Causal, Casual and Curious

Judea Pearl*

Physical and Metaphysical Counterfactuals: Evaluating Disjunctive Actions

https://doi.org/10.1515/jci-2017-0018

Abstract: The structural interpretation of counterfactuals as formulated in Balke and Pearl (1994a,b) [1, 2] excludes disjunctive conditionals, such as "had X been x_1 or x_2," as well as disjunctive actions such as $do(X = x_1 \text{ or } X = x_2)$. In contrast, the closest-world interpretation of counterfactuals (e.g. Lewis (1973a) [3]) assigns truth values to all counterfactual sentences, regardless of the logical form of the antecedent. This paper leverages "imaging" – a process of "mass-shifting" among possible worlds, to define disjunction in structural counterfactuals. We show that every imaging operation can be given an interpretation in terms of a stochastic policy in which agents choose actions with certain probabilities. This mapping, from the metaphysical to the physical, allows us to assess whether metaphysically-inspired extensions of interventional theories are warranted in a given decision making situation.

Keywords: counterfactuals, compound treatments, imaging, structural causal models

1 Introduction – Physical and metaphysical conceptions of actions

If the options available to an agent are specified in terms of their immediate consequences, as in "make him laugh," "paint the wall red," "raise taxes" or, in general, $do(X = x)$, then a rational agent is instructed to maximize the expected utility

$$EU(x) = \sum_y P_x(y)U(y) \tag{1}$$

over all options x. Here, $U(y)$ stands for the utility of outcome $Y = y$ and $P_x(y)$ – the focus of this paper - stands for the (subjective) probability that outcome $Y = y$ would prevail, had action $do(X = x)$ been performed and condition $X = x$ firmly established.

It has long been recognized that Bayesian conditionalization, i.e., $P_x(y) = P(y|x)$, is inappropriate for serving in eq. (1), for it leads to paradoxical results of several kinds (see [4, 5, pp. 108–109]). For example, patients would avoid going to the doctor to reduce the probability that one is seriously ill; barometers would be manipulated to reduce the likelihood of storms; doctors would recommend a drug to male and female patients, but not to patients with undisclosed gender, and so on. Yet the question of what function should substitute for $P_x(y)$, despite decades of thoughtful debates [6–8] seems to still baffle philosophers in the twenty-first century [9, 10].

Guided by ideas from structural econometrics [11–13], I have explored and axiomatized a conditioning operator called $do(x)$ [14] that captures the intent of $P_x(y)$ by simulating an intervention in a causal model of interdependent variables [15].

The idea is simple. To model an action $do(X = x)$ one performs a "mini-surgery" on the causal model, that is, a minimal change necessary for establishing the antecedent $X = x$, while leaving the rest of the model

***Corresponding author: Judea Pearl,** Computer Science Department, University of California, Los Angeles, Los Angeles, CA 90095-1596, USA, E-mail: judea@cs.ucla.edu

intact. This calls for removing the mechanism (i.e., equation) that nominally assigns values to variable X, and replacing it with a new equation, $X = x$, that enforces the intent of the specified action. This mini-surgery (which Lewis called "little miracle"), makes precise the idea of using a "minimal deviation from actuality" to define counterfactuals.

One important feature of this formulation is that $P(y|do(x))$ can be derived from pre-interventional probabilities provided one possesses a diagrammatic representation of the processes that govern variables in the domain [5, 16]. Specifically the post-intervention probability reads[1]:

$$P(x, y, z|do(X = x^*)) = \begin{cases} P(x, y, z)/P(x|z) & \text{if } x = x^* \\ 0 & \text{if } x \neq x^* \end{cases} \tag{2}$$

Here z stands for any realization of the set Z of "past" variables, y is any realization of the set Y of "future" variables, and "past" and "future" refer to the occurrence of the action event $X = x^*$.[2]

