2 Turing Award Lecture

Transcript of Judea Pearl's Turing Award Lecture, *The Mechanization of Causal Inference: A 'Mini' Turing Test and Beyond*, presented at the 26th Association for the Advancement of Artificial Intelligence (AAAI) Conference, held in Toronto, Canada, in July 2012. The transcript has been lightly edited for clarity. The introduction is by Professor Kelly Gotlieb.

Kelly Gotlieb. It is my great pleasure to welcome you to the ACM Alan Turing Lecture. This annual presentation is delivered by the winner of the ACM Alan Turing Award, which is named for the great British mathematician and computer scientist Alan M. Turing, the originator of the Turing Test, and whose 100th birthday we've been celebrating.

The Turing Award is often referred to as the "Nobel Prize of Computing," and is the most prestigious prize a computer scientist can receive; it carries a $250,000 prize generously provided by Intel and Google.

This year's recipient of the ACM Turing Award, and our lecturer this morning, is Judea Pearl, Professor of Computer Science and Statistics at the University of California in Los Angeles. He received this honor in recognition of his fundamental contribution to artificial intelligence as a result of the development of a calculus for probabilistic and causal reasoning.

So, you can see it is quite fitting that he addresses this audience at this conference, seeing he is one of the true pioneers in advancing both the science and the art of artificial intelligence. And I do not give the term "art" loosely because if you know any of Professor Pearl's works or books, you'll know that he is as much a philosopher as a scientist.

The subject for his talk this morning is "The Mechanization of Causal Inference: A 'Mini' Turing Test and Beyond." It is my privilege to introduce Judea Pearl.

Edited transcript of Lecture Video at https://amturing.acm.org/award_winners/pearl_2658896.cfm.

Judea Pearl. Thank you, Kelly, for a wonderful introduction. I'm very glad to be here. I did request to deliver the Turing lecture at AAAI because you, AAAI students and researchers, were with me at an early stage of this game, and deserve to hear a progress report about what happened in this adventure since we last played in the sandbox and built those castles together.

Also, I think it is important that I pay tribute to AAAI for nurturing my work when it was not exactly fashionable. I want to thank all of you for being partners in the development of the things I'm going to talk about: colleagues, co-authors, co-principal investigators, students, and reviewers. I do not know if I should thank my reviewers as well [LAUGHTER].

Three of my most important works were published in the proceedings of AAAI, so I would like to start with those.[1] The first, presented at AAAI 1982 in Pittsburgh, was my first paper on belief propagation in trees. The second was presented at AAAI 1994 and it was a paper with Adnan Darwiche on the *do calculus*. I'm sure that it wouldn't have been published in any other conference proceedings, in Statistics or any other field. The third was presented in the same conference, AAAI 1994, and it was the paper with Alexander Balke on "Probabilistic Evaluation of Counterfactual Queries." I chose those three papers because their titles are closely related to the names of the three-layer hierarchy of causal reasoning that we have today. They established a very solid kind of hierarchy that is rarely mixed, in the sense that you can syntactically tell if a sentence is probabilistic, causal, or counterfactual.

But this is not a lecture about my work; it is a lecture about Turing. So, let me start with Turing and his Turing Test in the article in *Mind Magazine* in 1950: a test that I think is an engine behind much of the work that is done in AI.

Turing's answer to the question of, "Can computers think?" was very simple. "Yes, if it acts like it thinks," where "acting" means that it provides reasonable answers to non-trivial questions about a story, a topic, or a situation. Many of us are working on mini-Turing Tests in various fields. I will consider questions that involve causal inference.

Here is how Turing described a hypothetical conversation with the machine. First was the question about poetry. And the answer, of course, is evasive, although with some human element to it: "I never could write poetry."

The second question is about arithmetic: "Can you add that and that," and the answer is also human. You pause for 30 seconds, and then you give the answer. This is also a very simple domain.

[1] The three AAAI papers that Judea Pearl is referring to are: "Reverend Bayes on inference engines: A distributed hierarchial approach" [Pearl 1982]; "Symbolic causal networks for reasoning about action and plans" [Darwiche and Pearl 1994]; and "Probabilistic evaluation of counterfactual queries" [Balke and Pearl 1994].

And then Turing said, “Let’s look at chess. Do you play chess?” “Yes.” “I have a King on my K1, and no other pieces; you have only King at K6 and Rook at R1. It is your move. What do you play?” And of course the machine answers after a pause, “Checkmate.”

