A Resource-Rational Approach to the Causal Frame Problem

Thomas F. Icard, III (icard@stanford.edu), **Noah D. Goodman** (ngoodman@stanford.edu)
Departments of Philosophy and Psychology, Stanford University

Abstract

The *causal frame problem* is an epistemological puzzle about how the mind is able to disregard seemingly irrelevant causal knowledge, and focus on those factors that promise to be useful in making an inference or coming to a decision. Taking a subject's causal knowledge to be (implicitly) represented in terms of directed graphical models, the causal frame problem can be construed as the question of how to determine a reasonable "submodel" of one's "full model" of the world, so as to optimize the balance between accuracy in prediction on the one hand, and computational costs on the other. We propose a framework for addressing this problem, and provide several illustrative examples based on HMMs and Bayes nets. We also show that our framework can account for some of the recent empirical phenomena associated with alternative neglect.

Keywords: frame problem, bounded-resource-rationality, causal reasoning, alternative neglect.

Introduction

To any inference or decision problem there is no *a priori* bound on what aspects of a person's knowledge may be usefully, or even critically, applied. In principle, anything could be related to anything. This challenge is sometimes referred to as the *frame problem*, characterized by Glymour (1987) as: "Given an enormous amount of stuff, and some task to be done using some of the stuff, what is the *relevant stuff* for the task?" (65). The question is foundational to reasoning and rationality. Part of what makes people so smart is the ability to solve the frame problem, ignoring those aspects of the world (and one's knowledge of it) that are irrelevant to the problem at hand, thereby simplifying the underlying reasoning task, turning an intractable problem into a tractable one.

Not all of the psychological literature paints a picture of human reasoners as so adept at disregarding only the irrelevant, however. In the literature on causal reasoning, there is a robust empirical finding that subjects often neglect causal variables, including those that are in principle accessible to the subject, which would sometimes allow the subject to make better, more accurate inferences. So called *alternative neglect* is an especially well documented phenomenon, in which subjects ignore alternative possible causes of some event (Fischhoff et al. 1978; Klayman and Ha 1987; Fernbach et al. 2011, *inter alia*), even when doing so leads to incorrect inferences. More generally, at least in the causal domain, subjects seem to consider "smaller" models of the world than would be relevant to the task at hand, given the subject's knowledge and reasoning abilities. This has led many to criticize the behavior as normatively objectionable. Perhaps people are ignoring too much of their knowledge.

We would like to suggest that alternative neglect and related phenomena may be natural consequences of a general mechanism for sifting the most pertinent information from all other knowledge—that is, for solving the frame problem with regard to causal knowledge. Assuming a person's causal knowledge can be represented (at least implicitly) in terms of a very large directed graphical model (or Bayes net), the *causal* frame problem arises because computations involving the entire model promise to be intractable. Somehow the mind must focus in on some "submodel" of the "full" model (including all possibly relevant variables) that suffices for the task at hand and is not too costly to use. In as far as a proper submodel may nonetheless neglect relevant causal information, this may lead to inaccuracy. We suggest that perhaps the mind tolerates local and occasional inaccuracy in order to achieve a more global efficiency. To substantiate this claim, we need a better understanding of what it is for a submodel to be more or less apt for a task, from the perspective of a reasoner with bounded time and resources. It is clear that human reasoners cannot consult an indefinitely detailed mental model of the world for every inference task. So what kind of simpler model *should* a reasoner consult for a given task?

This work follows a line of research in cognitive science concerned with *bounded* or *resource rationality* (Simon 1957; Gigerenzer and Goldstein 1996, *inter alia*), and specifically in the context of probabilistic models, and approximations thereto (Vul et al., 2014). In addition to inherent interest, it has recently been suggested that considerations of bounded rationality may play a methodological role in sharpening the search for reasonable accounts of the cognitive processes underlying inductive inference (Griffiths et al., 2014; Icard, 2014). However, in this tradition there has been more of a focus on the algorithm used for inference in a given model, and less attention paid to questions of model selection.

In this largely programmatic paper we offer a framework for addressing the causal frame problem by selecting rational submodels, provide several illustrative examples, and address some of the empirical findings concerning alternative neglect.

