Unintended Consequences: Updating Causal Models

Joseph Y. Halpern[1] , **Evan Piermont**[2] , **Marie-Louise Vierø**[3]
[1]Computer Science Department, Cornell University, Ithaca, USA.
[2]Department of Economics, Royal Holloway, University of London, UK.
[3]Department of Economics and Business Economics, Aarhus University, Denmark.
halpern@cs.cornell.edu, evan.piermont@rhul.ac.uk, mlv@econ.au.dk

Abstract

We examine how causal beliefs affect an agent's choices and how feedback on those choices leads to updated causal beliefs. Building on the structural-equations framework for modeling causality, we first examine the general problem of updating causal beliefs in the face of novel (and possibly inexplicable) data. We model an agent who is uncertain of the true causal model, and therefore entertains a probabilistic belief over the set of possible models. We then consider how causal beliefs influence choices by building a model of agency and utility on top of the usual structural-equations framework. Using these two components, we propose a notion of *steady state*, where the feedback received from an agent's optimal action, given her current beliefs about the true causal model, can be rationalized by those beliefs.

1 Introduction

Causal reasoning is difficult; agents routinely misunderstand, misinterpret, or simply overlook the true causal relationships between variables of interest. Examples of such erroneous judgments are plentiful, ranging from the pedestrian (superstitiously carrying a lucky charm) to the existential (misjudging humanity's effect on global climate). Because an agent will decide which actions to take based on her understanding of the *effect* of her actions, errors in causal reasoning are costly, leading to suboptimal or inefficient choices. An agent may entertain different causal models, each reflecting a different understanding of the causal relationships, and hold probabilistic beliefs over these. In dynamic or repeated environments, an agent who entertains incorrect causal judgements might observe evidence that contradicts her initial understanding, requiring her to re-evaluate her beliefs. However, as we show, this re-evaluation may not result in correct beliefs or stable behavior, even after an infinite number of interactions.

In this paper, we examine more generally how causal reasoning affects an agent's choices and how feedback on those choices leads to updated causal beliefs. In order to formalize causal reasoning, we draw on the literature on structural causal models à la [Pearl, 2000]. We first examine the general problem of updating causal beliefs in the face of novel (and possibly inexplicable) data. We model an agent who is uncertain of the true causal model, and therefore entertains a probabilistic belief over the set of possible models. To model the agent's reaction to inexplicable evidence (that is, when she observes outcomes that cannot be explained by any model in the support of her beliefs), we endow the agent with a *conditional probability system (cps)* [Popper, 1934; Popper, 1968; de Finetti, 1936], encoding the agent's behavior conditional on any (including potentially probability 0) observation.[1]

We then consider how causal beliefs influence choices by building a model of agency and utility on top of the usual structural-equations framework. We model an agent who can take actions that consist of intervening on a set of variables and making partial observations about the outcome of the intervention. The agent derives utility as a function of the values of the variables, and thus her ability to intervene has a direct consequence on her utility. In dynamic environments, the agent also derives value from her observation regarding the effect of interventions, as such information forms the basis of belief updating, and therefore allows for better future decision making.

Using these two components, we propose a notion of *steady state*, where the feedback received from an agent's optimal action, given her current beliefs about the true causal model, can be rationalized by those beliefs. In such a case, the optimal action does not induce the agent to revise her beliefs, and therefore, the action remains optimal. To get a sense of the relation between an agent's choice of action, the data that is generated, and the re-evaluation of beliefs, consider the following two examples. In both, the protagonist takes the incorrect action (as judged from the modeler's perspective) on the basis of their faulty causal reasoning; the difference lies in how the agents react to the observed evidence:

Example 1. *During British East India Company rule of colonial India, to combat the proliferation of dangerous serpents, the Governor came up with what appeared an ingenious solution: pay a bounty to any man who brought in the head of a dead cobra. This "cash for snakes" program worked for a*

[1]We interpret the beliefs conditional on probability 0 events as the analyst's model of the agent's behavior, and not as beliefs consciously held by the agent. In other words, the agent need not anticipate how she will react to inexplicable evidence.

while, but was followed by a marked uptick in the number of cobras plaguing the city. Indeed, entrepreneurial locals began breeding snakes for the express purpose of collecting the bounty, but, owing to lax security, many snakes escaped into the wild, dramatically growing the population.

The Governor *learns* that his action was misguided, since the observation that arises subsequent to his choice of action cannot be rationalized by his initial understanding. Our second example, by contrast, shows that an agent can remain in perpetual delusion, if the evidence generated by her action accords perfectly with her beliefs, despite the fact that those beliefs are incorrect.

Example 2. *Since time immemorial, members of the* Aition *tribe have woken before sunrise to perform an elaborate ritual to their God* Apotelesma*. They have never missed a single day, for they operate under the (fallacious) conviction that Apotelesma is vindictive and petty, and without their daily votive, she would not allow the sun to rise. Every day, in the daylight after their morning prayer, the Aitions conclude their model of the world must be correct as it accurately predicted the sunrise.*

1.1 Related Literature

Understanding the causal relationships between variables has been the focus of considerable work across many fields of study (see, for example, [Angrist and Pischke, 2009; Cunningham, 2021; Hernán and Robins, 2020; Morgan and Winship, 2007; Parascandola and Weed, 2001; Raina K. Plowright, 2008; Pearl, 2009; Pearl, 2000; Spirtes *et al.*, 1993], noting that this list is very cursory). Despite the considerable attention paid to both modeling and uncovering causal relationships in general, there has been much less focus on modeling subjective causal reasoning as a component of decision making. Notable exceptions include [Schenone, 2020; Ellis and Thysen, 2021; Alexander and Gilboa, 2023], who all relate subjective notions of causality to decision making, although not within the structural-equations framework.

Halpern and Piermont [2024] consider an agent who has a preference relation over interventions and provide axioms characterizing "causal sophistication," that is, when the agent's preferences can be rationalized by a utility function over variable values and a belief over structural equation models. Their representation theorem can be seen as a decision-theoretic foundation for the current framework, showing that the agent's utility function and (current) beliefs can be elicited from behavioral data.