So, these were the questions exemplified in Turing’s first paper: questions about various domains like arithmetic, poetry, and chess, all of which admit reasonable answers.

But then Turing talks about a “child machine,” which is essentially machine learning. “Why don’t we start with a child machine?” It should be easier, he said, because the child does not need as much background as we expect adults to have. “Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed.” I think that Turing underestimated the role that vision and motor action play even in high level intelligence. We know, for example, that metaphors taken from the child world play a tremendous role in the child’s ability to handle mathematics.

Turing then made some statements about the connection between machine learning and evolution, and said: “The survival of the fittest is a slow method for learning. The experimenter [the programmer], by exercise of intelligence, should be able to speed it up. How? By creating artificial mutations where they are needed. If he can trace the cause of some weakness, he can then probably think of a kind of mutation which will improve it.” Turing’s idea was that the programmer would be able to trace shortcomings of the program to where they matter, and fix them. There was a great vision here because it leads to the question: Why shouldn’t a machine, having a blue print of itself be able to pinpoint the root causes of a weakness, and change priority among competing computational resources?

I will explain to you why I chose causal reasoning to be a domain that deserves to be called a “Mini Turing Test.” For this, imagine that you have Turing’s experimental setting with an interrogator asking a machine questions. The questions, however, are limited mainly to three types or modalities: *What is?*, *What if?*, and *Why?*

The story, that I used many times in my 1988 book *Probabilistic Reasoning* [Pearl 1988] and the 2000 book *Causality* [Pearl 2000], is as follows: You get out of your house and you see the pavement. The pavement may be wet or dry, it may also be slippery or not, it may have rained or not, the season may be dry or wet, and the sprinkler may have been on or off. These are five binary variables that can be used to generate many simple stories connected to your everyday experience. The task is to tell a story to the machine and the machine has to answer questions corresponding to the three modalities.

One simple question: if the season is dry and the pavement is slippery, did it rain? You expect an answer like: "It is unlikely. It is more likely that the sprinkler was on, with a very slight possibility that the pavement is not even wet." There could indeed be other reasons for why the pavement is slippery. This is the kind of answer that you expect on the basis of observations alone.

Then comes a second question: "What if you see that the sprinkler is off?" A plausible answer is: "It is more likely then that it rained." This is reasonable; it is an example of what is called "explaining away."

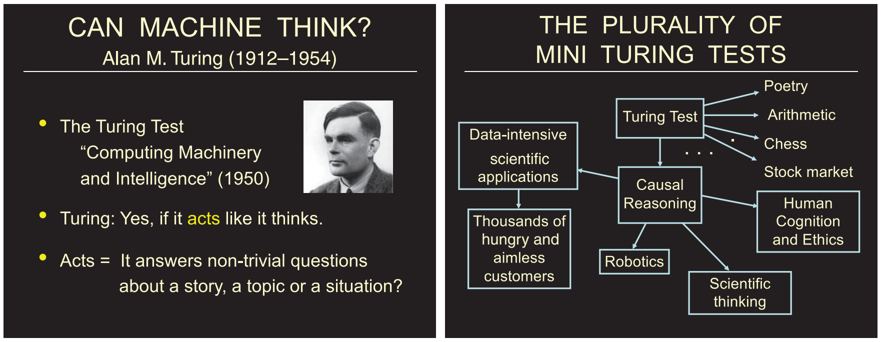
Now a question about actions: "Do you mean that if we actually turn the sprinkler on, then rain will be less likely?" And you want the machine to say, "No, there is a difference between seeing and doing; the likelihood of rain would remain the same but the pavement will surely get wet."

Finally, a question of counterfactual nature: "Suppose that you see that the sprinkler is on and the pavement is wet. Would the pavement be wet if the sprinkler were off?" I'll explain why I'm so hung up on counterfactuals, but first I would like you to answer the question instead of the machine. What I expect the machine to say is, "The pavement would be dry then, because the season is likely dry." Namely, you take the observation here, that the sprinkler is on, and you infer, "Oh, it must be a dry season." Then, if the sprinkler were off, the past remains the same but the future changes, so the justification should be: "Because the season is likely dry and the pavement is wet."

This is the kind of question/answer session that we expect for a toy problem. We all remember, however, Searle's argument of the Chinese room that says that answering questions does not mean that a machine thinks or even understands the questions. To prove his point, Searle imagines that the machine takes the questions in Chinese and answers them using a rule book, where every sentence in Chinese has the answer printed there in Chinese or in English. He concludes that the machine can't be said to understand Chinese just because it looks up the answers in the book.