Resource-Rational Submodels

Let $P(\mathbf{X})$ be a joint probability distribution over random variables $\mathbf{X} = X_1, X_2, \ldots$, and define a *query* to be a partition $\langle \mathbf{X}_q; \mathbf{X}_l; \mathbf{X}_e \rangle$ of $\mathbf{X}$ into *query variables*, *latent variables*, and *evidence variables*, respectively. A typical query task is to find values of $\mathbf{X}_q$ that maximize the conditional probability $P(\mathbf{X}_q \mid \mathbf{X}_e = \mathbf{v})$, marginalizing over $\mathbf{X}_l$. Clearly, the difficulty of this and related tasks scales with the number of variables. We will be interested in smaller models with fewer variables: a sublist $\mathbf{X}^*$ of $\mathbf{X}$ with associated distribution $P^*(\mathbf{X}^*)$, and partition $\langle \mathbf{X}_q; \mathbf{X}_l^*; \mathbf{X}_e^* \rangle$, so that only latent and evidence variables are ignored. The intention is for P^* to be close in structure to P but without the neglected variables. In each of the cases considered here (HMMs and Noisy-Or Bayes nets), there will be a canonical way of choosing P^* given $\mathbf{X}^*$.

Given $P(\mathbf{X})$ and $P^*(\mathbf{X}^*)$, there are at least two kinds of questions we would like to ask. The first of these captures how well an agent will fare by using the approximate submodel, as compared with the full model, holding fixed a procedure for using this model to choose an action. For instance, the agent might use this distribution to compute expected utility, or to *sample* from the model in order to approximate expected utility (see, e.g,. Vul et al. 2014). The second question asks how far off the approximate model is from the "true" model in its probabilistic predictions.[1]

1. Given a decision problem with action space $\mathcal{A}$ and utility function $U : \mathbf{X}_q \times \mathcal{A} \to \mathbb{R}$, and assuming fixed a (stochastic) choice rule Ψ_Q taking a distribution Q over $\mathbf{X}_q$ to a distribution on actions, what are the respective expected utilities of using P and P^* under (assumed "true") distribution P? That is, how great is the following difference Δ_{P,P^*}?

$$\Delta_{P,P^*} \;=\; \mathbb{E}_{\mathbf{x}\sim P}\,\mathbb{E}_{A\sim\Psi_P}\,U(\mathbf{x},A) - \mathbb{E}_{\mathbf{x}\sim P}\,\mathbb{E}_{A\sim\Psi_{P^*}}\,U(\mathbf{x},A)$$

2. How far is P^* from P in information distance, for the variables $\mathbf{X}_q$ of interest? That is, what is the Kullback-Leibler (KL) divergence between P and P^* with respect to $\mathbf{X}_q$?

$$KL(P \,||\, P^*) = \sum_{\mathbf{x}} P(\mathbf{X}_q = \mathbf{x} \mid \mathbf{X}_e = \mathbf{v}) \log \frac{P(\mathbf{X}_q = \mathbf{x} \mid \mathbf{X}_e = \mathbf{v})}{P^*(\mathbf{X}_q = \mathbf{x} \mid \mathbf{X}_e^* = \mathbf{v}^*)}$$

In general we will expect that $KL(P \,||\, P^*) > 0$, and $\Delta_{P,P^*} > 0$, indicating that the full model yields more accurate results. However, in line with other work on resource rationality, assuming the requisite computations using distribution P come with a greater cost than when using P^*, this difference in cost may well be worth the difference in accuracy or utility.

Suppose we have a cost function $c : \mathcal{P} \to \mathbb{R}^+$, assigning a real cost to each approximation $P^* \in \mathcal{P}$. For instance, c may simply charge a constant amount for each variable used in the associated submodel, or may be proportional to a graph property such as tree width of the corresponding graphical model. Given a set $\mathcal{P}$ of approximations, we can then ask for the *resource-optimal* approximation in either of the above two senses. For instance, with KL-divergence the distributions of interest include any $\tilde{P}$ that optimally trades off cost against KL-distance from the true distribution P:

$$\tilde{P} \;=\; \operatorname*{argmin}_{P^*\in\mathcal{P}} KL(P \,||\, P^*) + c(P^*)\,. \qquad (1)$$

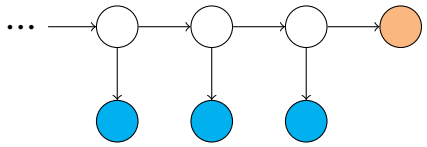
Notice the immediate result that any node X that is screened-off from $\mathbf{X}_q$ by $\mathbf{X}_e$ should be eliminated: doing so will not reduce the KL, but will improve the efficiency of inference.