This feature is perhaps the key for the popularity of graphical methods in causal inference applications. It states that the effects of common policies and interventions can be predicted without knowledge of the functional relationships (or mechanisms) among X, Y, and Z. The pre-interventional probability and a few the qualitative features of the graph are sufficient for determining the pot-intervention probabilities as in eq. (2). This unfortunately applies only to "simple interventions" of the type $do(x)$, where x is a specific value of X, or a conjunction of such values. We lose this property when we wish to evaluate more intricate policies such as "exercise at least 30 minutes daily," or "paint the wall either green or purple."[3] To properly evaluate a disjunctive intervention such as $do(x_1 \text{ or } x_2)$ one would need to specify what conditional probability $P_{(x_1 or x_2)}(x|z)$ would prevail under the policy, substitute it in the structural model, and proceed to compute the overall post-intervention probability $P_{(x_1 or x_2)}(x, y, z)$. Theoretically, the post-intervention probability $P_{(x_1 or x_2)}(x|z)$ may be totally unrelated to any pre-intervention information in our possession. Therefore, re-specification can be a fairly demanding cognitive task, since the disjunction may select any subset of X's values. This motivates us to seek the guidance of another perspective for possible approximations and shortcuts.

The philosophical literature spawned a totally different perspective on the probability function $P_x(y)$ in eq. (1). In a famous letter to David Lewis, Robert Stalnaker [19] suggested to replace conditional probabilities with probabilities of conditionals, i.e., $P_x(y) = P(x > y)$), where $(x > y)$ stands for counterfactual conditional "Y would be y if X were x." Using a "closest worlds" semantics, Lewis [20] defined $P(x > y)$ using a probability-revision operation called "imaging," in which probability mass "shifts" from worlds to worlds, governed by a measure of "similarity". Whereas Bayes conditioning $P(y|x)$ transfers the entire probability mass from worlds excluded by $X = x$ to all remaining worlds, in proportion to the latters' prior probabilities $P(\cdot)$, imaging works differently; each excluded world w transfers its mass individually to a select set of worlds $S_x(w)$ that are considered "closest" to w among those satisfying $X = x$ (see Figure 1). Joyce [21] used the "\" symbol, as in $P(y\backslash x)$, to denote the probability resulting from such imaging process.

In Pearl [5, p. 73] I have shown that the transformation defined by the $do(x)$ operator, eq. (2), can be interpreted as an imaging-type mass-transfer, if the following two provisions are met.

Provision 1: the choice of "similarity" measure is not arbitrary; worlds with equal histories should be considered equally similar to any given world.

---

1 The relation between P_x and P takes a variety of equivalent forms, including the back-door formula, truncated factorization, adjustment for direct causes, or the inverse probability weighing shown in eq. (2) [5, pp. 72–73]. I chose the latter form, because it is the easiest to motivate without appealing to graphical notation. Numerical examples and computational features of inverse probability weighing are described in [17, pp. 72–75].

2 I will use "future" and "past" figuratively; "affected" and "unaffected" (by X) are more accurate technically (i.e., descendants and nondescendants of X, in graphical terminology). The derivation of eq. (2) requires that processes be organized recursively (avoiding feedback loops); more intricate formulas apply to non-recursive models. See Pearl [15, pp. 72–73] or Spirtes et al. [16] for a simple derivation of this and equivalent formulas.

3 Hernán and VanderWeele [18] called such disjunctions "Compound Treatments."

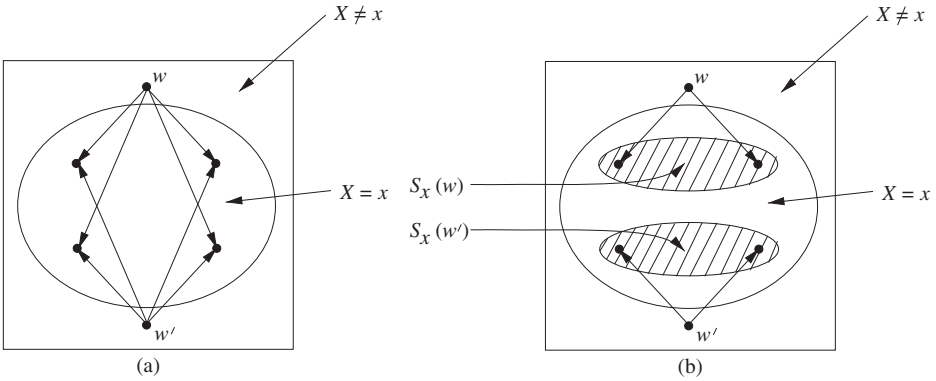

Figure 1: (a) Showing a world w (similarly w') excluded from the space $X = x$ transferring its mass $P(w)$ to all surviving worlds (in proportion to their prior probabilities). (b) Showing how an excluded world w transfers its mass $P(w)$ to a select set $S_x(w)$ of surviving worlds (still respecting their prior probabilities).