What Searle overlooks is the fact that there are not enough molecules in the universe to make up such a book, because of the huge number of questions that may be asked. "So what?," you may ask. "Just because you have combinatorial difficulty, you conclude that the machine thinks?" [AUDIENCE LAUGHS] The answer is "Yes," because when you have such a combinatorial problem to overcome, the only way to solve it is by taking advantage of the relevant constraints in the domain. And understanding and taking advantage of the relevant principles and constraints is what we mean by understanding.

Even for the sprinkler example, if, for the sake of argument, we consider ten binary variables and count the number of entries in the table that we would need



CAN MACHINE THINK?

Alan M. Turing (1912–1954)

- The Turing Test
  "Computing Machinery and Intelligence" (1950)
- Turing: Yes, if it acts like it thinks.
- Acts = It answers non-trivial questions about a story, a topic or a situation?

THE PLURALITY OF MINI TURING TESTS

Turing Test

Poetry

Arithmetic

Chess

Stock market

. . .

Data-intensive scientific applications

Causal Reasoning

Thousands of hungry and aimless customers

Human Cognition and Ethics

Robotics

Scientific thinking

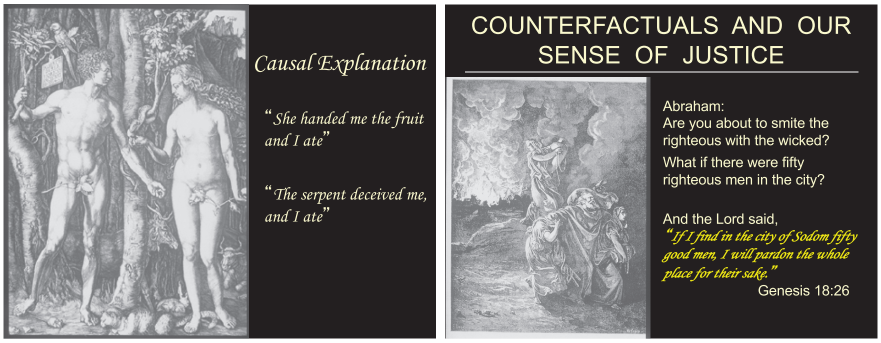
Figure 2.1 Turing Test and a plurality of mini-Turing Tests.

to use Searle's Chinese-book method, it turns out that we would need on the order of 1,000 entries just for the probability. We need to multiply this number by another 1,000 to get the probablities for all actions, and by an additional 1,000 to account for the counterfactual queries. So, we would need a billion-long table just to answer questions about the simple pavement story.

Yet even children can answer these questions quite intelligibly, and the question is, "How?" I'll argue that there are important principles and constraints that enable the child to answer questions about observations, interventions, and counterfactuals, but before getting there, I'd like to explain why I think that the causal conversation is important (Figure 2.1, "The plurality of mini Turing tests").

Causal reasoning is important because it is pervasive in human cognition and human ethics, and it is deeply entrenched in the cognitive development of children. In addition, causality is a building block of scientific thinking and crucial in robotics. Finally, and that's the reason I have spent more than 25 years of my life on causality, there are many data-intensive applications that can benefit from any new insight in causal reasoning. There are thousands of hungry and aimless customers, not hungry for money since they are well endowed—all the pharmacutical companies are part of this enterprise—but they are hungry for ideas, because causal reasoning has not been properly formalized in those fields. Thus, any insight that we get by trying to make a robot understand cause-and-effect could translate into methods that could save millions of lives and dollars in those fields.

Let me start with human cognition and ethics (Figure 2.2). I like to start with Adam and Eve—where else do you start? And you can see immediately that when God asked Adam, "Hey, did you eat from that tree?" Adam does not answer "yes" or "no." He says instead, "She handed me the fruit and I ate." You see: facts are for the gods; excuses are for men. [LAUGHTER] And Eve, of course, is no less expert in causal explanations, and says, "Don't blame me. The serpent deceived



Causal Explanation

"She handed me the fruit and I ate"

"The serpent deceived me, and I ate"

COUNTERFACTUALS AND OUR SENSE OF JUSTICE

Abraham:
Are you about to smite the righteous with the wicked?
What if there were fifty righteous men in the city?