In what follows we illustrate these ideas with three examples using familiar graphical models. The first example, of an HMM, demonstrates an extreme case of the frame problem in which the initial model is infinite. We show that the resource-optimal submodel is not only finite, but often quite small, and in many instances includes just a single node. The second example, of a causal Bayes net, shows that under a sampling scheme for decision making, the submodel actually outperforms the "ideal" full model in many cases, even without taking costs into account. Finally, the third example reveals that certain kinds of inferences may be subject to greater information loss resulting from neglect than others. Recent empirical literature shows that people respect this difference, suggesting that there may indeed be an element of resource rationality in alternative neglect behavior.

[1] Strictly speaking, the second can be seen a special case of the first, with a logarithmic scoring rule (Bernardo and Smith, 1994).

Hidden Markov Models

A Hidden Markov Model is given by a time-labeled sequence of state variables $\ldots, X_{-1}, X_0, X_1, \ldots$, with transition probabilities $P(X_{t+1} \mid X_t)$, and a sequence of evidence variables $\ldots, Y_{-1}, Y_0, Y_1, \ldots$, with emission probabilities $P(Y_t \mid X_t)$. In a typical inference task, after observing values of Y (blue), we are interested in the value of X_{t+1} (beige) at time $t+1$:



For instance, variables X might be whether there is high or low air pressure, while observations Y are of sun or clouds. While in principle determining X_0—today's weather—could depend on indefinitely earlier observations and states at times $t = -1, -2, \ldots$, one has the intuition that "looking back" only a few days should be sufficient for typical purposes.

For a first illustration, consider a simple HMM with binary variables X_t and Y_t, and probabilities as follows:

$$P(X_{t+1} = 1 \mid X_t = 1) = P(X_{t+1} = 0 \mid X_t = 0) = 0.9$$

$$P(Y_t = 1 \mid X_t = 1) = P(Y_t = 0 \mid X_t = 0) = 0.8$$

Our class $\mathcal{P}$ of approximate distributions includes all truncations of the model at variable X_{t-N}, in which case we assume the distribution $P^*(X_{t-N})$ is uniform. In Figure 1 is a graph showing the KL-distance between the full model and a submodel with only N previous time steps included. We

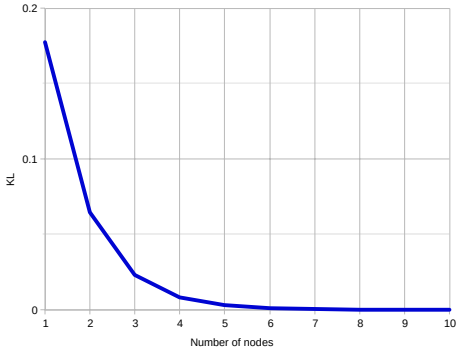


Figure 1: Dropoff in KL as function of number of nodes.

chose this particular model for illustration because the KL-distance is relatively high for the submodel with only one node. Nonetheless, even for this model, the value drops off rather dramatically with only a few additional nodes.

This model has a low mixing rate, as measured by the second eigenvalue (λ_2) of the transition matrix for the underlying Markov model (the transition probabilities). In general, a higher λ_2 value means a lower mixing rate, which means the past provides more information about the present. One might expect that in such cases it is more detrimental to ignore previous state variables. If we look at the graph (Figure 2) of KL-distances as a function of λ_2, holding fixed the observation probabilities as above, we see that this model (for which $\lambda_2 = 0.8$) is indeed near the higher end.

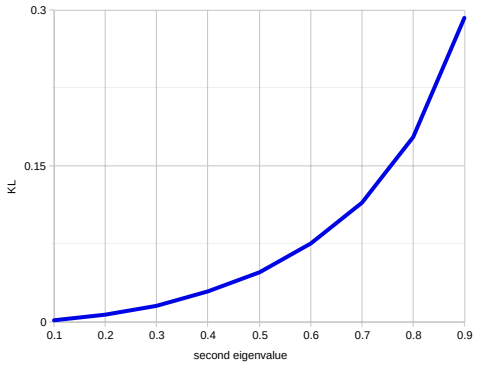


Figure 2: KL-distance for an approximate model with only one state variable, as a function of second eigenvalue λ_2.