Psychologists have studied how humans actually learn causal relationships. In simple, controlled environments, there is ample evidence that subjects' update their beliefs in a manner that approximates Bayesianism, albeit under some computational constraints [Griffiths and Tenenbaum, 2005; Kushnir and Gopnik, 2007; Bonawitz *et al.*, 2014; Coenen *et al.*, 2019]. It is less clear, however, how belief formation works in larger and less controlled settings, where the space of potential causal models is too large or unstructured to be considered all at once, or where the agent is simply unaware of some relevant relationships. Bramley et al. [2017] argue that people keep things simple, and "maintain only a single hypothesis about the global causal model, rather than a distribution over all possibilities and update their hypothesis by making local changes" (where, by "local changes", they mean "small changes to the causal model under consideration", such as adding more parents to a node or changing the direction of causality in one case).

Our paper lies within the more general literature on misspecified models, where an agent persistently entertains the wrong model governing data generation or the resolution of uncertainty. This topic has seen considerable recent interest within economics; see, for example, [Mailath and Samuelson, 2020; Eliaz *et al.*, 2020; Ellis and Thysen, 2021]. A consistent theme in these papers is that agents' beliefs will often fail to converge (and in particular, fail to converge to the truth). We contribute to this literature by focusing on two novel aspects: first, we model an agent's causal judgments, that is, her beliefs about the effect of her actions. Second, we consider the possibility that the agent abandons her initial beliefs in the face of inexplicable evidence. Model-misspecification has also been studied in strategic environments [Fudenberg and Levine, 1993; Eyster and Rabin, 2005; Esponda and Pouzo, 2016; Piermont and Zuazo-Garin, 2026]. While we do not consider strategic environments here, the extension of our framework to this setting seems straightforward.

Our approach also bears similarity to the literature on belief updating in the face of unawareness, e.g., [Karni and Vierø, 2013; Halpern and Piermont, 2020; Piermont, 2021]. In particular, we can interpret an agent who systematically neglects the causal influence of a particular variable as being unaware of the variable (or at least unaware of its relevance). For example, we could interpret the Governor in Example 1 as being unaware of the possibility that the local population might breed snakes.

2 Model

We begin by describing the general framework of structural causal models and how we model an agent's beliefs about them. At a high level, we assume a dynamic environment, where, in each period, a causal model stochastically determines the realization of all variables. The agent does not know the true causal model, but entertains probabilistic beliefs about it. Each period, the agent derives utility directly from the realized values of variables, and in addition, can take actions to intervene on variables and (partially) observe the effect of her intervention. Through this process she updates her beliefs.

The three main ingredients of our model, described in detail in the remainder of this section are: (a) a universe of variables and possible causal relations between them embodied by a set of possible structural equations; (b) a conditional probability system on this space of all structural equations that governs the agent's beliefs (and eventually her behavior); (c) the actions that are available to the agent and a notion of utility that directs the agent's choice of action.

2.1 Structural-Equations Models

We assume that the world is described by a set of variables. It is useful to split them into two sets: the *exogenous* vari-

ables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are determined by the exogenous variables and other endogenous variables. We assume that agents can intervene only on endogenous variables (but not necessarily all of them).

Let $\mathcal{U}$ and $\mathcal{V}$ denote the finite collection of exogenous and endogenous variables, respectively, that describe the environment. For each $X \in \mathcal{U} \cup \mathcal{V}$, let $\mathcal{R}(X)$ denote the finite set of values X can take. A *context* is a vector $\vec{u}$ of values for all the exogenous variables $\mathcal{U}$. Let $\mathcal{R}(\mathcal{U}) = \prod_{Y \in \mathcal{U}} \mathcal{R}(Y)$ consist of all contexts.

We use the standard vector notation to denote sets of variables or their values, that is, if $\vec{X} = (X_1, \ldots, X_n)$ and $\vec{Y} = (Y_1, \ldots, Y_m)$ are vectors, we write, in a standard extension of the usual containment relation, $\vec{X} \subseteq \mathcal{U} \cup \mathcal{V}$ to indicate that each $X_i \in \mathcal{U} \cup \mathcal{V}$, or, $\vec{X} \subseteq \vec{Y}$ to indicate that for each $i \leq n$, there is some $j \leq m$ such that $X_i = Y_j$. Similarly, we write $\vec{x} \in \mathcal{R}(\vec{X})$ if $\vec{x} \in \prod_{i \leq n} \mathcal{R}(X_i)$. As with the notation for random variables, uppercase letters refer to the names of variables and lowercase to their values. If $\vec{X} \subseteq \vec{Y}$ and $\vec{y} \in \mathcal{R}(\vec{Y})$, let $\vec{y}|_{\vec{X}}$ denote the restriction of $\vec{y}$ to $\vec{X}$. In particular, for a single variable X, $\vec{y}|_X$ denotes the X^{th} component of $\vec{y}$.

A *causal model* is a pair $M = \langle \mathcal{S}, \mathcal{F} \rangle$, where $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$ is a *signature* and $\mathcal{F}$ is a collection of structural equations that determine the values of endogenous variables based on the values of other variables. Formally, $\mathcal{F} = \{F_X\}_{X \in \mathcal{V}}$, where

$$F_X : \prod_{Y \in \mathcal{U} \cup (\mathcal{V} - \{X\})} \mathcal{R}(Y) \to \mathcal{R}(X).$$

Given $\mathcal{F}$, we say that X *depends on* Y if, there is some setting of the variables in $\vec{Z} = (\mathcal{U} \cup \mathcal{V}) - \{X, Y\}$ and values $y, y' \in \mathcal{R}(Y)$ such that $F_X(\vec{z}, y) \neq F_X(\vec{z}, y')$; that is, changing the value of Y changes the value of X. A causal model is *acyclic* if there are no cyclic dependencies. Following the literature, we restrict attention in this paper to acyclic models.

We can represent qualitative features of a causal model by a *causal graph*, where the nodes are labeled by distinct (exogenous or endogenous) variables, and there is an edge from X to Y if Y depends on X. Note that the graph of an acyclic model is acyclic. Let $\mathcal{M}$ denote the set of all acyclic models (note that $\mathcal{M}$ is finite).