Provision 2: the re-distribution of weight within each selection set S_w) is not arbitrary either, equally-similar worlds should receive mass in proportion to their prior probabilities. This tie-breaking rule is similar in spirit to the Bayesian policy.[4]

Regardless of how we define "similarity," the Bayesian tie-breaking rule (Provision 2) permits us to write a general expression for the probability function $P(w\backslash x)$ that results from imaging on x. It reads:

$$P(w\backslash x) = \sum_{w'} P(w')P(w|S_x(w')) \tag{3}$$

This compact formula, adopted from Joyce [22], is applicable to any selection function $S_x(w)$ and gives the final weight $P(w\backslash x)$ of both excluded and preserved worlds w.[5] Accordingly, for any two propositions A and B we can write:

$$P(B\backslash A) = \sum_{w\in B} \sum_{w'} P(w')P(w|S_A(w')$$

In this paper, I will first describe how the post-interventional probability in eq. (2) emerges from the imaging probability in eq. (3) and then examine a wider class of imaging operations that give rise to eq. (2). Using Provisions 1 and 2, I will then use imaging to extend the application of the $do(x)$ operator to a wider set of suppositions, beyond those defined by the structural model. Finally, I will demonstrate that caution need be exercised when metaphysical extensions are taken literally, without careful guidance of decision making considerations.

2 Action as imaging

To see how eq. (2) emerges from the mass transfer policy of eq. (3), let us associate a "world" with a given instantiation of the three sets of variable $\{X, Y, Z\}$ where X stands for the action variable in $do(X = x)$, Y

4 Joyce [22] labeled this mass transfer policy, "*Bayesianized* imaging," and noted that it violates a tacit assumption made in Gardenfors's proof that imaging should preserve mixtures [23, pp. 108–113].

5 This follows from two observations:

$$S_x(w') = w' \text{ if } w' \Rightarrow X = x$$

since every world is "closest" to itself, and

$$P(w|S_x(w')) = 0 \text{ if } w \Rightarrow X \neq x$$

since any such w is not a member in any $S_x(w')$ set.

stands for variables that are potentially affected by the action (i.e., descendants of X in the causal graph), and Z stands for all other variables in the model. A *world* w then would be a tuple (x, y, z).

Prior to the action, each world is assigned the mass $P(x, y, z)$. After the action $do(X = x^*)$ is executed, this mass must be re-distributed, since worlds in which $X \neq x^*$ must be ruled out. The post-action weight of w, $P(x, y, z|do(X = x^*))$, is equal to the old weight plus a supplement $P(w' \to w)$ that w receives from some worlds w' whose mass vanishes. Aside from satisfying $X \neq x^*$, each such w' must consider w to be its "most similar neighbor," that is, w must be a member of $S_{x^*}(w')$.

According to Provision 1 above, worlds in the most-similar set $S_{x^*}(w')$ should share with w' the entire past $(Z = z)$ up to the point where the action occurs.[6] This means that the supplement weight $P(w' \to w)$ that w receives in the transition comes from each and every world $w' = (x', y', z')$ that shares with $w = (x, y, z)$ its z component. The total weight in those w' worlds is

$$\sum_{w'|x' \neq x^*, z'=z} P(w') = P(X \neq x^*, z). \tag{4}$$

However, w does not receive all the probabilities, $P(w')$ that w' is prepared to discharge, because w is in competition with other z-sharing worlds in the $X = x^*$ subspace (the space of surviving worlds). Since weight is distributed in proportion to the competitors prior weight, the fraction that w receives from each w' is therefore:

$$P(w)/\sum_{y} P(x^*, y, z') = P(x^*, y, z')/P(x^*, z') = P(y|x^*, z). \tag{5}$$

Multiplying eqs. (4) and (5), the total weight transferred to w from all its w' contributors is

$$\begin{aligned} P(y|x^*, z)P(X \neq x^*, z) &= P(y|x^*, z)P(z)(1 - P(x^*|z)) \\ &= P(y|x^*, z)P(z) - P(x^*, y, z). \end{aligned}$$

and adding to this w's original weight, $P(x^*, y, z)$, we finally obtain:

$$\begin{aligned} P(w\backslash x) &= P(y|x^*, z)P(z) \\ &= P(y, x^*, z)/P(x^*|z) \end{aligned} \tag{6}$$

which coincides with eq. (2).

