table to pARIs in [35]. In the subjective probability theory developed by De Finetti, conditional probabilities are atomic and that they correspond to conditional events [31].

Causal biconditionals and pARIs-like causal intuition likely have a more direct relationship, as suggested in [36]. The causal relationship between cause and effect may be recognised when both directions of conditional—'if cause, then effect' (predictive) and 'if effect, then cause' (diagnostic)—are acceptable (i.e., $P(\text{effect}|\text{cause})$ and $P(\text{cause}|\text{effect})$ are both high). Considering only the predictive conditional is sufficient for predicting the occurrence of effect from that of cause, but it does not exclude the possibility that $P(\text{effect}|\text{cause})$ is high just because $P(\text{effect})$ is high and cause is not strongly preventive of effect. Evaluating $P(\text{cause}|\text{effect})$ at the same time may be an effective strategy for considering the necessity of the cause for the effect.

In a Bayesian framework, Lu et al. point out that people favour sparse and strong causes [13]. If we interpret a cause C as being strong when $P(E|C)$ is high, and sparse when $P(C|E)$ is high, it can be modelled as a high biconditional probability. One situation in which sparseness can be modelled as $P(C|E)$ being high is when multiple causes $C_1, C_2, \ldots$ are mutually exclusive. $P(C_i|E)$ being high means that all $P(C_{j\neq i}|E)$ are low. In this sense, the form of pARIs can be considered a simple expression of people's priors on causal relationships.

6.3. Computational cost of causal induction models

Forming a conditional probability would elicit some cognitive cost through the Ramsey test [37] or alike. Two conditional probabilities are required to compute ΔP and DFH, whereare pARIs needs a single conditional probability. Still, pARIs accounts for two reciprocal conditionals, consistent with the general tendency that single (rather than plural) hypotheses are preferred (e.g. the singularity principle by Evans [38]). More complicated or computationally demanding Bayesian models (e.g., causal support in [12] and the SS model [13]) are not excluded as cognitive models. However, the kind of Bayesian computations that can realistically or effectively be approximated in our brains must be investigated further.

In all cases, pARIs is advantageous as a cognitive model because it is a simple, single conditional probability that integrates properties like those of ΔP and DFH. This makes a bigger difference when we consider many-to-one or many-to-many cause-effect correspondences, which are the computational goal of causal search to be done in the observational stage of causal induction.

6.4. Observational causal induction as similarity judgment

As we have seen in the previous section, there are several indices of similarity and other concepts equivalent to pARIs. Moreover, DFH is also equivalent to a similarity indices called cosine similarity (or the Ochiai similarity index [39]). It is mathematically interesting that the correlation coefficient $\phi \in [-1, 1]$ becomes the similarity index DFH $\in [0, 1]$ under the limit operation, as in Eq. 8 (equation (2) in [7]). pARIs is also equivalent to the Jaccard similarity index [40], which is commonly used in statistical analysis including cluster analysis in biology and sociology (e.g., [41]), and gives rise to a proper distance metric between sample sets [42]. These facts suggest that humans judge statistical similarity between the occurrences of C and E when they observe the statistical information to verify a correlation between the two events. It is consistent with the hypothesis that Chater and Oaksford proposed: the human ability to learn

causal relationships remarkably quickly—sometimes even in a single instance—depends on the feature similarity of the candidate cause and the effect [6].

Succeeding the study by Hattori and Oaksford [7], this study suggested that people respond with statistical similarity to ascertain the strength of causal relationships in a realistic setting with a small sample and no other cues (spatial or temporal information, prior knowledge, or causal models). According to premilinary simulation results in [18], similarity indices—especially pARIs—contributes to fast and accurate inference of correlation under the rarity assumption, at least when compared with correlation (ϕ) and regression (ΔP) indices.

6.5. Future work

Establishing observational causal induction as a stastical independence test, and the pARIs model will require further theoretical and empirical arguments. To expand the theoretical aspect, we must situate the observational stage into a more comprehensive framework of causal inference. One possibility is a causal search in which the topology (graph structure) of a causal Bayes network, rather than the edge weights (strength), is established [34; 12]. In the PC algorithm for causal search [43], which is a representative constraint-based algorithm for constructing sparse causal graphs, the complete undirected graph of all the variables is formed, and then undirected edges are removed if they are (conditionally) independent. The directionality of the edges between nodes is determined only after the removal of undirected edges. Our hypothesis is that people use pARIs- or DFH-like intuition to quickly test the independence quickly under assumptions like rarity of events, explaining why DFH and pARIs are symmetric or undirected (i.e. invariant under exchange of cause and effect), and may completely resolve some of the rationality issues posed in [13].