A *setting* is a pair $(M, \vec{u})$ consisting of a model and a context. A setting $(M, \vec{u})$ determines the value of all variables, i.e., there is a unique value for each variable that simultaneously satisfies all the structural equations in M and agrees with the context $\vec{u}$. We write $(M, \vec{u}) \models \vec{Y} = \vec{y}$ for the vector $\vec{y} \in \mathcal{R}(\vec{Y})$ if $\vec{y}$ is the value of $\vec{Y}$ in $(M, \vec{u})$. See [Halpern, 2000] for details about the semantics of structural models.

We denote the intervention on variable $X \in \mathcal{V}$ that sets its value to $x \in \mathcal{R}(X)$ by $X \leftarrow x$, and likewise for a set of variables $\vec{X} \leftarrow \vec{x}$. This constructs a new *counterfactual* causal model (over the same set of variables, and consistent with the same causal graph) where the equation for each variable $X \in \vec{X}$ is just the constant function $X = x$. We denote the counterfactual model constructed by starting with M and performing intervention $\vec{X} \leftarrow \vec{x}$ as $M_{\vec{X} \leftarrow \vec{x}}$.

2.2 Beliefs and Updating

There are two sources of uncertainty: about the true model (that is, about the causal relationship between variables) and about the context (that is, about the realizations of exogenous variables). We assume that there exists a true (but unknown) model $M^\star$ that is fixed across time, so that the true causal relationships between variables remains the same at every period. By contrast, we assume that the context is stochastic, drawn i.i.d. at every period according to a probability distribution over the set of contexts, $\pi \in \Delta(\mathcal{R}(\mathcal{U}))$. Therefore, the realization of the variables (without any intervention) is the outcome of $(M^\star, \vec{u})$ where $\vec{u}$ is distributed according to π. The asymmetry in how we model these two sources of uncertainty captures the difference between uncertainty that arises because of persistent randomness in the system (governed by π) and the uncertainty that arises because of the agent's epistemic state (governed by the agents beliefs about the true model, as described below).

The agent may not know the true causal model $M^\star$, but we assume that the agent does know π, at least to the extent of determining those exogenous variables she is aware of.[2] Thus, if the agent learned the true model, she would learn the true *distribution* over variable values. She would still be unable, however, to deterministically predict the outcome, owing to the randomness of the context.[3]

We now describe how we model beliefs about the true model. Let $E \subseteq \mathcal{M}$ be a non-empty subset of models. Then a *conditional probability system (cps) over* E is a family of probability measures on the non-empty subsets of E, indexed by non-empty subsets of E; that is, $\mu = \{\mu_F\}_{\emptyset \neq F \subseteq E}$, such that for all $F'' \subseteq F' \subseteq F \subseteq E$, we have

CP1. $\mu_F(F) = 1$

CP2. $\mu_F(F'') = \mu_{F'}(F'') \times \mu_F(F')$

The measure μ_F represents the conditional beliefs given evidence of the event F. The two consistency requirements state that beliefs are as dynamically coherent as possible: (CP1) states that the evidence is treated as fact while (CP2) extends a standard probabilistic identity to ensure coherence for different distributions μ_F.

A cps also allows beliefs conditional on events of unconditional probability 0, where the unconditional probability is given by μ_E. To simplify notation, we denote μ_E by $\bar{\mu}$. The probability $\bar{\mu}$ completely governs the (contemporaneous) behavior of the agent, as she does not anticipate observing inexplicable evidence; the other components of the cps determine only her future behavior should her contemporaneous beliefs prove incorrect.

[2] The set of causal models considered by the agent may exclude some variables, including some of the exogenous variables. Formally, we assume that the agent knows the marginal of π on those variables that are included in the models she considers possible.

[3] Conceptually, it is straightforward to also allow the agent to be uncertain about the distribution π over the realization of contexts. We do not include this additional layer of uncertainty as it increases complexity and is largely unrelated to the main focus of the paper.

Shortly, we will endow the agent with the ability to take actions that consist of intervening on some set of variables and observing the effect of this intervention on (a different) set of variables. The information generated by such an action is an *observation*; formally an *observation* is a pair $o = \langle \vec{X} \leftarrow \vec{x}, \vec{Y} = \vec{y} \rangle$ where $\vec{X} \subseteq \mathcal{V}$ is the set of variables that are intervened on to be set to $\vec{x} \in \mathcal{R}(\vec{X})$ and $\vec{Y} \subseteq \mathcal{U} \cup \mathcal{V}$ is the set of variables that are observed to take values $\vec{y} \in \mathcal{R}(\vec{Y})$ as a result of the intervention. We require that if $Y \in \vec{X} \cap \vec{Y}$, then $\vec{y}|_Y = \vec{x}|_Y$. Let $\mathcal{O}$ denote the set of all observations.

Let the agent's beliefs be given by a cps μ. How should the agent update her beliefs subsequent to the observation $o = \langle \vec{X} \leftarrow \vec{x}, \vec{Y} = \vec{y} \rangle$? In an abuse of notation, let

$$\pi(M, o) := \pi(\{\vec{u} \mid (M_{\vec{X} \leftarrow \vec{x}}, \vec{u}) \models \vec{Y} = \vec{y}\})$$

denote the π-probability that model M would generate the observation o. The agent can rule out any model not in the set

$$E(o) := \{M \in \mathcal{M} \mid \pi(M, o) > 0\}.$$

If the agent observes o, then she has *learned* the event $E(o)$. For example, let $E_0 = \mathcal{M}$ be the set of all models. After observing o_1, the agent will learn $E(o_1)$, and entertain new beliefs modeled by a cps over $E_1 = E(o_1) \cap E_0$; after another observation o_2, the agent will learn $E(o_2)$, and the new cps will be over $E_2 = E(o_2) \cap E_1 = E(o_2) \cap E(o_1) \cap E_0$, etc.

When observing o, however, the agent learns more that *just* $E(o)$: the relative likelihood of different models producing the observation yields further belief updating. For any $F \subseteq E(o)$, let μ^o_F be the distribution defined by

$$\mu^o_F(M) := \frac{\mu_F(M)\pi(M, o)}{\sum_{M' \in F} \mu_F(M')\pi(M', o)},$$

which captures the posterior probability of models using each conditional distribution as a potential prior.

We can thus define an 'updated' cps as

$$\mu^o := \{\mu^o_F\}_{F \subseteq E(o)},$$

which can easily be checked to be a well-defined cps over $E(o)$ whenever $E(o) \cap E \neq \emptyset$.