To summarize, we have established the identity

$$P(w\backslash x) = P(w|do(x)) \tag{7}$$

which provides an imaging-grounded justification for the inverse-probability weighting that characterizes the $do(x)$ operator and, conversely, a decision-making justification for the imaging operator.

3 Imaging as an extrapolation principle

It is important to note that the fraction of weight that w receives from w', $P(y|x^*, z)$, is the same for any weight-contributing world w' that sees w as its closest neighbor, for they all share the same z. This means that

6 This follows from measuring similarity not by appearance, but rather by the number of mechanism modifications (or "miracles") necessary for establishing $X = x^*$ (see [5, p. 239]). Clearly worlds in which history differs from ours at several points in time require more modifications than one in which history remains in tact and only the last mechanism before the action is perturbed.

the same result would obtain had we not insisted that each w' individually delivers its weight to its closest neighbors, but allow instead for all the z'-sharing w's to first pool their weights together and then deliver that pooled weight onto the recipients in $\{w : x = x^*, z = z'\}$ in proportion to their prior weights $P(x^*, y, z)$. There is in fact no way of telling which way weight is being transferred in the transition, whether it is accomplished through an individual transfer, as depicted in Figure 2(a) or through "pooled" transfer, as in Figure 2(b).

The reason for this ambiguity lies in the coarseness of the propositions $X = x$ to which the $do(x)$ operator is applicable. In structural models, such propositions are limited to elementary instantiations of individual variables, and to conjunctions of such instantiations, but do not include disjunctions such as $do(X = x \text{ or } Y = y)$ or $do(X \neq x)$. The structural definition, which invokes equation removal, insists on having a unique solution for all variables,[7] before and after the intervention, and cannot allow therefore for ambiguity in the form of the disjunction $do(X = x \text{ or } X = x')$.

This limitation prevents us from excluding one single world $w' = (x', y', z')$ and watching how its weight is being distributed according to eq. (2). To do so would require us to compute the distribution $P(x, y, z|do(\neg x', \neg y', \neg z'))$, which is not definable by the surgery operation of structural models, since negations cannot be expressed as conjunctions of elementary propositions $X = x$, $Y = y$.

If we take imaging as an organizing principle, more fundamental than the structural account, we can easily circumvent this limitation and use eq. (3) together with Provision 1 to compute $P(B|do(A))$ for any

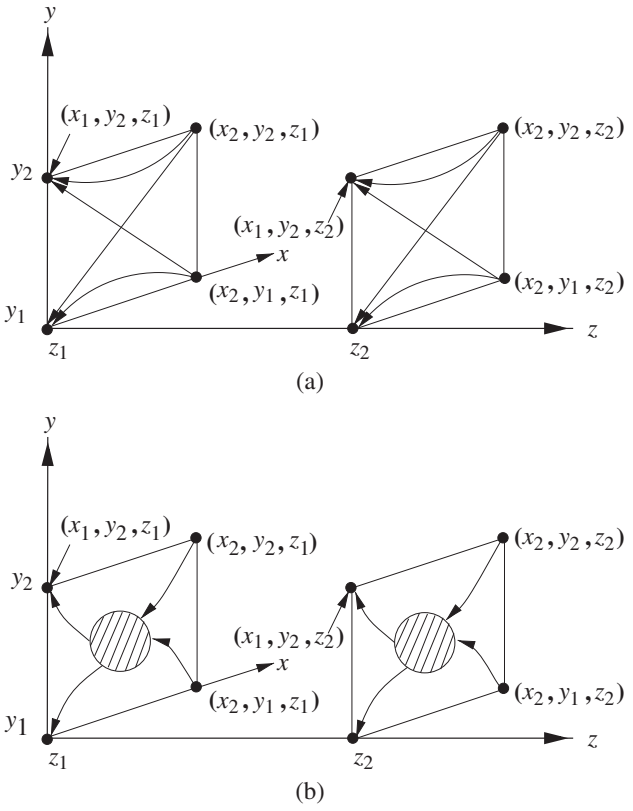

Figure 2: Imaging under the action $do(X = x_1)$. (a) using individual mass transfer and (b) using pooled mass transfer; the two are indistinguishable in the structural account.