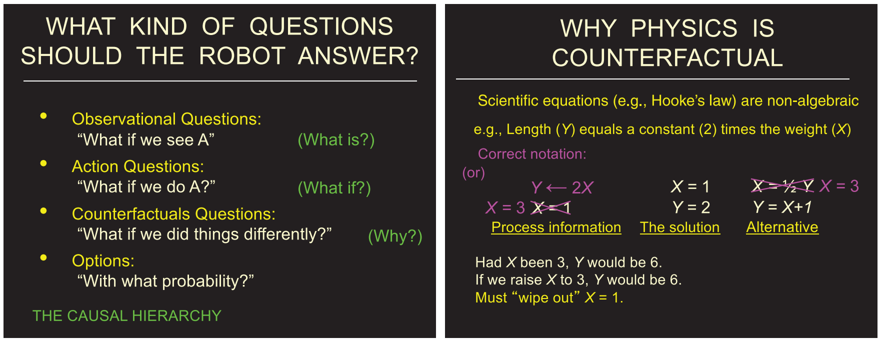
And the Lord said,
"If I find in the city of Sodom fifty good men, I will pardon the whole place for their sake."
Genesis 18:26

Figure 2.2 Causes, counterfactuals, and our sense of justice.

me and I ate." Thus causal reasoning plays a key role in our sense of justice, and in the need to pass the buck to somebody else. [LAUGHTER]

You also remember when God told Abraham that he is about to destroy the cities of Sodom and Gomorrah, and Abraham said, "Are you about to smite the righteous with the wicked? You can't do that. What if there were 50 righteous men in the city?" Here, you have the first counterfactual in the Bible. [AUDIENCE LAUGHS] "What if there were 50?" And look what God says: "If I find in the city of Sodom 50 good men, I will pardon the whole place for their sake." Do you think that Abraham gave up at that point? No. He got down and said, "What about 45?" [AUDIENCE LAUGHS] "Are you going to make a big fuss for five people?" And God says, "No, I ain't gonna destroy it," and then he goes down to 40, and then 30, and 20, and 10, and you know what happened. The rest is history, and the question, of course, is what kind of game this is. Did Abraham doubt the ability of God to count or to distinguish the righteous from the wicked? No. Abraham was the first scientist: he tried to find a general rule. "Where is the threshold?" "What is the general rule for collective punishment?" [AUDIENCE LAUGHS] In that sense, he was the first scientist, because what is science all about? It is about the general rules; not about specific events.

So, here I go to science to prove to you that counterfactuals are indeed the basis for science. We all used to do problems in physics, for example, using Hooke's law, which tells you that the length of the string Y is equal to a constant, say 2, times the weight X it supports. So if X is one kilogram, we have two equations: $Y = 2X$ and $X = 1$ (Figure 2.3). You may think that finding the length of the string Y is just arithmetic: you solve the two equations with the two unknowns, and obtain the values $Y = 2$ and $X = 1$. The question is: are the equations $Y = 2X$ and $X = 1$, and the equations $Y = 2$ and $X = 1$ equivalent? They are of course algebraically equivalent, as they have the same solution, but I will argue that they are not equivalent,



WHAT KIND OF QUESTIONS SHOULD THE ROBOT ANSWER?

- Observational Questions:
  "What if we see A" (What is?)
- Action Questions:
  "What if we do A?" (What if?)
- Counterfactuals Questions:
  "What if we did things differently?" (Why?)
- Options:
  "With what probability?"

THE CAUSAL HIERARCHY

WHY PHYSICS IS COUNTERFACTUAL

Scientific equations (e.g., Hooke's law) are non-algebraic

e.g., Length (Y) equals a constant (2) times the weight (X)

Correct notation:
(or)

$Y \leftarrow 2X$ $X = 1$ ~~$X = ½ Y$~~ $X = 3$
$X = 3$ ~~$X = 1$~~ $Y = 2$ $Y = X+1$
Process information The solution Alternative

Had X been 3, Y would be 6.
If we raise X to 3, Y would be 6.
Must "wipe out" $X = 1$.

Figure 2.3 The causal hierarchy and why physics is counterfactual.

because the equations on the left can answer questions that the ones on the right cannot.

For illustrating this difference, consider actually a system of equations $X = Y/2$ and $Y = X + 1$ which has the same solution $Y = 2$ and $X = 1$, along with the following question: "If we raise the weight X to 3, what would be the length Y?" In the first system of equations $Y = 2X$ and $X = 1$, which captures Hooke's law and the unit body weight, the counterfactual question "if X had been 3" has the answer $Y = 6$, which can be obtained by wiping out the equation $X = 1$ and replacing it by $X = 3$. The new system of two equations, modified by the new information, gives us the answer $Y = 6$.