We conclude this section with a few special cases:

- The distribution π is trivial, that is, a point mass on a single context: We call such a state of affairs *deterministic*.
- The supports of each μ_F is a singleton. After any conditioning event, the agent is certain of a particular model. The results of Bramley et al. [2017] suggest that this is what people actually do.
- The cps is over $\mathcal{M}$ and $\mu_{\mathcal{M}}$ is fully supported. The agent entertains all possible models and is never surprised. We could model such an agent without the cps machinery.
- The support of each μ_F is the set of all models defined on a causal graph $\mathbf{g}$. After an observation the agent is certain about the direction of causal influence, but not the exact details of the structural equations. She only re-evaluates her world view when she cannot explain the data with her given graph $\mathbf{g}$.

2.3 Agency and Utility

Each period, the agent can take an action; formally, an *action* is an intervention paired with a set of variables to observe: $a = (\vec{X} \leftarrow \vec{x}, \vec{Y})$. The interpretation is that by taking the action a the agent intervenes so as to set the variables $\vec{X} \subseteq \mathcal{V}$ to the values $\vec{x} \in \mathcal{R}(\vec{X})$. In addition, the action a allows the agent to (partially) observe the effect of the intervention; she observes the value $\vec{y} \in \mathcal{R}(\vec{Y})$ that results from the intervention. Let A denote the set of all actions available to the agent.

Given a setting $(M, \vec{u})$, each action $a = (\vec{X} \leftarrow \vec{x}, \vec{Y})$ determines the observation that would be made: if a is taken, then given $(M, \vec{u})$, the agent will expect to observe

$$o^a_{M,\vec{u}} := (\vec{X} \leftarrow \vec{x}, \vec{Y} = \vec{y},)$$

where $\vec{y} \in \mathcal{R}(\vec{Y})$ is the unique vector of values such that $(M_{\vec{X} \leftarrow \vec{x}}, \vec{u}) \models \vec{Y} = \vec{y}$. Let $\mu^a_{M,\vec{u}} := \mu^{o^a_{M,\vec{u}}}$ denote the cps that results from updating on this observation.

Finally, let $\textsc{Util} \in \mathcal{V}$ denote a distinguished variable which we interpret as utility. In an abuse of notation, we can treat $\textsc{Util}$ as a utility function over outcomes of variables as given by settings. This is defined by $\textsc{Util}(M, \vec{u}) = x$ for the $x \in \mathcal{R}(\textsc{Util})$ such that $(M, \vec{u}) \models \textsc{Util} = x$. While it is not required by what follows, it is perhaps reasonable to assume that for all $(\vec{X} \leftarrow \vec{x}, \vec{Y}) \in A$, we have $\textsc{Util} \in \vec{Y}$; that is, independent of whatever else is observed, the agent can determine her own utility.

For an action $a = (\vec{X} \leftarrow \vec{x}, \vec{Y}) \in A$, let $M^a = M_{\vec{X} \leftarrow \vec{x}}$, be the counterfactual model induced by the action a. With this notation, we can further extend $\textsc{Util}$ to capture the agent's (one-shot) value of taking action a, given her beliefs μ:

$$\textsc{Util}(a, \mu) := \sum_{\mathcal{M} \times \mathcal{R}(\mathcal{U})} \textsc{Util}(M^a, \vec{u})\, \pi(\vec{u})\, \bar{\mu}(M).$$

The decision maker, however, does not only care about her one-shot value but also her ability to continue to make choices. Given a discount factor of $\delta \in (0, 1)$, the agent values the action a according to its one-shot utility and its information value. This leads to our recursive definition of $\textsc{Val}$, as is standard in infinite horizon dynamic programming.

$$\textsc{Val}(a, \mu) := \sum_{\mathcal{M} \times \mathcal{R}(\mathcal{U})} \big(\textsc{Util}(M^a, \vec{u}) + \delta \max_{a' \in A} \textsc{Val}(a', \mu^a_{M,\vec{u}})\big)\pi(\vec{u})\bar{\mu}(M).$$

This is well defined on the basis that $\textsc{Util}$ is bounded and $\delta < 1$ (see, e.g., Theorem 4.6 of [Stokey and Lucas Jr., 1989]). The value of taking action $a \in A$ today is the expectation of

- $\textsc{Util}(a, \mu)$ — the immediate payoff associated with the intervention according to today's utility, and
- $\delta \max_{a' \in A} \textsc{Val}(a', \mu^a_{M,\vec{u}})$, the discounted value of the same problem tomorrow, except that beliefs will be updated according to the additional observation made. Recall that $o^a_{M,\vec{u}}$ is the observation that is realized when action a is taken given model M and context $\vec{u}$, so $\mu^a_{M,\vec{u}}$ is the agent's updated beliefs conditional on what she observes as part of her chosen action.

Note that we are assuming, by nature of the "max" operator in the continuation value (i.e., tomorrow's utility), that the agent optimally uses whatever information arrives to maximize the continuation value. Note, the *actual* observation made by the agent is determined by $M^\star$. So, if $M^\star$ is not in the support of $\bar{\mu}$, it is possible that the true observation $o^a_{M^\star,\vec{u}}$ does not coincide with any $o^a_{M,\vec{u}}$ considered with positive probability; that is, $\bar{\mu}(E(o^a_{M^\star,\vec{u}})) = 0$. This is *not* considered possible by the agent (i.e., it is a 0-probability event, and therefore does not contribute to the choice of $a \in A$), but is possible from the modeler's perspective. However, the agent can still condition on the event using her cps.

3 Steady State

We now introduce the idea of a steady state, where the optimal action today does not alter beliefs, so that it remains optimal tomorrow. Therefore no new information enters the system and actions are repeated. Using the notation of the previous section, the environment (given a set of variables and ranges) is described by the tuple

$$(M^\star, \pi, A, \textsc{Util}, \delta).$$

Given this, a *steady state* is a pair (μ, a) consisting of a cps μ, and an action a such that:

- $a \in \arg\max_{a' \in A} \textsc{Val}(a', \mu)$;
- $\bar{\mu} = \bar{\mu}^a_{M^\star,\vec{u}}$ for $\vec{u} \in \mathcal{R}(\mathcal{U})$ with $\pi(\vec{u}) > 0$.