7 See Halpern [24] for extensions to non-unique solutions.

arbitrary propositions A and B. This would give:

$$P(B|do(A)) = \sum_{w|w\in B} \sum_{w'} P(w')P(w|S_A(w')) \tag{8}$$

where $S_A(w')$ is the set of all A-worlds in for which $z = z'$.

Proponents of metaphysical principles would probably welcome the opportunity to overcome various limitations of the $do(x)$ operator and extend it with imaging-based extrapolations, beyond the decision making context for which it was motivated. In [5, Chapter 7], I indeed used such an extension to interpret counterfactuals with non-manipulable antecedents. For example, to define statements such as "She would have been hired had she not been a female," in which it is difficult to imagine a physical action $do(female)$, I proposed the symbolic removal of a fictitious equation $Gender = u_g$; the result is identical of course to eq. (3). Joyce [22] has also noted that imaging can answer problems on which the $do(x)$ operator is silent, and his example (Berkson paradox) falls well within the structural definition of non-manipulative counterfactuals (see [5, p. 206]).

Such extensions, which go from interventional to non-interventional counterfactuals are fairly safe, for the human mind interprets the two types of sentences through the same mental machinery. The interpolation proposed in eq. (8) however is much more powerful, for it assigns a concrete formal interpretation to disjunctive action $do(A \text{ or } B)$ for which no structural definition exists.[8]

In the next section I will demonstrate what this interpretation amounts to in the context of decision making. I will further argue though that such extensions should be taken with caution; limitations imposed by structural models are there for a reason – to keep us tuned to physical reality.

4 Imaging and disjunctive actions

Assume we are given three variables, $X = \{x_1, x_2, x_3\}$, $Y = \{y_1, y_2\}$, $Z = \{z_1, z_2\}$ such that X is affected by Z and Y is affected by both X and Z, as shown in Figure 3.

We wish to compute $P(y_1|do(x_2 \text{ or } x_3))$ from the prior probability $P(x, y, z)$, but since $do(x_2 \text{ or } x_3)$ is not defined, we resort to $P(y_1\backslash x_2 \text{ or } x_3)$ instead, as given in eq. (8).

Following the derivation in Section 3 we know that, in every $Z = z$ stratum, each of the four surviving worlds $\{(x, y, z) : x \in (x_2, x_3), y \in (y_1, y_2)\}$ receives a fraction $P(x, y, z)/P(X \neq x_1, Z = z)$ of the weight released by the two excluded worlds

$$\{(x', y', z') : x' = x_1, y' \in \{y_1, y_2\}, z' = z\}$$

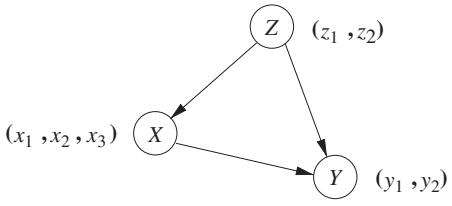

Figure 3: Given $P(x, y, z)$, find $P(y_1|do(x_2 \text{ or } x_3))$.

8 James Joyce called my attention to the fact that proponents of imaging may not see themselves committed to eq. (8) because this equation presupposes a determinate value for $S_{A \text{ or } B}(w')$ for each w'. Therefore eq. (8) may only be used in contexts where there is enough information to fix the $S_A(w')$, which would take a great deal of information to ascertain. In the sequel, I will assume that imaging analysts are committed to eq. (8).

The final mass in each surviving world will therefore be:

$$\begin{aligned} P(x, y, z \backslash x_2 \text{ or } x_3) &= P(x, y, z)/P((x_2 \text{ or } x_3)|z) \\ &= P(x, y, z)/[P(x_2|z) + P(x_3|z)] \end{aligned} \tag{9}$$

strongly reminiscent of the inverse probability formula of the standard $do(x)$ operator (eq. (2)), and amounts to Bayesian conditioning in each stratum of Z.