If we now factor in the cost of including more nodes in the approximate HMM, we can determine, for different values of λ_2, and for different assumptions about cost of a node, what the optimal number of nodes to include will be, in line with Equation (1) above. To give one (arbitrary, but illustrative) example, let us assume the cost of an additional node to be 0.02, i.e., that this cost is equivalent to the utility of $\frac{1}{50}$ more bits of information. For relatively low values of λ_2, it is not

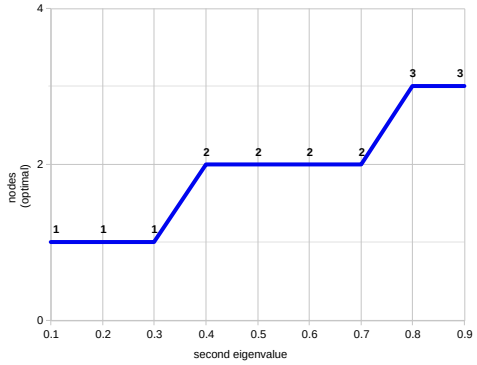


Figure 3: Optimal number of nodes, given a cost of 0.02 per node, as a function of second eigenvalue λ_2.

worth the cost to include more than a single state variable in the model (see Fig. 3). This is perhaps not surprising, given the low KL-values in Fig. 2. However, even in models with significantly higher KL-distances in general, it does not pay to include more than one or two additional nodes. As we increase or decrease the cost c of a node, the graph becomes less flat and flatter, respectively. For instance, provided that $c > 0.148$, the optimal number is 1 for all these values of $\lambda_2 = 0.1, \ldots, 0.9$. Decreasing the cost would increase the optimal number for larger values of λ_2, but for any $c > 0$ this number is of course still finite.

As Fig. 3 indicates, the optimal number of nodes to include in an HMM is not only finite, but can typically be quite small, in line with ordinary intuition.

Neglecting Alternative Causes

We next consider a simple causal model under a so called Noisy-Or parameterization (Cheng, 1997), in which each cause has independent causal power to bring about the effect. Suppose we have binary causal variables $\mathbf{X}$, taking on values 0 or 1, and conditional probabilities given in terms of weights $\theta_{Y,X}$ codifying the influence of parent Y on a variable X—in particular $\theta_{Y,X}$ gives the probability of Y causing X when Y is active—and a "background bias" parameter β:

$$P\big(X \mid \mathbf{pa}(X)\big) \quad = \quad 1 - \Big((1-\beta) \prod_{Y \in \mathbf{pa}(X)} (1-\theta_{Y,X})^{Y}\Big)$$

This model has the convenient property that deleting nodes from the graph still leaves us with a well-defined distribution. Hence the family of submodels is immediate from the full model (without having to introduce a proxy uniform distribution as in the previous example).

Suppose in particular we have variables A, B, C, D with weights θ_1, θ_2, and θ_3, as depicted on the left in Figure 4.

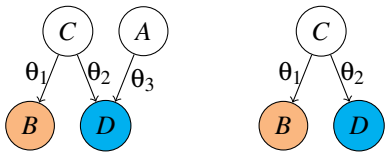


Figure 4: Full Model versus Partial Submodel

Imagine, for instance, a simple scenario in which these variables correspond to:

A: "Mary has lost her usual gloves"

B: "Mary has her bicycle with her"

C: "Mary is going cycling"

D: "Mary is wearing cycling gloves"

Observing that Mary is wearing cycling gloves makes it more likely that she is going cycling, and therefore that she has her bike with her. But this is attenuated by the alternative

possible cause, that she lost her other gloves. Our question in this case is, how much worse will a reasoner fare by ignoring alternative cause A (that Mary has lost her gloves), that is, by using the smaller submodel in Figure 4, on the right?

To assess this question, we can look at both the difference in expected utility and the KL-divergence between using the "true" distribution P and the approximate distribution P^*, for B given $D = 1$. Table 1 below presents example KL-values for different settings of model parameters: the priors on C and A, and the three weights θ_1, θ_2, and θ_3 (here and throughout this subsection, we set $\beta = 0.05$).