The system of equations $X = Y/2$ and $Y = X + 1$, on the other hand, has the same solutions as the equations $Y = 2X$ and $X = 1$, but if we apply the same method for answering the counterfactual query, and replace the equation $X = Y/2$ by $X = 3$, we obtain the answer $Y = 4$, which is wrong.

Every child in high school, when he or she solves physics problems, engages in counterfactual reasoning of this sort. The child knows which equations to write, which equations to wipe out, and which ones to keep. They keep the one that conveys the generic rule and wipe out the ones that are merely boundary conditions and subject to the antecedent of the counterfactual. If this is the case, the equality sign that we saw before in the equation $Y = 2X$ for expressing Hooke's law does not really represent an algebraic equality but something closer to an assignment statement in a programming language.

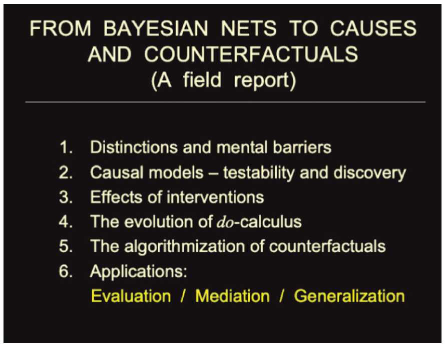
You can imagine that Nature, before determining the length of the spring, looks around for all variables that might possibly affect the length. She looks at the weight and says, "Ah, that is the one," then consults the weight on the spring, and finally determines the value of the length. So, this is the conception of Nature in physics: Nature looks at some variables, goes through some process, and then assigns values to other variables. If that is so, then modeling Nature requires a different kind of algebra because the process involves wiping out equations. That is the meaning of arrows in the structure of causal graphs; it is a description of the strategy used by Nature.

The role of counterfactuals and causation in human reasoning has not escaped philosophers. Already at the time of the Greeks in 430 BC, Democritus said, "I would rather discover one causal relationship than be king of Persia." King of Persia at that time was not exactly a dangerous occupation like it is today. [LAUGHTER] And Hume, of course, looked at that and said, "What is this idea of causation? I've got to solve it." And he came out with the conception that causation is not a gift of the gods, but something that we learn from experience. Here is a famous paragraph: "We remember to have seen that species of object called 'flame,' and to have felt the species of sensations we call 'heat.' Without any further ceremony, we call the one 'cause' and the other 'effect.'" So, it is a matter of determining regularity in nature that makes us come up with the label "cause." There are obvious difficulties to that conception, of course, but the fact that generations of philosophers have stumbled on the difficulty of explaining what "cause" is, brings us to ask: "What gives us the audacity, here in AI, to think that we can add another iota to this long debate?"

The answer is simply that we do not have the luxury to philosophize. We need to build robots that understand what went wrong in the laboratory or the kitchen, and if they do not learn it by themselves, we need to teach them, so that they can act properly and answer queries about cause/effect relationships. And this is not a trivial thing to do because now the puzzles that philosophers have faced translate into engineering problems. The question of, "How do we acquire causal information from the environment?" is translated into, "How do we people conclude that the sprinkler caused the pavement to get wet?" And the question of "How do we people conclude that the sprinkler caused the pavement to get wet?" translates into, "How should a robot use causal information received from its creator-programmer to understand or to answer queries properly?"

The use of causal information may look trivial but it is not, because if you just follow the rules you get unexpected results. If the input is "If the grass is wet, then it rained" and "If you break this bottle, the grass will get wet," you do not want an output such as "If we break the bottle, then it rained." So, just rule-chaining is not going to do the work for us; we need something more.

