Annals for Mathematics and Artificial Intelligence, 61(1):29-39, 2011.
DOI: 10.1007/s10472-011-9247-9

TECHNICAL REPORT
R-360
June 2011

The algorithmization of counterfactuals

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

June 28, 2011

Abstract

Recent advances in causal reasoning have given rise to a computation model that emulates the process by which humans generate, evaluate and distinguish counterfactual sentences. Though compatible with the "possible worlds" account, this model enjoys the advantages of representational economy, algorithmic simplicity and conceptual clarity. Using this model, the paper demonstrates the processing of counterfactual sentences on a classical example due to Ernest Adam. It then gives a panoramic view of several applications where counterfactual reasoning has benefited problem areas in the empirical sciences.

Keywords: causal reasoning - counterfactuals - conditional logic

1 Introduction

One of the most striking phenomenon in the study of conditionals is the ease and uniformity with which people evaluate counterfactuals. To witness, the majority of people would accept the statement: S_1: "If Oswald didn't kill Kennedy, someone else did," but few, if any, would accept its subjunctive version: S_2: "If Oswald hadn't killed Kennedy, someone else would have."

For students of conditionals, these canonical examples (attributed to Ernst Adams, 1975) represent a compelling proof of the ubiquity of the indicative/subjunctive distinction, and of the amazing capacity of humans to process, evaluate and form consensus about counterfactuals.

Yet, not many students of conditionals asked the next question: How do we, humans, reach such consensus? More concretely, what mental representation permits such consensus to emerge from the little knowledge we have about Oswald, Kennedy and 1960's Texas, and what algorithms would need to be postulated to account for the swiftness, comfort and confidence with which such judgments are issued.

While it is generally acknowledged that reducing a theory to algorithmic details is helpful in maintaining clarity and facilitating communication among researchers, I submit that it serves a deeper purpose. Any theory of counterfactuals, be it of the possible worlds or "truth functional" variety should be deemed incomplete, until it is algorithmitized in sufficient details to allow a robot to correctly evaluate sentences on which humans agree. My contention rests on the observation that philosophers themselves rate the plausibility of theories by one and only one criterion: compatibility with human discourse. It seems to me, therefore, that a theory that cannot explain the computational realizability of its claims, has a much greater chance of deviating from its professed aim and, however appealing, cannot acquire the credence of an uncoached theory, running independent of its author-interpretor.

In Section 2 of this paper, I will present a formal model and simple algorithms that reliably interpret indicative and subjunctive conditionals, thus illustrating the basic elements of counterfactual reasoning. Section 3 will cast these algorithms in the context of the general theory of structural counterfactuals. In Section 4, I will demonstrate how this model has given rise to an effective methodology of causal inference in several of the empirical sciences, and how it has helped resolve practical questions, from policy evaluation and mediation analysis to generalizing conclusions across experimental studies. Finally, in the conclusion section, I will briefly compare the structural account of counterfactuals to the "possible worlds" account of Lewis (1973) and defend my preference of the former.

2 Oswald's Conditionals: Models and Algorithms

My basic thesis (Pearl, 2000) is that counterfactuals are generated and evaluated by symbolic operations on a model that represents an agent's beliefs about functional relationships in the world. The procedure can be viewed as a concrete implementation of Ramsey's idea (Ramsey, 1929), according to which a conditional is accepted if the consequent is true after we add the antecedent (hypothetically) to our stock of beliefs and make whatever minimal adjustments are required to maintain consistency (Arlo-Costa, 2007). In the indicative case, we simply add the antecedent A as if we received a new evidence that affirms its truth and discredits whatever previous evidence we had for its negation. In the subjunctive case, we establish the truth of A by changing the model itself.

Taking Kennedy's assassination as a working example, the model needed for evaluating the sentence S_1: "If Oswald didn't kill Kennedy, someone else did" is shown in the graph of Fig. 1. The symbols OS, SE, and KD represent the propositional variables "Oswald killed Kennedy," "Someone else killed Kennedy," and "Kennedy is dead," respectively, and the symbols M_{OS} and M_{SE} stand for the corresponding "motivations" (including all necessary enabling conditions) for each of the putative killers.[1] To complete the model, the arrows in the graph are annotated with the functions (double implication) that relate the corresponding variables to each other.