$P(C)$	$P(A)$	θ_1	θ_2	θ_3	$KL(P \,\|\| P^*)$
0.5	0.5	1	1	1	0.594
0.5	0.5	1	0.5	0.9	0.467
0.5	0.5	1	0.5	0.5	0.270
0.5	0.5	1	0.9	0.5	0.261
0.3	0.1	1	1	1	0.121
0.5	0.5	0.9	0.1	0.5	0.081
0.3	0.1	0.9	0.1	0.5	0.023
0.5	0.5	0.1	0.1	0.1	0.000

Table 1: Example KL-values, with $\beta = 0.05$.

Table 1 in fact shows settings near the higher end. We can also calculate the approximate *average* KL-divergence, over values of $P(C)$ and $P(A)$ at 0.1 intervals (0.1, 0.2, etc.), and 0.01 intervals for $\theta_1, \theta_2, \theta_3$ (thus, over 7 million parameter settings): for this model, it is 0.060. Averaging over parameters with $P(A)$ and $P(C)$ fixed at 0.5 gives a similar average KL-divergence of 0.059. Thus, the KL-value is typically well under one-tenth of a bit of information lost. Nonetheless, if confidence in estimation is important, or if very fine-grained decisions are called for, using the submodel may be detrimental in this case.

However, following question 1 from above, we may also consider how detrimental using a submodel will be for action choice in specific decision problems. For the EU calculation, suppose our agent is making a guess based on a single sample from either distribution P or P^*,[2] and that utility 1 is achieved for a correct response, 0 for incorrect. Example calculations are summarized in Table 2. As with KL, we can also compute the (approximate) average difference in EU, which for this model is 0.024. That is, on average over all parameter settings, a sampling agent will only suffer about $\frac{1}{50}$ of a utile by using the simpler submodel. The cost of including the additional variable A would therefore need to be extremely low, relative to utility in the given decision problem, to merit its presence in the model.

Evidently, when making a binary, sample-based decision, using the smaller submodel does not greatly reduce one's success probability. In fact, as shown in Table 1, for many parameter settings the agent actually fares better by using the simpler model ($\Delta_{P,P*} < 0$). Take the first parameter setting, for example. In this case $P(B = 1 \mid D = 1) \approx 0.673$, whereas $P^*(B = 1 \mid D = 1) \approx 0.955$. The submodel drastically overestimates the probability of B by ignoring the alternative cause, as reflected in the very high KL-value in Table 1. However, insofar as the true probability is significantly above 0.5, if the subject is going to make a decision by drawing a single sample from this distribution, such overestimation turns out to be advantageous, since the subject is more likely to choose the more probable outcome. Such advantages would only be compounded by the reduction in computational cost resulting from the simpler model.

$P(C)$	$P(A)$	θ_1	θ_2	θ_3	$\Delta_{P,P*}$
0.5	0.5	1	1	1	−0.097
0.5	0.5	1	0.5	0.9	−0.074
0.5	0.5	1	0.5	0.5	−0.087
0.5	0.5	1	0.9	0.5	−0.096
0.3	0.1	1	1	1	−0.073
0.5	0.5	0.9	0.1	0.5	−0.007
0.3	0.1	0.9	0.1	0.5	0.011
0.5	0.5	0.1	0.1	0.1	0.006

Table 2: Example EU-values, with $\beta = 0.05$.

We can tentatively conclude from this small case-study that alternative neglect—even when it results in less accurate judgments, which it certainly does—can still be a very reasonable, indeed resource-rational, strategy.

Predictive versus Diagnostic Reasoning

Resource-rational analysis of submodel choice and alternative neglect predicts these phenomena to occur, at least to a first approximation, when they would result in an optimal balance of outcome expected utility and computation cost, as outlined above. To what degree is this prediction born out by empirical data on alternative neglect?

One of the more robust recent findings in the causal reasoning literature is that subjects tend to neglect alternatives to a much greater extent in *predictive* reasoning than in *diagnostic* reasoning (Fernbach et al., 2011; Fernbach and Rehder, 2013). Most of the experiments in this work evince three variables A, B, C as in Figure 5. The left diagram depicts a

Figure 5: Predictive versus Diagnostic Inference

predictive inference, where effect B is queried given evidence that cause C is active. On the right is a diagnostic inference,

[2] This assumption is more apt in more complicated models, where computing exact estimates would be harder. We consider this kind of rule simply for illustration, and contrast with information distance, which would be more closely aligned with an agent (non-noisily) maximizing expected utility across decision problems.