And what is that something more? Before we get there, let me provide an outline of what I'm going to talk about (Figure 2.4). I'm going to talk about the three-level hierarchy first. The question "What if I see" is about probability and beliefs. The question "What if I do?" is about actions and interventions. Finally, the question



FROM BAYESIAN NETS TO CAUSES
AND COUNTERFACTUALS
(A field report)

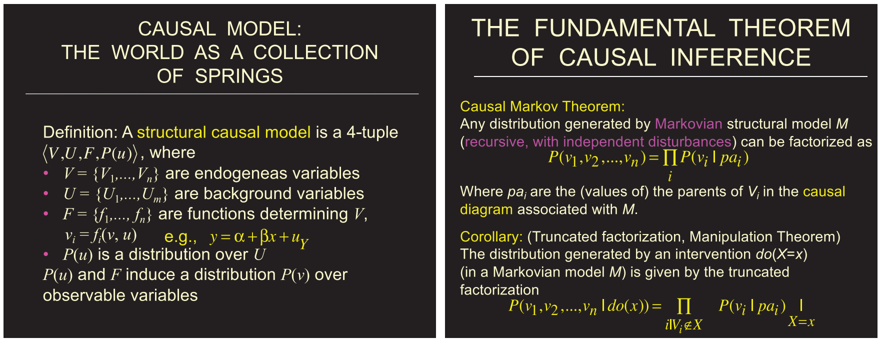
1. Distinctions and mental barriers
2. Causal models – testability and discovery
3. Effects of interventions
4. The evolution of *do*-calculus
5. The algorithmization of counterfactuals
6. Applications:
   Evaluation / Mediation / Generalization

Figure 2.4 Roadmap: From Bayesian nets to causality and counterfactuals.

"What if I did things differently?" is about counterfactuals. You can decorate these questions with probabilities; namely, how likely are the answers, but that's not essential.

The following is a field report of the journey that we took from the old days of Bayesian networks to causality and counterfactuals. We have to understand the distinctions and mental barriers that stood in our way. We have to talk also about what makes a model causal as opposed to something else, how a causal model can be tested, and how causal models and data are connected. If a model has testable implications, then you can hope to discover or learn the model from data. A model that does not have any testable implication cannot be discovered from data. Then I'll talk about three themes: the effects of interventions, the evolution of the *do*-calculus, and the algorithmization of counterfactuals. I'll also talk about applications: evaluation of plans and policies, mediation (i.e., distinguishing between direct and indirect causes), and generalization.

I start with the basic statistical problem and the paradigm that rules statistical thinking and most of machine learning. The idea is that someplace behind the scenes there is a Santa Claus called the "joint probability distribution" that occasionally, when he or she is gracious enough, spits out data. Our job is to infer Santa Claus's properties: some aspect $Q(P)$ of the joint distribution function P from the data; for example, we might want to estimate the mean, come up with a classifier, or decide whether a customer who bought Product A will also buy Product B. This kind of question is neat and well-formulated because it can be neatly encapsulated in the language of probability theory. We even have a short sentence to express this question: "Find the conditional probability of B given A," with conditional probability coming all the way from Reverend Bayes 250 years ago. The function P can be a very complex distribution defined on many variables, some continuous and some binary, and so on. Although this is not a simple computational problem,



CAUSAL MODEL:
THE WORLD AS A COLLECTION
OF SPRINGS

Definition: A structural causal model is a 4-tuple $\langle V, U, F, P(u)\rangle$, where

- $V = \{V_1, \ldots, V_n\}$ are endogeneas variables
- $U = \{U_1, \ldots, U_m\}$ are background variables
- $F = \{f_1, \ldots, f_n\}$ are functions determining V, $v_i = f_i(v, u)$ e.g., $y = \alpha + \beta x + u_Y$
- $P(u)$ is a distribution over U

$P(u)$ and F induce a distribution $P(v)$ over observable variables

THE FUNDAMENTAL THEOREM
OF CAUSAL INFERENCE

Causal Markov Theorem:
Any distribution generated by Markovian structural model *M* (recursive, with independent disturbances) can be factorized as

$$P(v_1, v_2, \ldots, v_n) = \prod_i P(v_i \mid pa_i)$$

Where pa_i are the (values of) the parents of V_i in the causal diagram associated with *M*.

Corollary: (Truncated factorization, Manipulation Theorem)
The distribution generated by an intervention *do*(*X*=*x*) (in a Markovian model *M*) is given by the truncated factorization

$$P(v_1, v_2, \ldots, v_n \mid do(x)) = \prod_{i \mid V_i \notin X} P(v_i \mid pa_i) \Big|_{X=x}$$

Figure 2.5 Structured causal models and truncated factorization.

the paradigm is clear enough. Causal reasoning, however, deals with a different paradigm.

You ask a question, for instance, “Infer whether customers who bought Product A would buy Product B if we double the price.” So, here we get up in the morning, whimsically greedy, and wonder what would happen if we raised the price. And we ask the question, “What will the probability of B given A be after we do something that perhaps has not been done before, like doubling the price of the product.” This is not even an aspect of the probability distribution *P*; observing that the price has doubled (and what has happened as a consequence) is very different from doubling the price and seeing the consequence.