To interpret the indicative conditional S_1: 'If Oswald didn't kill Kennedy, someone else did" we start by assigning truth values to variables that are known (or believed) to be true in the story. In our case, although S_1 does not state so explicitly, the evaluation is

[1] The purpose of the M variables will become clear in the sequel.

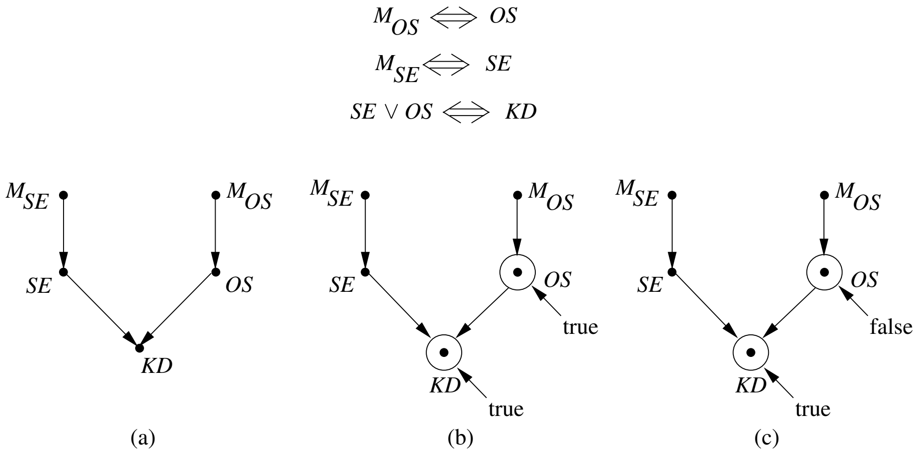

Figure 1: Evaluating an indicative conditional. State of knowledge (a) prior to learning that Oswald killed Kennedy, (b) after learning about Oswald's killing, and (c) after supposing that Oswald did not kill Kennedy.

predicated upon the common knowledge that Kennedy was in fact killed. Explicating that knowledge, S_1 can be written

$$S_1 : KD \ \wedge \ \neg OS \Rightarrow SE \tag{1}$$

In words, given that Kennedy is dead (KD) and that Oswald did not kill Kennedy ($\neg OS$) it must be that someone-else killed him (SE).

The truth value of S_1 can then be established by propagating truth values in the graphical model. Starting with the knowledge that KD and OS are true (Fig. 1(b)), we instantiate OS to its new truth value, *false*, and propagate these values to the rest of the variables in the theory (Fig. 1(c)), concluding with

$$\begin{aligned} SE &= \mathit{true}. \\ M_{SE} &= \mathit{true} \\ M_{OS} &= \mathit{false} \end{aligned}$$

We can also conduct a probabilistic analysis of S_1 by assigning probabilities to the root variables M_{SE} and M_{OS} and conclude, using Bayes formula, that

$$P(SE|\neg OS, K) = P(SE|\neg OS \wedge (SE \vee OS)) = P(SE|SE) = 1 \tag{2}$$

In words, regardless of the prior probabilities of the motivational variables M_{SE} and M_{OS}, S_1 is confirmed with probability 1.

The evaluation of the subjunctive conditional S_2 ("If Oswald hadn't killed Kennedy, someone else would have") triggers a different procedure. In addition to assuming that Oswald did in fact kill Kennedy, $KD \wedge OS$, S_2 calls for rolling back history as we know it, and rerun it under different conditions where, for some unknown reason, Oswald refrains from shooting Kennedy. This three-step procedure is illustrated in Fig. 2. Figure 2(a)