To compute our target quantity, we sum over z and obtain:

$$\begin{aligned} P(y_1 \backslash x_2 \text{ or } x_3) &= \sum_z P(y_1, z, x_2 \text{ or } x_3)/P(x_2 \text{ or } x_3|z) \\ &= \sum_z P(z)[P(y_1|z, x_2)P(x_2|z) + P(y_1|z, x_3)P(x_3|z)]/[P(x_2|z) + P(x_3|z)] \end{aligned} \tag{10}$$

This formula can be given a simple interpretation in terms of a stochastic intervention policy: An agent instructed to perform the action $do(x_2 \text{ or } x_3)$ first observes the value of Z, then chooses either action $do(x_2)$ or $do(x_3)$ with probability ratio $P(x_2|z) : P(x_3|z)$.

We see that the interpretation engendered by imaging reflects a commitment to specific interventional policy that may or may not be compatible with the intent of the action $do(x_2 \text{ or } x_3)$.

In the next section we will see that the silence of the structural theory vis a vis disjunctive actions is not a sign of weakness but rather a wise warning to potential ambiguity that deserves the attention of rational agents.

5 Restaurants and taxi drivers in the service of imaging

Consider the sentence:

> "The food was terrible, we should have asked the taxi driver to drop us at any of the other two restaurants in town."

Let the proposition $X = x_i$, $i = 1, 2, 3$ stand for "eating at the *ith* restaurant", and let $Y = y_1$ stand for "fine food." Assume that the quality of the various restaurants is encoded by the conditional probability $P(y|x)$, with $x \in (x_1, x_2, x_3)$, $y \in (y_1, y_2)$. Similarly, the likelihood that a freely-choosing taxi driver would drop us at any of the three restaurants in town is given by $P(x)$. We ask whether the imaging interpretation of $P(y|do(x_2 \text{ or } x_3))$ (eqs. (3), (7), or (10)) would provide an adequate evaluation of the sentence above.

The first question to ask is whether the information available is sufficient for calculating the probability of being dropped off at restaurant x_2 (similarly, x_3) were we to instruct the driver to avoid restaurant x_1. The structural theory says: no, and the imaging theory says: yes. The former argues that there is nothing in the information at hand that dictates how a taxi driver would behave once her space of options shrinks from three to two alternatives. The imaging theorist argues that it is highly unlikely (though possible) that a driver who prefers x_2 to x_3 when three options are available would change her preference under two options. Therefore, in the absence of information to the contrary, we should appeal to eq. (3) and, since all worlds are equally similar (i.e., sharing the same past, $Z = \{0\}$,) eq. (9) leads us to the Bayesian solution:

$$P(x_2|do(x_2 \text{ or } x_3)) = P(x_2|x_2 \text{ or } x_3) = P(x_2)/[P(x_2) + P(x_3)] \tag{11}$$

which preserves not only preferences but ratios as well.[9]

9 Note that both theories agree that, in principle, Bayesian conditionalization is inadequate for evaluating the probability sought in our story. The reason being that Bayesian conditionalization represents indicative conditionals (e.g., knowing that we were not dropped of at restaurant x_1, how likely is it that we will end up eating at x_2) while our conditional is subjunctive (e.g., if we were to

The structural theorist is not happy with this solution, and claims that, even if we assume "ratio preservation" under shrinking options, that does not guarantee proper evaluation of the hypothetical $P(y|do(x_2 \text{ or } x_3))$ because, even if ratios are preserved by every individual taxi driver, they may not be preserved in probability. To back up this claim he presents a numerical example, showing that, two different assumptions about drivers' behavior, both consistent with the information available, produce drastically different values for $P(x_2|do(x_2 \text{ or } x_3))$.

Suppose there are two types of taxi drivers in town. Type 1, designated $Z = z_1$, drop all customers at restaurant x_3 i.e.,

$$P(x|z_1) = \begin{cases} 1 & \text{if } x = x_3 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Type 2 drivers, designated $Z = z_2$, are not on the payroll of restaurant x_3, and follow the following pattern:

$$P(x|z_2) = \begin{cases} 8/9 & \text{if } x = x_1 \\ 1/9 & \text{if } x = x_2 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