The counterfactual “had we doubled the price” is thus not an aspect or property of the Santa Claus. So, what is it? It is a property of a data-generating model that is behind the joint probability. As before, the joint probability spits out data, we get the samples, and we need to infer some property, but of what? Not of *P*, but of the data-generating model. This is the invariant strategy of Nature that I talked about before, sometimes called a “mechanism,” “recipe,” “law,” or “protocol”—all are counterfactual notions—by which Nature assigns values to variables.

This simple idea is torture for a statistician because it takes a leap of imagination to think of Nature rather than experiments or measurements. It is a traumatic experience for people outside artificial intelligence; I would like you to be aware of that if you ever talk to an outsider. [AUDIENCE LAUGHS]

Once we go there, let’s generalize it. Let’s imagine that the whole world is just a collection of springs. So, the model is fueled by a collection of functions that assign values to variables. Every variable is assigned a value that is a function of the other variables in the system (Figure 2.5). Some of the variables are exogenous; you do not care about their causes, but only about their effects. The rest are endogenous. And our job is to encode this on a machine so that the machine can provide reasonable and plausible answers to our reasonable questions.

The equations that we had for the spring example are typical: after Nature spends some time, maybe a billionth of a second, looking at *X*, multiplying it by constant, adding to it some noise, and deciding that *Y* deserves the value *y* (great work, mother Nature!), our job is to decipher the strategy of Nature. If this sounds too ambitious, at the very least we should be able to answer counterfactual queries if we have enough data.

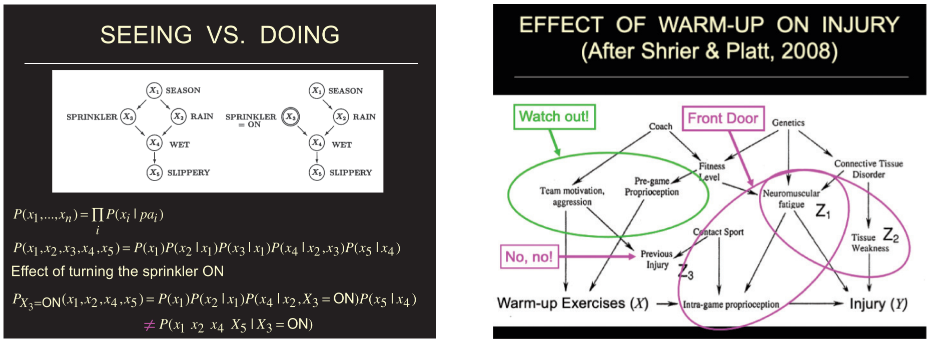
Let us illustrate this by considering a familiar digital circuit diagram. The circuit is an oracle for counterfactuals because if you look at the circuit you can answer a counterfactual question like "What if I were to replace this OR gate with an AND gate?" or "What if I were to connect this node *Y* to a power supply of 5 volts?" Even though the circuit designer never anticipated such crazy questions and events, the engineer glancing at the circuit has the ability to contemplate the answers and compute them correctly.

Where does this ability come from? It comes from some fundamental properties of the collection of functions and equations in the causal model. The fundamental one, from which everything else eventually derives, is that, if you happen to be lucky and your equations are recursive (no cycles there), and the disturbances happen to be independent of each other, then regardless of the functions that you have there and regardless of the distributions of disturbances, you can say something about the probability distribution of what you observe. So, the structure of that collection of springs determines something very basic in your distribution function, which has the form of a product and represents conditional independencies (Figure 2.5).

And from that comes the next corollary, which is the ability to answer questions about interventions. Once you have this product form, if somebody asks you, "And what if I take an action?," the answer comes from the truncated factorized product (Figure 2.5). This is the same factorized product as before, but we delete from the product those variables that are forced to a constant (by the interventation) because those variables no longer listen to their parents.

Here is our sprinkler example again (Figure 2.6). Before you act, you have the diamond structure shown in the figure, which corresponds to the set of equations shown. But once you take an action like turning the sprinkler on, you must remove the causal influence of the variable Season on the variable Sprinkler, as Mr. Sprinkler no longer listens to its parent, and instead becomes enslaved to your muscles, which set the variable to a value.