where the causal variable C is queried given evidence B. In this simple model, we can fold $P(A)$ and $\theta_{B,A}$ into a single parameter θ_2, so that A effectively has prior probability 1, and the resulting conditional probabilities can be simplified to:

$$P(B \mid C) \;=\; \theta_1 + \theta_2 - \theta_1\theta_2$$

$$P(C \mid B) \;=\; 1 - \big(1 - P(C)\big)\Big(\theta_2 \;/\; \big(P(C)\theta_1 + \theta_2 - P(C)\big)\Big)$$

The finding in Fernbach et al. (2011) is that subjects routinely ignore variable A in predictive inference tasks, and thereby consistently make low estimates of $P(B \mid C)$. In diagnostic inference tasks, however, subjects show sensitivity to strength and number of alternatives, consequently making more accurate judgments. Indeed, there is a longer reaction time for diagnostic than for predictive inferences, and only in the diagnostic condition is there dependency of reaction time on number of alternatives (Fernbach and Darlow, 2010). Fernbach and Rehder (2013) verified that this asymmetry between diagnostic and predictive reasoning is robust; in particular, it seems not to be due to availability or memory limitations.

In other words, subjects seem to be reasoning with a submodel (ignoring variable A) in the predictive case, but not in the diagnostic case. How detrimental would it be to neglect A for these two types of inference? Consider first KL-divergence (question 2). Without a background bias term (as in the previous example), for the diagnostic case ignoring variable A will lead to the conclusion that C has probability 1, since it is the only possible cause. In that case, the KL-divergence is infinite. With a positive bias term β, we can make the KL-divergence finite, but it will still be large if the bias is small. For instance, with 1% chance of B happening spontaneously ($\beta = 0.01$), the average value of $KL(P \,||\, P^*)$ for the diagnostic inference is already 1.740, extremely high. With $\beta = 0.05$, it is 0.916.

By contrast, the average value of $KL(P \,||\, P^*)$ in the predictive case (even without a bias term, which would further decrease the average KL value) is only 0.357. This is a general observation about this particular small causal graph, which is the one implicated in many studies of "elemental causal induction." While it may be difficult to assess these KL-values absolutely, we can confirm that there is a substantial difference between the two types of inference. Indeed, on average one can expect to make much worse predictions in the diagnostic case than in the predictive case; an agent balancing computation cost with accuracy would then be expected to neglect the alternative more in predictive reasoning than in diagnostic reasoning.

How does this look from the perspective of expected utility, again assuming a single-sample-based agent? Table 3 shows the differences in expected utility for several parameter settings between the EU of using the true distribution and the approximate distribution (ignoring variable A), for both the predictive and diagnostic cases. In some cases, Δ_{pred} is indeed smaller than Δ_{diag}, meaning that the agent suffers less in expected utility when using the smaller submodel. However, there are also cases where Δ_{pred} is greater than Δ_{diag}.

$P(C)$	θ_1	θ_2	Δ_{pred}	Δ_{diag}
0.1	0.3	0.9	1.083	1.345
0.5	0.9	0.9	0.176	−0.041
0.5	0.5	0.1	0.010	−0.144
0.2	0.3	0.2	−0.034	0.345
0.1	0.3	0.1	−0.036	0.550

Table 3: Differences in expected utility between the true and approximate distributions, for predictive and diagnostic inferences (with $\beta = 0.05$ for diagnostic cases).

Indeed, the average Δ_{pred} value, over 0.01 intervals for all parameters, is 0.198—relatively large, and significantly greater than Δ_{diag}, whose average is 0.044. Again, this means that a subject drawing a single sample will on average lose about $\frac{1}{5}$ of a utile for predictive inferences on this graph, versus only about $\frac{1}{25}$ for diagnostic inferences.

This is in sharp contrast to an agent that decides based on exact inference (or equivalently, based on many samples). Based on the earlier results for KL-distance, we can compare action selection directly for agents that accurately maximize expected utility from the full- or sub-model. It turns out that for such agents the average difference in expected utility over all parameter settings is indeed significantly greater in the diagnostic case (0.378, versus 0.174 in the predictive case). If resource rationality is the correct explanation of these cases of alternative neglect, we would then have to conclude that participants are not using a single sample but rather many (or more generally are using some algorithm for computing closer approximations to the exact probabilities). Indeed, the models in Fig. 5 are rather simple and one might expect that inference for these models is relatively easy. At any rate, positing that one can well-approximate the true probability for a given model, if one were to ignore causal variables routinely in one type of case (diagnostic or predictive) but not the other, it would be most rational to do so in the predictive case, as subjects in fact do.