The first condition states that the agent is choosing an optimal action *given her beliefs*. The second condition states that, given this action and the true model, the realized observation does not change her *unconditional* beliefs—hence the same action will remain optimal in the future.

The character of a steady state is quite straightforward in deterministic environments (i.e., when π is trivial), as shown by the following two remarks, neither of which holds in the more general stochastic environment, as we show below.

Remark 1. *Suppose that π is trivial. If (μ, a) is a steady state, then $a \in \arg\max_{a' \in A} \textsc{Util}(a', \mu)$.*

Proof. Let $\vec{u}$ be the unique context in the support of π. Then, since observing $o^a_{M^\star,\vec{u}}$ does not change her beliefs, it must be that $o^a_{M,\vec{u}} = o^a_{M^\star,\vec{u}}$ for every model M in the support of $\bar{\mu}$. As such, with $\bar{\mu} \times \pi$-probability 1, $\textsc{Val}(\cdot, \mu^a_{M^\star,\vec{u}}) = \textsc{Val}(\cdot, \mu)$. The remark follows immediately. □

This remark does not hold if the context is stochastic. In particular, it is possible that an agent would choose a statically suboptimal action because she thinks there is some probability it will be informative (i.e., there are some contexts in which it would allow her to discriminate between models), while in fact (i.e., according to $M^\star$) it is not informative for any context.

Our second remark shows that, in a deterministic environment, the agent's dynamic behavior will converge to a steady state.

Remark 2. *Assume that π is trivial and let $\vec{u}$ be the unique context in the support of π. Let μ be any cps over E for some E containing the true model $M^\star$. Starting with $\mu_0 = \mu$, define inductively the following sequence of actions and beliefs: for each $i \geq 1$, take*[4]

- $a_i \in \arg\max_{a' \in A} \textsc{Val}(a', \mu_{i-1})$;
- $\mu_i = (\mu_{i-1})^{a_i}_{M^\star,\vec{u}}$.

Then (μ_i, a_i) converges to a steady state as $i \to \infty$.

Proof. Given the triviality of π, $\bar{\mu}_i = \bar{\mu}_{i-1}$ if and only if $o^{a_i}_{M,\vec{u}} = o^{a_i}_{M^\star,\vec{u}}$ for every model in the support of $\bar{\mu}_{i-1}$; as such, any belief change corresponds to the agent excluding some previously considered model. Since there is a finite number of models, it follows that beliefs are eventually constant; let $k > 0$ be such that for all $i \geq k$, $\bar{\mu}_i = \bar{\mu}_k$. Then by construction we have

- $a_{k+1} \in \arg\max_{a' \in A} \textsc{Val}(a', \mu_k)$ and
- $\bar{\mu}_k = \bar{\mu}_{k+1} = (\bar{\mu}_k)^{a_{k+1}}_{M^\star,\vec{u}}$,

and so, a steady state. □

As mentioned above, Remark 2 does not hold in general, when there is uncertainty about the context. In fact, not only might repeated maximization not lead to a steady state, it does not necessarily even lead to stable behavior. In other words, it is possible that the agent's optimal action cycles between alternatives forever. We provide a simple example of such an environment in Example 3.

4 Examples

To get a sense of what a steady state entails, we first formalize the examples from the introduction within this framework. We also provide an example of the failure of convergence for non-deterministic environments. As it is relatively simpler, we begin with Example 2.[5]

Example 2. *(continued) There are three endogenous variables, $\mathcal{V} = \{\textsc{Util}, \textsc{pray}, \textsc{sun}\}$ and no exogenous variables $\mathcal{U} = \emptyset$. Both (non-utility) variables have range $\{0, 1\}$. The first, PRAY, represents whether the Aitions make their daily votive ($\textsc{pray} = 1$ if they do) and the second, SUN, whether the sun rises ($\textsc{sun} = 1$ if it does). Consider the following two structural models (which differ only in the equation that governs SUN):*

$$M^\star : \textsc{pray} = 0, \quad \textsc{sun} = 1, \quad \textsc{Util} = (10 \times \textsc{sun}) - \textsc{pray}$$

and

$$M^\dagger : \textsc{pray} = 0, \quad \textsc{sun} = \textsc{pray}, \quad \textsc{Util} = (10 \times \textsc{sun}) - \textsc{pray}.$$

The true model is $M^\star$. The Aitions, however, consider $M^\dagger$. Let μ be a cps such that $\bar{\mu}(M^\dagger) = 1$ and, for any E with $M^\star \in E$ and $M^\dagger \notin E$, $\mu_E(M^\star) = 1$.

[4] We must assume ties are broken consistently: if $\textsc{Val}(\cdot, \mu_j) = \textsc{Val}(\cdot, \mu_{j'})$ then $a_{j+1} = a_{j'+1}$.

[5] In these examples, for ease of exposition, we omit the exogenous variables, which determine the relevant endogenous variables in the obvious way.

The Aitions have two actions at their disposal:

$$a_0 = (\text{PRAY} \leftarrow 0, \mathcal{V}) \quad \textit{and} \quad a_1 = (\text{PRAY} \leftarrow 1, \mathcal{V}).$$

The associated model-dependent predicted observations are:[6]

$$o^0_{m^\star} = \langle \text{PRAY} \leftarrow 0, (\text{PRAY} = 0, \text{SUN} = 1)\rangle,$$
$$o^0_{m^\dagger} = \langle \text{PRAY} \leftarrow 0, (\text{PRAY} = 0, \text{SUN} = 0)\rangle,$$
$$o^1_{m^\star} = \langle \text{PRAY} \leftarrow 1, (\text{PRAY} = 1, \text{SUN} = 1)\rangle,$$
$$o^1_{m^\dagger} = \langle \text{PRAY} \leftarrow 1, (\text{PRAY} = 1, \text{SUN} = 1)\rangle.$$

The observation generated by action a_1 does not depend on the model, whereas the observation generated by a_0 is model dependent. Thus, starting from an initial belief $\bar{\mu}$, the action a_1 generates an observation that does not update beliefs at all; that is $\bar{\mu}^{a_1}_{M^\star} = \bar{\mu}$. On the other hand, observing $o^0_{m^\star} = \langle \text{PRAY} \leftarrow 0, (\text{PRAY} = 0, \text{SUN} = 1)\rangle$ cannot be explained by $\bar{\mu}$, leading to wholesale belief revision to $\mu_E(o^0_{m^\star})$, and so, given the structure of μ, adoption of correct beliefs.