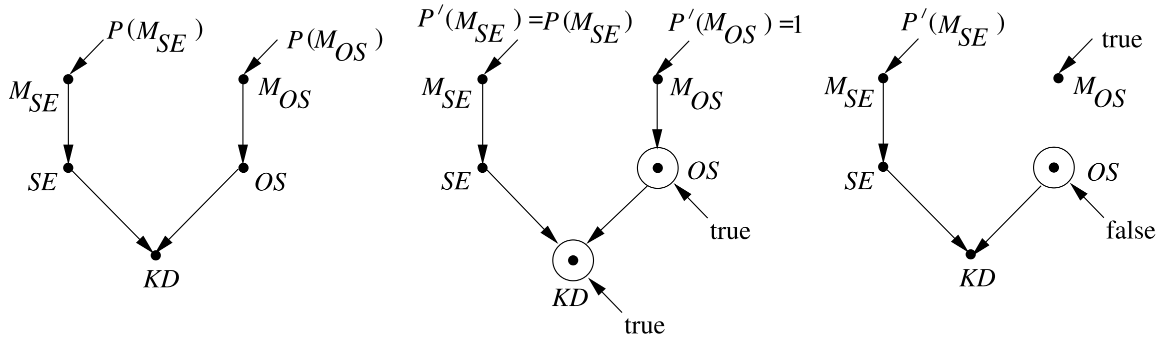

Figure 2: (a) Generic belief state, (b) Belief state after learning that Oswald killed Kennedy ($OS \wedge KD$), (c) Belief state assuming Oswald had refrained from killing Kennedy.

describes our generic belief state prior to learning that Oswald killed Kennedy. The root variables M_{SE} and M_{OS} are annotated with their prior probabilities $P(M_{SE})$ and $P(M_{OS})$. Upon learning that Oswald killed Kennedy (Fig. 2(b)) these probabilities are updated with the new evidence to yield:

$$P'(M_{SE}) = P(M_{SE}|KD, OS) = P(M_{SE})$$
$$P'(M_{OS}) = P(M_{OS}|KD, OS) = 1$$

Step 2 in the evaluation of S_2 calls for erasing the truth values of KD and OS, severing the link $M_{OS} \rightarrow OS$ and instantiating OS to *false* (Fig. 2(c)) to satisfy the antecedent of S_2. Finally, we need to compute the posterior probability $P'(S)$, based on the newly established priors, $P'(M_{OS})$ and $P'(M_{SE})$, and the newly established fact $OS = \mathit{false}$. This can readily be accomplished using the functional relationships in the model, yielding

$$P'(SE) = P(M_{SE}). \tag{3}$$

In other words, the probability that someone else would have killed Kennedy is the same as the probability that a random person would have kill Kennedy in 1963 Texas; our current knowledge about Kennedy assassination is totally irrelevant. The key difference between (2) and (3) lies in holding KD *true* in the former case but leaving it uncommitted in the latter.

This analysis is predicated on the assumption that there is no collusion between Oswald and another potential assassin (SE), nor any correlation in their behavior. Assuming however that Oswald and others are motivated by some public resentment, R, to President Kennedy's policies. In such a case (represented in Fig. 3) Kennedy's assassination as we know it lends evidence to the hypothesis that public resentment was a factor to reckon with or, at the very least, deserving a higher probability than what it garnered before the assassination. This is shown in Fig. 3(b), where the facts $OS = true$ and $KD = true$ are used to increase $P'(R)$ higher than $P(R)$. In the next phase of the evaluation, Fig. 3(c), we need to compute the updated probability $P'(KD)$ based on the fact $OS = false$ and the newly updated prior $P'(R) = P(R|M_{OS}) > P(R)$. Not surprisingly, any reasonable assumption about $P(M_{SE}|R)$ and $P(M_{OS}|R)$ would yield an increased probability for KD,