Empirically, we can investigate whether people are trying to test independence, and when they do so, which factors contribute to switching behaviours and goals [9]. To compare DFH and pARIs, we can set the stimuli so that the predictions by the two models diverge. DFH depends on the value of $N(C, \neg E) \cdot$

$N(\neg C, E)$ whereas pARIs does not, so we can keep $N(C, \neg E) + N(\neg C, E)$ invariant and vary $N(C, \neg E) \cdot N(\neg C, E)$.

7. Conclusion

Adopting the framework of the two-stage theory of human causal judgment [7], we have proposed a computational analysis of the first, observational stage of causal induction. According to this analysis, when an extreme rarity assumption is valid, it is rational to use the product of the two reciprocal conditional probabilities of two events to detect non-independence between them. The DFH model is the square root of this measure, and the pARIs rule effectively approximates it. We replicated the meta-analysis in [7] and showed that DFH and pARIs both fit the data well, even though the latter is simpler. We then pointed out two problems in the simulations as a rational analysis of causal induction in [7]. We showed that a crucial problem can be solved if we adopt pARIs as the model derived from the theory. The difference between DFH and pARIs is, as far as the two models represent causal biconditionals between cause and effect, that the former is merely the geometric mean of the two conditional probabilities, whereas the latter has rational grounding as the probability of a biconditional event. Adopting the rarity assumption, which has been proven important in covariation assessment [17] and other reasoning tasks (e.g. [16]), we conducted a new computational analysis and proposed a new model consistent with the theories of compound conditionals that further supports the new paradigm psychology of reasoning advocated by David Over [44].

8. Acknowledgments

We thank David Over for, firstly, pointing out that the pARIs rule is equivalent to the probability of a biconditional event, in a JSPS-ANR CHORUS Franco-Japanese project (BTAFDOC) meeting on 5 March 2012 in Paris, and for other generous support. We also thank Jean Baratgin, Mike Oaksford, and Nicole Cruz for helpful discussions and suggestions, and Masasi Hattori for

providing us with the raw data of the meta-analysis. Part of this work was supported by a Grant-in-Aid for Scientific Research 17H04696 from the Japan Society for the Promotion of Science.

References

[1] S. A. Sloman, D. Lagnado, Causality in Thought, Annual Review of Psychology 66 (1) (2015) 223–247, ISSN 0066-4308, doi:\bibinfo{doi}{10.1146/annurev-psych-010814-015135}, URL `http://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010814-015135`.

[2] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building Machines That Learn and Think Like People, Behavioral and Brain Sciences 40 (2017) 1–72.

[3] Y.-P. Gunji, T. Takahashi, M. Aono, Dynamical infomorphism: Form of endo-perspective, Chaos, Solitons and Fractals 22 (5) (2004) 1077–1101.

[4] T. L. Griffiths, J. B. Tenenbaum, Theory-based causal induction., Psychological Review 116 (4) (2009) 661–716.

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

[6] N. Chater, M. Oaksford, Mental Mechanisms: Speculations on Human Causal Learning and Reasoning, in: K. Fiedler, P. Juslin (Eds.), Information Sampling and Adaptive Cognition, Cambridge University Press, 210–236, 2006.

[7] M. Hattori, M. Oaksford, Adaptive Non-Interventional Heuristics for Covariation Detection in Causal Induction: Model Comparison and Rational Analysis, Cognitive Science 31 (5) (2007) 765–814, ISSN 0364-0213.

[8] M. Hattori, D. E. Over, I. Hattori, T. Takahashi, J. Baratgin, Dual frames in causal reasoning and other types of thinking, The Thinking Mind: A Festschrift for Ken Manktelow (2016) 98–114.

[9] I. Hattori, M. Hattori, D. E. Over, T. Takahashi, J. Baratgin, Dual frames for causal induction: the normative and the heuristic, Thinking & Reasoning 23 (3) (2017) 292–317.

[10] H. M. Jenkins, W. C. Ward, Judgment of Contingency between Responses and Outcomes, Psychological Monographs: General and Applied 79 (1) (1965) 1–17.

[11] P. W. Cheng, From covariation to causation: A causal power theory., Psychological Review 104 (2) (1997) 367–405.

[12] T. L. Griffiths, J. B. Tenenbaum, Structure and strength in causal induction., Cognitive Psychology 51 (4) (2005) 334–84.

[13] H. Lu, A. L. Yuille, M. Liljeholm, P. W. Cheng, K. J. Holyoak, Bayesian Generic Priors for Causal Learning, Psychological Review 115 (4) (2008) 955–984.

[14] T. Takahashi, Y. Kohno, K. Oyo, Causal induction heuristics as proportion of assumed-to-be rare Instances (pARIs), in: Proceedings of the 7th International Conference on Cognitive Science (ICCS2010), 361–362, 2010.