Finally, given $\bar{\mu}$, the immediate value of each action is given by $\text{UTIL}(a, \mu) = \text{UTIL}(M^{\dagger,a})$*, so that* $\text{UTIL}(a_1, \mu) = 9$*, whereas* $\text{UTIL}(a_0, \mu) = 0$*. So, (μ, a_1) is a steady state: given their initial beliefs, the Aitions will continue to pray every morning, and consequently, never update their beliefs.*[7]

In a similar manner, we can formalize Example 1:

Example 1. *(continued) There are five variables, $\mathcal{V} = \{\text{UTIL}, \text{SNAKE}, \text{BOUNTY}, \text{BREED}, \text{HUNT}\}$, where* SNAKE *represents the snake population (with range $\{0, 1, 2\}$),* BOUNTY *represents the existence of a bounty (with range $\{0, 1\}$),* HUNT *represents whether locals hunt snakes (with range $\{0, 1\}$), and* BREED *represents whether locals breed snakes (with range $\{0, 1\}$).*

Consider the following two structural models.

$$M^\star : \begin{aligned} &\text{BOUNTY} = 0, \quad \text{HUNT} = \text{BOUNTY},\\ &\text{BREED} = \text{BOUNTY},\\ &\text{SNAKE} = 1 - \text{HUNT} + (2 \times \text{BREED}),\\ &\text{UTIL} = -(2 \times \text{SNAKE}) - (\text{BOUNTY} \times \text{HUNT}),\end{aligned}$$

and

$$M^\dagger : \begin{aligned} &\text{BOUNTY} = 0, \quad \text{HUNT} = \text{BOUNTY},\\ &\text{BREED} = 0, \quad \text{SNAKE} = 1 - \text{HUNT},\\ &\text{UTIL} = -(2 \times \text{SNAKE}) - (\text{BOUNTY} \times \text{HUNT}).\end{aligned}$$

Again, the true model is $M^\star$, but the Governor considers $M^\dagger$. In $M^\dagger$, the variable BREED *is considered irrelevant (i.e., the agent is unaware of or inattentive to that variable). As in Example 2, let μ be a cps such that $\bar{\mu}(M^\dagger) = 1$, and, for any E with $M^\star \in E$ and $M^\dagger \notin E$, $\mu_E(M^\star) = 1$. The Governor must choose between*

$$a_1 = (\text{BOUNTY} \leftarrow 1, (\text{SNAKE}, \text{HUNT})) \quad \textit{and}$$
$$a_0 = (\text{BOUNTY} \leftarrow 0, (\text{SNAKE}, \text{HUNT})),$$

with the associated predicted observations

$$o^1_{m^\star} = \langle \text{BOUNTY} \leftarrow 1, (\text{SNAKE} = 2, \text{HUNT} = 1)\rangle,$$
$$o^1_{m^\dagger} = \langle \text{BOUNTY} \leftarrow 1, (\text{SNAKE} = 0, \text{HUNT} = 1)\rangle,$$
$$o^0_{m^\star} = \langle \text{BOUNTY} \leftarrow 0, (\text{SNAKE} = 1, \text{HUNT} = 0)\rangle,$$
$$o^0_{m^\dagger} = \langle \text{BOUNTY} \leftarrow 0, (\text{SNAKE} = 1, \text{HUNT} = 0)\rangle.$$

As in Example 2, the observation generated by one action (here a_0) is model independent, whereas the observation generated by the other action (here a_1) depends on the model. In contrast to Example 2, however, the action chosen under the initial beliefs is the revealing action. To see this, note that, since the Governor places probability 1 on $M^\dagger$, he sees no value in exploration, and thus chooses his action based only on the immediate payoff. Since $\text{UTIL}(a_1, \mu) = -1$ *and* $\text{UTIL}(a_0, \mu) = -2$*, the Governor institutes the bounty, expecting the snake population to decline. However, the actual observation is $o = o^1_{m^\star}$: an increase in the snake population (and a realized utility of -5).*

This observation cannot be explained by his current model, so the Governor would revise his beliefs to $\mu_{E(o)}$. Under his new beliefs, his optimal action is to retract the bounty, which corresponds to steady-state behavior.

Notice that in this example, when the value of BREED is fixed to 0, the equations of $M^\star$ reduce to the equations of $M^\dagger$. That is, one could view $M^\dagger$ as a specification of $M^\star$ where a particular variable is ignored (i.e., is implicitly deemed unchangeable, is believed to bear no causal relation to any utility-relevant variable, or lies outside of the agent's awareness). Interpreting $M^\dagger$ as capturing the beliefs of an agent who is unaware of BREED and $M^\star$ as encoding her understanding after becoming more aware, then this equational relation between the two models embodies a principle akin to the principle of *Reverse Bayesianism* [Karni and Vierø, 2013] in state-space models of awareness. That is, when awareness expands, the previously-held relationships between variables remain as a special case of the new expanded relationships.

Example 3. *Let $\mathcal{U} = \{Y\}$ be a single variable with range $\mathcal{R}(Y) = \{0, 1, \ldots, 99\}$, and π the uniform distribution over outcomes. There are two endogenous variables, $\mathcal{V} = \{\text{UTIL}, X\}$, where X is the only variable that can be intervened on, and* UTIL *is the agent's utility with ranges $\mathcal{R}(X) = \{r, l, m\}$ and $\mathcal{R}(\text{UTIL}) = \{0, 1\}$. We consider three causal models:*

$$M^r : \quad X = m, \quad \text{UTIL} = \begin{cases} 1 \textit{ if } X = r \textit{ and } Y \geq 25\\ 1 \textit{ if } X = l \textit{ and } Y \geq 50\\ 0 \textit{ otherwise;}\end{cases}$$

$$M^l : \quad X = m, \quad \text{UTIL} = \begin{cases} 1 \textit{ if } X = r \textit{ and } Y \geq 50\\ 1 \textit{ if } X = l \textit{ and } Y \geq 25\\ 0 \textit{ otherwise;}\end{cases}$$

$$M^\star : \quad X = m, \quad \text{UTIL} = \begin{cases} 1 \textit{ if } Y \geq 50\\ 0 \textit{ otherwise.}\end{cases}$$

Suppose that there are two actions:

$$a_r = (X \leftarrow r, \text{UTIL}) \textit{ and } a_l = (X \leftarrow l, \text{UTIL}).$$

[6] Since there are no exogenous variables, we suppress the subscript for the context. Further, we abuse notation and write o^i rather than o^{a_i} for the action superscript.