10% of taxi drivers are of type 1 and 90% of type 2, thus $P(z_1) = 0.1 = 1 - P(z_2)$. Accordingly, the prior probabilities for X calculate to:

$$\begin{aligned} P(x) &= \sum_z P(x|z)P(z) \\ &= \begin{cases} 0.80 & \text{for } x = x_1 \\ 0.10 & \text{for } x = x_2 \\ 0.10 & \text{for } x = x_3 \end{cases} \end{aligned} \tag{14}$$

and the Bayesian answer gives equal probabilities to restaurants x_2 and x_3:

$$\begin{aligned} P(x_2|x_2 \text{ or } x_3) &= 0.50 \\ P(x_3|x_2 \text{ or } x_3) &= 0.50 \end{aligned} \tag{15}$$

On the other hand, if imaging is applied, eq. (10) gives a 9 to 1 preference to restaurant x_2:

$$\begin{aligned} P(x_2|do(x_2 \text{ or } x_3)) &= \sum_z P(z)P(x_2|z)/P(x_2 \text{ or } x_3|z) \\ &= \sum_z P(z)[P(x_2|z)/[P(x_2|z) + P(x_3|z)] \\ &= 0.90 \end{aligned} \tag{16}$$

while

$$P(x_3|do(x_2 \text{ or } x_3)) = 0.10$$

This proves, claims the structural analyst, that contrary to eq. (11), $P(x)$ tells us nothing about which restaurant we are likely to end up in if we allow each driver to follow her own preferences and average over

avoid x_1) or interventional (e.g., if we forbade the driver from x_1). The imaging analyst, however is willing to compromise, arguing that, if Bayesianized imaging works for definitive action, it should work for disjunctive action as well.

all drivers. What started with as equal prior probabilities eqs. (14) and (15) turns out to be a 9:1 preference, just by explicating the behavior of each driver. (Reversal of probability ratios is of course easy to demonstrate by assuming, for example $P(z_1) = 0.20$). This drastic change will be reflected in the expected food quality $P(y_1|do(x_2 \text{ or } x_3)$. Even if we assume that Y is independent of Z conditional on X, the fact that $P(x_2|do(x_2 \text{ or } x_3))$ depends so critically on the distribution of driver types, amounts to saying that the information available is insufficient for calculating the target quantity. Yet imaging commits to the Bayesian answer eq. (11) if we are not given $P(x|z)$. Such sensitivity does not occur in the calculation of non-disjunctive interventions like $P(y_1|do(x_2))$; regardless of the story about taxi drivers and their preferences, as long as we average over types, we get the same answer

$$P(y_1|do(x_2)) = P(y_1|x_2).$$

The imaging analyst replies that, while he appreciates the warning that the structural theory gives to decision makers, imaging is an epistemic theory and, as such, it views sensitivity to mechanisms as a virtue, not a weakness. If an agent truly believes the story about the two types of taxi drivers, it is only rational that the agent also believes in the 9:1 probability ratio. If, on the other hand, the agent has no basis for supposing this story over other conspiratorial theories, the Bayesian answer is the best one can expect.

There is nothing new in sensitivity to processes, argues the imaging analyst, it is commonplace even in the structural context. We know, for example, that the probability of causation (i.e., the probability that Y would be different had X been different [5]) is sensitive to the mechanism underlying the data. Yet the structural theory does not proclaim this probability "undefined." On the contrary, it considers it well-defined in fully specified models, and "unidentified" in a partially specified models, where some aspects of the underlying mechanisms are not known. In contrast, counterfactuals with disjunctive antecedents are deemed "undefined" even in fully specified structural models.

Here, the structural analyst replies that disjunctive counterfactuals are defined, albeit in the form of an interval, with

$$P(y|do(x_2 \text{ or } x_3)) \in [P(y|do(x_2)), P(y|do(x_3))]$$

and, if one insists on obtaining a definitive value for $P(y|do(x_2 \text{ or } x_3))$, a fully-specified model should take into account how each agent reacts to shrinking options – the Bayesian assumption that probability ratios should be preserved by default, at the population level, is utterly ad hoc.

The conversation would probably not end here, but the paper must.

6 Conclusions

The structural account of actions and counterfactuals provides a decision theoretic justification for two provisions associated with imaging operations: (i) worlds with equal histories should be deemed equally similar, and (ii) ties are broken in a Bayesian fashion. Extending these provisions beyond the context of decision making led to plausible interpretation of non-manipulative counterfactuals using either the structural or the possible-worlds accounts. However, extensions to disjunctive actions were shown to require assumptions that one may not be prepared to make in any given situation. This paper explicates some of these assumptions and helps clarify the relationships between the structural and imaging accounts of counterfactuals. Extensions to decision making in the context of linear models are developed in [25].