We might therefore tentatively conclude that, at a certain level of grain, subjects are ignoring variables in a reasonable way. However, this does not yet say anything about resource-rationality at the level of individual inferences. In fact, Fernbach and Rehder (2013) have shown that subjects exhibit neglect even in cases where the mistake is rather serious, resulting in egregiously wrong predictions. One possible explanation is that subjects are optimizing grain of representation only at a high level of abstraction, in terms of general features of the inference problem (e.g., whether it is predictive or diagnostic). This hypothesis merits further empirical investigation, as well as further theoretical consideration, for instance, by incorporating elements of metalevel control (Icard, 2014; Lieder et al., 2014) into the framework. The analysis offered here, we believe, promises a useful starting point for understanding how rational submodels might be selected online, and for addressing this question of level of grain.

Further Questions and Directions

The *resource-rational submodel* is that submodel of the true, larger model which an agent ought to use for a given purpose, balancing costs with accuracy. As we have seen, this approach to the causal frame problem does already shed light on a number of phenomena: it agrees with our pretheoretical intuition that only a few previous time steps should matter in an HMM; it interacts in subtle ways with different assumptions about choice rules, witness the single-sampling agent in the second example; and it retrodicts general alternative neglect patterns observed in human causal reasoning.

The claim we would like to make is that, somehow or other, the mind is able to focus attention on the predicted submodels for the task at hand. Our analysis, however, is from a "god's eye" point of view, and leaves open how is the mind able to select the right submodel online, at inference time. There is relevant empirical work which might start to provide a process-level explanation of submodel selection. For example, across many situations subjects seem to "add" a new causal variable when faced with an apparent contradiction (Park and Sloman, 2013). At the same time, work in AI suggests useful heuristics for similar resource optimization problems (e.g., Wick and McCallum, 2011).

A number of extensions to our analysis suggest themselves. One could consider other types of submodels, e.g., not by eliminating nodes, but by cutting links. One could also consider more seriously the interaction of submodel choice with inference algorithm choice. As hinted in the previous section, it would be useful to consider what the optimal *combination* of submodel together with number of samples will be for a given model (thus combining the analysis here with that in Vul et al. 2014). Assuming smaller submodels will allow for more samples per unit of time (or energy), and given that more samples lead to more accurate predictions with respect to that model, there is a substantive question of what combination is optimal in any case.

While many questions remain open, we hope to have made some progress toward illuminating a general solution to the causal frame problem. We believe the question is central to our understanding of human reasoning. As Fodor remarked, "The frame problem goes very deep; it goes as deep as the analysis of rationality" (Fodor, 1987).

Acknowledgements

Thanks to Andreas Stuhlmüller and Long Ouyang for comments on a draft of this paper.

References

Bernardo, J. M. and Smith, A. M. (1994). *Bayesian Theory*. John Wiley and Sons.

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

Fernbach, P. M. and Darlow, A. (2010). Causal conditional reasoning and conditional likelihood. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pages 1088–1093.

Fernbach, P. M., Darlow, A., and Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140:168–185.

Fernbach, P. M. and Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument and Computation*, 4(1):64–88.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, 4(2):330–344.

Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In Pylyshyn, Z. W., editor, *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, pages 139–149. Ablex.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–699.

Glymour, C. (1987). Android epistemology and the frame problem. In Pylyshyn, Z. W., editor, *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, pages 65–75. Ablex.

Griffiths, T. L., Lieder, F., and Goodman, N. D. (2014). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*. forthcoming.

Icard, T. F. (2014). Toward boundedly rational analysis. In Bello, P., Guarini, M., McShane, M., and Scassellati, B., editors, *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 637–642.

Klayman, J. and Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2):211–228.

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N. J., and Griffiths, T. L. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in Neural Processing Systems*.

Park, J. and Sloman, S. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, 67:186–216.

Simon, H. A. (1957). *Models of Man*. Wiley.

Vul, E., Goodman, N. D., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4):699–637.

Wick, M. L. and McCallum, A. (2011). Query-Aware MCMC. In *25th Conference on Neural Information Processing Systems*, pages 2564–2572.