[7] Note also that they place probability 0 on ever receiving any information, so the choice of action is determined entirely by the immediate value: $\text{VAL}(a_1, \mu) = \frac{9}{1-\delta}$ and $\text{VAL}(a_0, \mu) = 0$

Note the following simple observations: first, for beliefs sufficiently concentrated on M^r (resp., M^l), the action a_r (resp., a_l) maximizes expected utility; second, since the agent cannot observe Y, there are no inexplicable outcomes; all three models can produce any combination of X and Util. *As such, she will never completely discard any model that is in the support of her initial beliefs.*

Finally, assume the support of the agent's initial beliefs, $\bar{\mu}$, is $\{m^r, m^l\}$. In this case, her behavior will switch between the two actions infinitely often. To see this, assume to the contrary that her behavior was eventually constant. Without loss of generality, assume she eventually chooses a_r forever. Under the true model $M^\star$, this produces the outcome $(X \leftarrow r, \text{Util} = 1)$ with probability $\frac{1}{2}$ and the outcome $(X \leftarrow r, \text{Util} = 0)$ with probability $\frac{1}{2}$. This is the expected distribution of observations under M^l, but not *under M^r; thus, with π-probability 1, with sufficiently many observations, her beliefs will become arbitrarily concentrated on M^l; however, for such beliefs, a_r is no longer optimal.*

5 Introspective Unawareness

The agents in the framework laid out above entertain a classical exploration/exploitation trade-off: given their initial uncertainty about the true causal model, the forward looking agents maximizing $\text{Val}(\cdot, \mu)$ can take statically suboptimal actions to gather information about the causal structure, enabling better future decisions. The perceived value of exploration is determined by the agent's current beliefs, $\bar{\mu}$, and so she acts with certainty that the true model is contained in the support of $\bar{\mu}$. This is true even if the agent has repeatedly abandoned prior beliefs in the face of surprising observations; she never considers the possibility that her *current* beliefs might be misspecified or incomplete.

In this section, we augment the model above slightly to accommodate an introspectively unaware agent, who entertains the possibility that her beliefs are incomplete. If the agent was aware of which relevant models she was failing to consider, she could simply shift some small probability mass onto these models, expanding the support of her beliefs to capture all relevant possibilities. By way of contrast, we assume that the agent is (a) unaware of the models outside the support of $\bar{\mu}$, but (b) aware that she is unaware. This parallels awareness of unawareness in state-space models [Karni and Vierø, 2017]. The agent therefore cannot predict what it is that she might observe that would make her abandon her current world view. Nonetheless, the relevant parameters may still be identified from choice behavior [Karni and Vierø, 2025].

In essence, the idea is that the agent considers the possibility of encountering some inexplicable evidence, and conjectures about the potential likelihood and value of doing so, but cannot explicitly determine what the world would look like afterwards. To model this, we introduce two new objects into our model.

- Let v^* denote the agent's subjective value of learning something unforeseeable, capturing the agent's best guess of the (discounted expected) value of a paradigm shift that causes her to abandon her current beliefs.
- Let $\xi : A \to [0,1]$ denote a function that represents the agent's beliefs about the likelihood of each action yielding a surprising outcome, and let $\tau : [0,1]^A \times \mathcal{O} \to [0,1]^A$ describe how the agent updates her beliefs about the likelihood of surprise.

With these two pieces of notation, we can consider the introspective agent's version of Val:

$$\text{Val}^*(a, \mu, \xi) = \xi(a)v^* + (1 - \xi(a)) \sum_{\mathcal{M} \times \mathcal{R}(\mathcal{U})} \Big(\text{Util}(M^a, \vec{u}) + \delta \max_{a' \in A} \text{Val}^*(a', \mu^a_{M,\vec{u}}, \tau(\xi, o^a_{M^\star,\vec{u}})) \Big) \pi(\vec{u})\, \bar{\mu}(M).$$

The value of an action a is the $\xi(a)$ convex combination of v^*—the value subjectively assigned to unknown outcomes—and the expected value of the problem under the distribution $\bar{\mu}$. This second component is *almost* the same as Val as defined in Section 2.3, except that the continuation value also allows for the possibility of discovering a novel outcome, where the probabilities of discovery have been updated by τ.

Notice that the appeal of the novel outcome only kicks in when the known part of the model-space provides low expected payoff. In line with the finding of Bonawitz et al. [2014], actions that yield low "objective" payoff but high chance of revealing something novel will only be desirable when actions with higher objective payoff are unavailable.

We can therefore recast our notion of a steady state using v^*: Call a tuple (μ, a, ξ) an *introspective steady state* if

- $a \in \arg\max_{a' \in A} \text{Val}^*(a', \mu, \xi)$
- $\bar{\mu} = \bar{\mu}^a_{M^\star,\vec{u}}$ for $\vec{u} \in \mathcal{R}(\mathcal{U})$ with $\pi(\vec{u}) > 0$, and
- $\tau(\xi, o^a_{m^\star,\vec{u}}) = \xi$ for $\vec{u} \in \mathcal{R}(\mathcal{U})$ with $\pi(\vec{u}) > 0$.

The first two conditions are direct analogs of the conditions of a (non-introspective) steady state. The third also requires that the probabilities of discovery remain unchanged.

6 Conclusion

In Sections 2 and 3, there are no behavioral markers of off-path behavior; choices are dictated entirely by $\bar{\mu}$ (and its on-path evolution), as the decision maker places 0 probability on entertaining incorrect beliefs. Nonetheless, we do believe there are normative criteria for choosing one cps over another. For example, the agent may may prioritize structural stability (e.g., retaining the same causal graph when possible), or may want to move to beliefs that are 'close' to her old beliefs (e.g., changing as few structural equations as possible). We hope to formalize these ideas in future work.