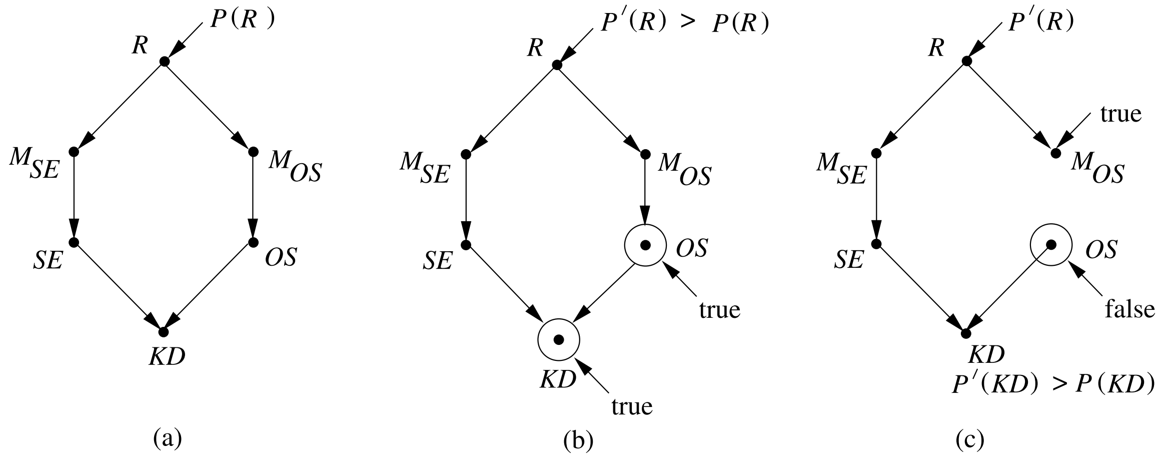

Figure 3: Belief states in the "public-resentment" theory. (a) Prior belief state, (b) Belief state after learning that Oswald killed Kennedy, showing increased probability of "Resentment," (c) Belief state assuming Oswald refrained from killing Kennedy; still, the probability that someone else would have killed him has increased, in view of what we know.

meaning that S_2 cannot be ruled out entirely. According to this "public resentment" theory, it is quite likely that, "had Oswald not killed Kennedy, someone else would have."

A speaker who seriously believes in S_2 is aiming to convey valuable information to the listener. For example, S_2 might convey the speaker's belief in the existence of public resentment to Kennedy prior to Kennedy's assassination. Or, the purpose of stating S_2 might be to convey the speaker's surprise at the intensity of that resentment, as revealed by the assassination. Whatever the aim of the speech act, it is clear that counterfactual statements, be they indicative or subjunctive, convey valuable information of either personal or factual nature.

In the next sections I will briefly describe how this theory of counterfactuals emerged in the empirical sciences and the role it played in resolving practical problems in planning and decision making.

3 An Outline of the Structural Theory

The analysis illustrated in the preceding section is part of a general theory of counterfactuals that I named "structural" (Pearl, 2000, Chapter 7) in honor of its origin in the structural equation models developed by econometricians in the 1940-50's (Haavelmo, 1943; Simon, 1953; Hurwicz, 1950; Marschak, 1953).

At the center of the theory lies a "structural model," M, consisting of two sets of variables, U and V, and a set F of functions that determine how values are assigned to each variable $V_i \in V$. Thus for example, the equation

$$v_i = f_i(v, u)$$

describes a physical process by which Nature *examines* the current values, v and u, of all variables in V and U and, accordingly, *assigns* variable V_i the value $v_i = f_i(v, u)$. The variables in U are considered "exogenous," namely, background conditions for which no explanatory mechanism is encoded in model M. Every instantiation $U = u$ of the exogenous variables uniquely determines the values of all variables in V and, hence, if we assign a probability $P(u)$ to U, it defines a probability function $P(v)$ on V.

The basic counterfactual entity in structural models is the sentence: "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$. The key to interpreting counterfactuals is to treat the subjunctive phrase "had X been x" as an instruction to make a "minimal" modification in the current model, so as to ensure the antecedent condition $X = x$. Such a minimal modification amounts to replacing the equation for X by a constant x, as we have done in Fig. 2(c). This replacement permits the constant x to differ from the actual value of X (namely $f_X(v, u)$) without rendering the system of equations inconsistent, thus allowing all variables, exogenous as well as endogenous, to serve as antecedents.

Letting M_x stand for a modified version of M, with the equation(s) of X replaced by $X = x$, the formal definition of the counterfactual $Y_x(u)$ reads:

$$Y_x(u) \stackrel{\Delta}{=} Y_{M_x}(u). \tag{4}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x. Galles and Pearl (1998) and Halpern (1998) have given a complete axiomatization of structural counterfactuals, embracing both recursive and non-recursive models. (see also Pearl, 2009b, Chapter 7).

Since the distribution $P(u)$ induces a well defined probability on the counterfactual event $Y_x = y$, it also defines a joint distribution on all Boolean combinations of such events, for instance '$Y_x = y$ AND $Z_{x'} = z$,' which may appear contradictory, if $x \neq x'$. For example, to answer retrospective questions, such as whether Y would be y_1 if X were x_1, given that in fact Y is y_0 and X is x_0, we need to compute the conditional probability $P(Y_{x_1} = y_1 | Y = y_0, X = x_0)$ which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model.