Acknowledgment: I am indebted to [the late] Horacio Arlo-Costa for bringing this topic to my attention and for his advice in writing this note. James M. Joyce has provided many insightful comments.

Funding: This research was partially supported by grants from NSF #IIS-1302448 and #1527490, ONR #N000-14-17-S-B001, and DARPA #W911NF-16-1-0579.

References

1. Balke A, Pearl J. Counterfactual probabilities: computational methods, bounds, and applications. In: de Mantaras RL, Poole D, editors. Uncertainty in artificial intelligence, vol. 10. San Mateo, CA: Morgan Kaufmann, 1994a: 46–54.
2. Balke A, Pearl J. Probabilistic evaluation of counterfactual queries. In: Proceedings of the twelfth national conference on artificial intelligence, vol. I. Menlo Park, CA: MIT Press, 1994b: 230–237.
3. Lewis, D. Counterfactuals. Cambridge, MA: Harvard University Press, 1973a.
4. Skyrms B. Causal necessity. New Haven: Yale University Press, 1980.
5. Pearl J. Causality: models, reasoning, and inference, 2nd ed. New York: Cambridge University Press, 2000[2009].
6. Cartwright N. How the laws of physics lie. Oxford: Clarendon Press, 1983.
7. Harper W, Stalnaker R, Pearce G. Ifs. Dordrecht: D. Reidel, 1981.
8. Jeffrey R. The logic of decisions. New York: McGraw-Hill, 1965.
9. Arlo-Costa, H. The logic of conditionals. In: Zalta EN, editor. The Stanford encyclopedia of philosophy, 2007. Available at: http://plato.stanford.edu/entries/logic-conditionals/.
10. Weirich, P. Causal decision theory. In: Zalta EN, editor. The Stanford encyclopedia of philosophy, 2008. Available at: http://plato.stanford.edu/archives/win2008/entries/decision-causal/.
11. Haavelmo, T. The statistical implications of a system of simultaneous equations. Econometrica 1943;11:1–12. Reprinted in Hendry DF, Morgan MS, editors. The foundations of econometric analysis. Cambridge University Press, 1995: 477–490.
12. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. New York: Springer-Verlag, 1993.
13. Strotz R, Wold H. Recursive versus nonrecursive systems: an attempt at synthesis. Econometrica 1960;28:417–427.
14. Pearl J. Causal diagrams for empirical research. Biometrika 1995;82:669–710.
15. Pearl J. Causality: models, reasoning, and inference, 2nd ed. New York: Cambridge University Press, 2009.
16. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search, 2nd ed. Cambridge, MA: MIT Press, 2001.
17. Pearl J, Glymour M, Jewell N. Causal inference in statistics: a primer. New York: Wiley, 2016.
18. Hernán M, VanderWeele T. Compound treatments and transportability of causal inference. Epidemiology 2011;22:368–377.
19. Stalnaker R. Letter to David Lewis. In: Harper WL, Stalnaker R, Pearce G, editors. Ifs. Dordrecht: D. Reidel, 1972[1981]: 151–152.
20. Lewis D. Counterfactuals and comparative possibility. In: Harper WL, Stalnaker R, Pearce G, editors. Ifs. Dordrecht: D. Reidel, 1973b[1981]: 57–85.
21. Joyce J. The foundations of causal decision theory. Cambridge, MA: Cambridge University Press, 1999.
22. Joyce J. Causal reasoning and backtracking. Philos Stud 2009;147:139–154, 2010 (print).
23. Gardenfors P. Causation and the dynamics of belief. In: Harper W, Skyrms B, editors. Causation in decision, belief change and statistics, vol. II. Kluwer Academic Publishers, 1988: 85–104.
24. Halpern J. Axiomatizing causal reasoning. J Artif Intell Res 2000;12:317–337.
25. Kuroki M. Counterfactual reasoning with disjunctive knowledge in a linear structural equation model. Tech rep. Department of Data Science, Institute of Statistical Mathematics, Japan. Submitted to *Journal of Causal Inference*. Available at: https://arxiv.org/abs/1707.09506/.