[15] P. A. White, Making Causal Judgments From the Proportion of Confirming Instances : The pCI Rule, Journal of Experimental Psychology: Learning, Memory, and Cognition 29 (4) (2003) 710 –727.

[16] M. Oaksford, N. Chater, A Rational Analysis of the Selection Task as Optimal Data Selection, Psychological Review 101 (4) (1994) 608–631.

[17] C. R. M. McKenzie, L. A. Mikkelsen, A Bayesian view of covariation assessment, Cognitive Psychology 54 (1) (2007) 33–61.

[18] T. Takahashi, K. Oyo, Biconditional probability as a model of symmetry inference and causal induction from small samples, in: Proceedings of the 31st Annual Meeting of the Japanese Cognitive Society, 141–148, 2014.

[19] K. Lober, D. R. Shanks, Is causal induction based on causal power? Critique of Cheng (1997)., Psychological review 107 (1) (2000) 195–212.

[20] J. R. Anderson, C. F. Sheu, Causal inferences as perceptual judgements., Memory & cognition 23 (4) (1995) 510–24.

[21] M. J. Buehner, P. W. Cheng, D. Clifford, From covariation to causation: a test of the assumption of causal power., Journal of experimental psychology. Learning, memory, and cognition 29 (6) (2003) 1119–1140.

[22] J. R. Anderson, The Adaptive Character of Thought, Hillsdele, NJ: Erlbaum, 1990.

[23] N. Cowan, The magical number 4 in short term memory. A reconsideration of storage capacity, Behavioral and Brain Sciences 24 (4) (2001) 87–186.

[24] J. Yokokawa, T. Takahashi, Inductive inference of causal relationship from uncertain co-occurrence information, in: Proceedings of the 76th National Convention of IPSJ, 1T–6, 2014.

[25] K. Tanaka, N. Namiki, K. Oyo, T. Takahashi, Causal induction by observation with uncertain occurrences, in: Proceedings of the 29th Annual Conference of the Japanese Society for Artificial Intelligence, 1E3–1in, 2015.

[26] J. Baratgin, D. E. Over, G. Politzer, Uncertainty and the de Finetti tables, Thinking & Reasoning 19 (3-4) (2013) 308–328.

[27] V. McGee, Conditional Probabilities and Compounds of Conditionals, The Philosophical Review 98 (4) (1989) 485–541.

[28] S. Kaufmann, Conditionals Right and Left: Probabilities for the Whole Family, Journal of Philosophical Logic 38 (1) (2009) 1–53.

[29] A. Gilio, G. Sanfilippo, Conditional Random Quantities and Compounds of Conditionals, Studia Logica 102 (4) (2014) 709–729.

[30] G. Sanfilippo, N. Pfeifer, D. E. Over, A. Gilio, Probabilistic inferences from conjoined to iterated conditionals, International Journal of Approximate Reasoning 93 (October) (2018) 103–118.

[31] B. de Finetti, Foresight: Its logical laws, its subjective sources (original publication, 1937), in: H. Kyburg, H. E. Smokier (Eds.), Studies in subjective probability, Wiley, New York, 55–118, 1964.

[32] I. R. Goodman, H. T. Nguyen, E. A. Walker, Conditional Inference and Logic for Intelligent Systems: A Theory of Measure-Free Conditioning, Amsterdam: North-Holland, 1991.

[33] D. E. Over, C. Hadjichristidis, J. S. B. T. Evans, S. J. Handley, S. a. Sloman, The probability of causal conditionals., Cognitive psychology 54 (1) (2007) 62–97.

[34] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, New York, NY, USA, 2000.

[35] J. Baratgin, G. Politzer, D. E. Over, T. Takahashi, Three-valued Truth Tables and the Psychology of Ignorance, manuscript under review .

[36] C. Gauffroy, P. Barrouillet, Heuristic and analytic processes in mental models for conditionals : An integrative developmental theory, Developmental Review 29 (4) (2009) 249–282.

[37] F. P. Ramsey, General Propositions and Causality, in: H. A. Mellor (Ed.), Philosophical Papers, Cambridge University Press, 1990.

[38] J. S. B. T. Evans, Hypothetical Thinking: Dual Processes in Reasoning and Judgement, Psychology Press, 2007.

[39] A. Ochiai, Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, Bull. Jpn. Soc. Sci. Fish 22 (9) (1957) 526–530.

[40] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bulletin de la Société Vaudoise des Sciences Naturelles 37 (1901) 547–579.

[41] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, 1990.

[42] G. Gilbert, Distance between Sets, Nature 239 (5368) (1972) 174–174.

[43] P. Spirtes, C. Glymour, An Algorithm for Fast Recovery of Sparse Causal Graphs, Social Science Computer Review 9 (1) (1991) 62–72.

[44] D. E. Over, New paradigm psychology of reasoning, Thinking & Reasoning 15 (4) (2009) 431–438.