In general, the probability of the counterfactual sentence $P(Y_x = y|e)$, where e is any propositioned evidence, can be computed by the 3-step process (illustrated in Section 2);

Step 1 (***abduction***): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (***action***): Replace the equations corresponding to variables in set X by the equations $X = x$.

Step 3 (***prediction***): Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, Step 1 explains the past (U) in light of the current evidence e; Step 2 bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

It can be shown (Pearl, 2000, p. 76) that this procedure can be given an interpretation in terms of "imaging" (Lewis, 1973) – a process of "mass-shifting" among possible worlds – provided that (a) worlds with equal histories should be considered equally similar and

(b) equally-similar worlds should receive mass in proportion to their prior probabilities (Pearl, 2000, pp. 76). Because "similarities" are thus shaped by causal-temporal priorities, the structural account does not suffer from classical paradoxes that plague "similarity by appearance" (Taylor and Dennett, 2011). For example, the sentence "Had Nixon pressed the button, a nuclear war would have started" is accepted as true, despite Fine's (1975) "more similar" scenario in which someone had disconnected the switch. Fine's scenario is not minimally sufficient to ensure the antecedent "pressed the button."

4 Summary of Applications

Since its inception (Balke and Pearl, 1995) this counterfactual model has provided mathematical solutions to a vast number of lingering problems in policy analysis and retrospective reasoning. In the context of decision making, for example, a rational agent is instructed to maximize the expected utility

$$EU(x) = \sum_y P(Y_x = y)U(y) \tag{5}$$

over all options x. Here, $U(y)$ stands for the utility of outcome $Y = y$ and $P(Y_x = y)$ stands for the probability that outcome $Y = y$ would prevail, had action $do(X = x)$ been performed and condition $X = x$ firmly established.[2]

The central question in many of the empirical sciences is that of *identification*: Can we predict the effect of a contemplated action $do(X = x)$ or, in other words, can the post-intervention distribution, $P(Y_x = y)$, be estimated from data generated by the pre-intervention distribution, $P(z, x, y)$? Clearly, since the prospective counterfactual Y_x is generally not observed, the answer must depend on the agent's model M and then the question reduces to: Can $P(Y_x = y)$ be estimated from a combination of $P(z, x, y)$ and a graph G that encodes the structure of M.

This problem has been solved by deriving a precise characterization of what Skyrms (1980) called "KD-partition," namely, a set S of observed variables that permits $P(Y_x = y)$ to be written in terms of Bayes conditioning on, or, "adjusting for" S:

$$P(Y_x = y) = \sum_s P(y|x, s)P(s).$$

The solution came to be known as the back-door criterion (Pearl, 1995), stating (roughly) that a set S of variables is admissible for adjustment if it "blocks" every path between X and Y that ends with an arrow into X. In Fig. 3(a), for example, the effect of M_{SE} on KD can be predicted from pre-intervention data once we conditioned on R (or M_{OS}, or OS) because the latter "blocks" the back-door bath

$$M_{SE} \leftarrow R \rightarrow M_{OS} \rightarrow OS \rightarrow KD.$$

[2] Equation (5) represents the dictates of Causal Decision Theory (CDT) Stalnaker (1972); Lewis (1973); Gardenfors (1988) and Joyce (1999) – the pitfalls of Evidential Decision Theory are well documented (see (Skyrms, 1980; Pearl, 2000, pp. 108–9)), and need not be considered.

Tian and Pearl (2002) and Shpitser and Pearl (2007) further expanded this result and established a criterion that permits (or forbids) the assessment of $P(Y_x = y)$ by any method whatsoever.