Another direction we are pursuing is to characterize the exact boundary demarcating convergence to a steady-state. A main obstacle for convergence, as underscored by Example 3, is when an action a is optimal conditional on model M, but under the true model $M^\star$ taking a results in observations that make M less likely. Eliminating this yields convergence, but this is not necessary. Another interesting question regards *convergence in distribution*: the agent's behavior may not converge to a constant action but rather to a distribution over actions.

Acknowledgments

We would like to thank seminar audiences at the Philosophy, Politics, and Economics Society London Meeting and at Time, Uncertainties, and Strategies XI. Vierø gratefully acknowledges financial support from Aarhus University Research Foundation.

References

[Alexander and Gilboa, 2023] Yotam Alexander and Itzhak Gilboa. Subjective causality. *Revue économique*, 74(4):619–633, 2023.

[Angrist and Pischke, 2009] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

[Bonawitz *et al.*, 2014] Elizabeth Bonawitz, Stephanie Denison, Alison Gopnik, and Thomas L. Griffiths. Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, 74:35–65, 2014.

[Bramley *et al.*, 2017] Neil R. Bramley, Peter Dayan, Thomas L. Griffiths, and David A. Lagnado. Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, 124(3):301, 2017.

[Coenen *et al.*, 2019] Anna Coenen, Azzurra Ruggeri, Neil R. Bramley, and Todd M. Gureckis. Testing one or multiple: How beliefs about sparsity affect causal experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(11):1923, 2019.

[Cunningham, 2021] Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, 2021.

[de Finetti, 1936] Bruno de Finetti. Les probabilités nulles. *Bulletins des Science Mathématiques (première partie)*, 60:275–288, 1936.

[Eliaz *et al.*, 2020] Kfir Eliaz, Ran Spiegler, and Yair Weiss. Cheating with models. *American Economic Review: Insights*, 2020.

[Ellis and Thysen, 2021] Andrew Ellis and Heidi Christina Thysen. Subjective causality in choice, 2021. Available at http://arxiv.org/abs/2106.05957.

[Esponda and Pouzo, 2016] Ignacio Esponda and Demian Pouzo. Berk–Nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84(3):1093–1130, 2016.

[Eyster and Rabin, 2005] Erik Eyster and Matthew Rabin. Cursed equilibrium. *Econometrica*, 73(5):1623–1672, 2005.

[Fudenberg and Levine, 1993] Drew Fudenberg and David K. Levine. Self-confirming equilibrium. *Econometrica*, pages 523–545, 1993.

[Griffiths and Tenenbaum, 2005] Thomas L. Griffiths and Joshua B. Tenenbaum. Structure and strength in causal induction. *Cognitive Psychology*, 51(4):334–384, 2005.

[Halpern and Piermont, 2020] Joseph Y. Halpern and Evan Piermont. Dynamic partial awareness. In *Principles of Knowledge Representation and Reasoning: Proc. Seventeenth International Conference (KR '20)*, 2020.

[Halpern and Piermont, 2024] Joseph Y. Halpern and Evan Piermont. A representation theorem for causal decision making. In *Proc. 21st International Conference on Principles of Knowledge Representation and Reasoning (KR '24)*, 2024.

[Halpern, 2000] Joseph Y. Halpern. Axiomatizing causal reasoning. *Journal of A.I. Research*, 12:317–337, 2000.

[Hernán and Robins, 2020] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman and Hall/CRC, 2020.

[Karni and Vierø, 2013] Edi Karni and Marie-Louise Vierø. "Reverse Bayesianism": A choice-based theory of growing awareness. *American Economic Review*, 103(7):2790–2810, 2013.

[Karni and Vierø, 2017] Edi Karni and Marie-Louise Vierø. Awareness of unawareness: A theory of decision making in the face of ignorance. *Journal of Economic Theory*, 168:301–328, 2017.

[Karni and Vierø, 2025] Edi Karni and Marie-Louise Vierø. Measurements of attitudes toward unawareness. 2025.

[Kushnir and Gopnik, 2007] Tamar Kushnir and Alison Gopnik. Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology*, 43(1):186, 2007.

[Mailath and Samuelson, 2020] George J. Mailath and Larry Samuelson. Learning under diverse world views: model-based inference. *American Economic Review*, 110(5):1464–1501, 2020.

[Morgan and Winship, 2007] Stephen Morgan and Christopher Winship. *Counterfactuals and Causal Inference*. Cambridge University Press, 2007.

[Parascandola and Weed, 2001] Mark Parascandola and Douglas L. Weed. Causation in epidemiology. *Journal of Epidemiology & Community Health*, 55(12):905–912, 2001.

[Pearl, 2000] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[Pearl, 2009] Judea Pearl. Causal inference in statistics. *Statistics Surveys*, 3:96–146, 2009.

[Piermont and Zuazo-Garin, 2026] Evan Piermont and Peio Zuazo-Garin. Misspecified information in dynamic games, 2026. Available at https://arxiv.org/abs/2105.06772.

[Piermont, 2021] Evan Piermont. Unforeseen evidence. *Journal of Economic Theory*, 193, 2021.

[Popper, 1934] Karl R. Popper. *Logik der Forschung*. Julius Springer Verlag, Vienna, 1934.

[Popper, 1968] Karl R. Popper. *The Logic of Scientific Discovery*. Hutchison, London, 2nd edition, 1968. The first version of this book appeared as *Logik der Forschung*, 1934.

[Raina K. Plowright, 2008] Michael E. Gorman Peter Daszak Janet E. Foley Raina K. Plowright, Susanne H. Sokolow. Causal inference in disease ecology: investigating ecological drivers of disease emergence. *Frontiers in Ecology and the Environment*, 6(8):420–429, 2008.

[Schenone, 2020] Pablo Schenone. Causality: a decision theoretic foundation, 2020. Available at http://arxiv.org/abs/1812.07414.

[Spirtes *et al.*, 1993] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

[Stokey and Lucas Jr., 1989] Nancy L. Stokey and Robert E. Lucas Jr. *Recursive Methods in Economic Dynamics*. Harvard University Press, 1989.