Prospective counterfactual expressions of the type $P(Y_x = y)$ are concerned with predicting the average effect of hypothetical actions and policies and can, in principle, be assessed from experimental studies in which X is randomized. Retrospective counterfactuals, on the other hand, like S_2 in the Oswald scenario, consist of variables at different hypothetical worlds (different subscripts) and these may or may not be testable experimentally. In epidemiology, for example, the expression $P(Y_{x'} = y'|x, y)$ may stand for the fraction of patients who recovered (y) under treatment (x) that would not have recovered (y') had they not been treated (x'). This fraction cannot be assessed in experimental study, for the simple reason that we cannot re-test patients twice, with and without treatment. A different question is therefore posed: which counterfactuals can be tested, be it in experimental or observational studies. This question has been given a mathematical solution in (Shpitser and Pearl, 2007). It has been shown, for example, that in linear systems, $E(Y_x|e)$ is estimable from experimental studies whenever the prospective effect $E(Y_x)$ is estimable in such studies. Likewise, the counterfactual probability $P(Y_{x'}|x)$, also known as the effect of treatment on the treated (ETT) is estimable from observational studies whenever an admissible S exists for $P(Y_x = y)$ (Shpitser and Pearl, 2009).

Retrospective counterfactuals have also been indispensable in conceptualizing direct and indirect effects (Baron and Kenny, 1986; Robins and Greenland, 1992; Pearl, 2001), which require nested counterfactuals in their definitions. For example, to evaluate the direct effect of treatment $X = x'$ on individual u, un-mediated by a set Z of intermediate variables, we need to construct the nested counterfactual $Y_{x',Z_x(u)}$ where Y is the effect of interest, and $Z_x(u)$ stands for whatever values the intermediate variables Z would take had treatment not been given.[3] Likewise, the average *indirect effect*, of a transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z, hypothetically, to whatever value it would have attained had X been set to $X = x'$.

This counterfactual formulation has enabled researchers to derive conditions under which direct and indirect effects are estimable from empirical data (Pearl, 2001; Petersen et al., 2006) and to answer such questions as: "Can data prove an employer guilty of hiring discrimination?" or, phrased counterfactually, "what fraction of employees owes its hiring to sex discrimination?"

These tasks are performed using a general estimator, called the Mediation Formula (Pearl, 2001, 2009a, 2011), which is applicable to nonlinear models with discrete or continuous variables, and permits the evaluation of path-specific effects with minimal assumptions regarding the data-generating process.

Finally, as the last application, I point to a recent theory of "transportability" (Pearl and Bareinboim, 2011) which provides a formal solution to the century-old problem of "external validity" (Campbell and Stanley, 1966); i.e., under what conditions can experimental findings be transported to another environment, how the results should be

[3] Note that conditioning on the intermediate variables in Z would generally yield the wrong answer, due to unobserved "confounders" affecting both Z and Y. Moreover, in non linear systems, the value at which we hold Z constant will affect the result (Pearl, 2000, pp. 126-132).

calibrated to account for environmental differences, and what measurements need be taken in each of the two environments to license the transport.

The impact of the structural theory in the empirical sciences does not prove, of course, its merits as a cognitive theory of counterfactual reasoning. It proves nevertheless that in the arena of policy evaluation and decision making the theory is compatible with investigators states of belief and, whenever testable, its conclusions have withstood the test of fire.

5 Conclusions

In (Pearl, 2000, pp 239) I remarked: "In contrast with Lewis's theory, [structural] counterfactuals are not based on an abstract notion of similarity among hypothetical worlds; instead they rest directly on the mechanisms (or "laws," to be fancy) that govern those worlds and on the invariant properties of those mechanisms. Lewis's elusive "miracles" are replaced by principled mini-surgeries, $do(X = x)$, which represent a minimal change (to a model) necessary for establishing the antecedent $X = x$ (for all u). Thus, similarities and priorities—if they are ever needed—may be read into the $do(\cdot)$ operator as an afterthought (see (Pearl, 2000, Eq. (3.11)) and (Goldszmidt and Pearl, 1992)), but they are not basic to the analysis."

This paper started with the enigma of consensus: "What mental representation permits such consensus to emerge from the little knowledge we have about Oswald, Kennedy and 1960's Texas, and what algorithms would need to be postulated to account for the swiftness, comfort and confidence with which such judgments are issued." The very fact that people communicate with counterfactuals already suggests that they share a similarity measure, that this measure is encoded parsimoniously in the mind, and hence that it must be highly structured.

Using Oswald's counterfactuals as an example, this paper proposes a solution to the consensus enigma. It presents conceptually clear and parsimonious encoding of knowledge from which causes, counterfactuals, and probabilities of counterfactuals can be derived by effective algorithms. It carries therefore the potential of teaching robots to communicate in the language of counterfactuals and eventually acquire an understanding of notions such as responsibility and regret, pride and free will.

Acknowledgments

I am grateful to Daniel Dennett and Christopher Taylor for discussing the structural theory of counterfactuals and thus enticing me to write this exposition. This research was partially supported by grants from NIH #1R01 LM009961-01, NSF #IIS-0914211 and #IIS-1018922, and ONR #N000-14-09-1-0665 and #N00014-10-1-0933.

References

Adams, E. (1975). *The Logic of Conditionals*. D. Reidel, Dordrecht, Netherlands.

Arlo-Costa, H. (2007). The logic of conditionals. In *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.). URL = <http://plato.stanford.edu/entries/logic-conditionals/>.

Balke, A. and Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 11–18.

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51** 1173–1182.

Campbell, D. and Stanley, J. (1966). *Experimental and Quasi-Experimental Designs for Research.* R. McNally and Co., Chicago, IL.

Fine, K. (1975). Review of Lewis' counterfactuals. *Mind* **84** 451–458.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science* **3** 151–182.

Gardenfors, P. (1988). Causation and the dynamics of belief. In *Causation in Decision, Belief Change and Statistics II* (W. Harper and B. Skyrms, eds.). Kluwer Academic Publishers, 85–104.

Goldszmidt, M. and Pearl, J. (1992). Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning* (B. Nebel, C. Rich and W. Swartout, eds.). Morgan Kaufmann, San Mateo, CA, 661–672.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

Halpern, J. (1998). Axiomatizing causal reasoning. In *Uncertainty in Artificial Intelligence* (G. Cooper and S. Moral, eds.). Morgan Kaufmann, San Francisco, CA, 202–210. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

Hurwicz, L. (1950). Generalization of the concept of identification. In *Statistical Inference in Dynamic Economic Models* (T. Koopmans, ed.). Cowles Commission, Monograph 10, Wiley, New York, 245–257.

Joyce, J. (1999). *The Foundatins of Causal Decision Theory.* Cambridge University Press, Cambridge, MA.

Lewis, D. (1973). Counterfactuals and comparative possibility. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.). *Ifs*, D. Reidel, Dordrecht, pages 57–85, 1981.

Marschak, J. (1953). Economic measurements for policy and prediction. In *Studies in Econometric Method* (W. C. Hood and T. Koopmans, eds.). Cowles Commission Monograph 10, Wiley and Sons, Inc., 1–26.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York. Second ed., 2009.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146. <http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference.* 2nd ed. Cambridge University Press, New York.

Pearl, J. (2011). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. Tech. Rep. R-363, <http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf>, Department of Computer Science, University of California, Los Angeles. To appear in C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.), *Causality: Statistical Perspectives and Applications*. Forthcoming, 2011.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. Tech. Rep. R-372, <http://ftp.cs.ucla.edu/pub/stat_ser/r372-a.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, Proceedings of AAAI-2011.

Petersen, M., Sinisi, S. and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology* **17** 276–284.

Ramsey, F. (1929). General propositions and causality. In *Philosophical Papers* (F.P. Ramsey (Author) and H.A. Mellor, ed.). Cambridge University Press, Cambridge, 145–153.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Vancouver, BC, Canada, 352–359. Also, *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Shpitser, I. and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Montreal, Quebec.

Simon, H. (1953). Causal ordering and identifiability. In *Studies in Econometric Method* (W. C. Hood and T. Koopmans, eds.). Wiley and Sons, Inc., New York, NY, 49–74.

Skyrms, B. (1980). *Causal Necessity.* Yale University Press, New Haven.

Stalnaker, R. (1972). Letter to David Lewis. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.), *Ifs*, D. Reidel, Dordrecht, pages 151–152, 1981.

Taylor, C. and Dennett, D. (2011). Who's *still* afraid of determinism? Rethinking causes and possibilities. In *The Oxford Handbook of Free Will* (R. H. Kane, ed.). Oxford University Press, New York. Forthcoming.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, 567–573.