This formalism for actions did not germinate in AI, but originated with an economist, Haavelmo. In 1943, he considered the problem of modeling


SEEING VS. DOING

$P(x_1,...,x_n) = \prod_i P(x_i \mid pa_i)$

$P(x_1,x_2,x_3,x_4,x_5) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2,x_3)P(x_5 \mid x_4)$

Effect of turning the sprinkler ON

$P_{X_3=\text{ON}}(x_1,x_2,x_4,x_5) = P(x_1)P(x_2 \mid x_1)P(x_4 \mid x_2,X_3 = \text{ON})P(x_5 \mid x_4)$

$\neq P(x_1\ x_2\ x_4\ X_5 \mid X_3 = \text{ON})$

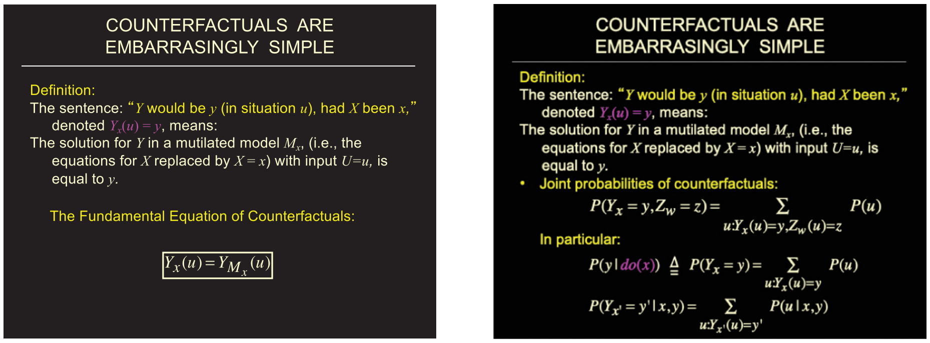
EFFECT OF WARM-UP ON INJURY
(After Shrier & Platt, 2008)

Figure 2.6 Structured causal models of two of the examples in the text.

government interventions in the economy, like fixing a price or imposing taxes, and he had the idea to model the effects of the actions by introducing changes in the equations. If the government does something like keeping a price constant, a term is added to the corresponding equation to balance the other terms, so that the price remains constant. Later on, this manipulation was replaced by Strotz and Wold, who "wiped out" the relevant equation and replaced it with a constant assignment. Then Spirtes, Glymour, and Scheines transformed this manipulation into a graphical surgery procedure, where you wipe out the arrows going into the manipulated variable, resulting in the truncated factorization. I took this all very seriously and said, "We have a new calculus that deserves algebraic support," translated it into the *do*-calculus, and then applied it to counterfactuals. That has been the evolution of these ideas. Now we also have the unification with the Neyman–Rubin account in statistics, which also handles causality with counterfactuals.

How are counterfactuals handled, and what is the general model for counterfactuals? This is all very simple (Figure 2.7). You mutilate your model to take care of the antecedent of the counterfactual, and you solve the equation in the mutilated model. There's nothing else to it; it's embarrassingly simple. In this Definition, I simply say symbolically what I said verbally: you are in possession of a calculus because you have a semantics for joint counterfactuals. For any set of variables $X, Y, Z, \dots$, you can find the joint probability of Y taking a value y had X been x, and simultaneously, Z taking value z had W been w, and so on. The semantics determines the probability of any such sentence.

Specifically, the sentences can involve actions with the "*do*" operator and attributions, like "What is the likelihood that a patient would be alive today had he not taken the drug, given that in fact he is dead and he took the drug?" This is a sentence in the language, and the semantics is there. If you have the model,



COUNTERFACTUALS ARE EMBARRASINGLY SIMPLE

Definition:
The sentence: "Y would be y (in situation u), had X been x,"
denoted $Y_x(u) = y$, means:
The solution for Y in a mutilated model M_x, (i.e., the equations for X replaced by $X = x$) with input $U=u$, is equal to y.

The Fundamental Equation of Counterfactuals:

$$\boxed{Y_x(u) = Y_{M_x}(u)}$$

COUNTERFACTUALS ARE EMBARRASINGLY SIMPLE

Definition:
The sentence: "Y would be y (in situation u), had X been x,"
denoted $Y_x(u) = y$, means:
The solution for Y in a mutilated model M_x, (i.e., the equations for X replaced by $X = x$) with input $U=u$, is equal to y.

- Joint probabilities of counterfactuals:

$$P(Y_x = y, Z_w = z) = \sum_{u: Y_x(u)=y, Z_w(u)=z} P(u)$$

In particular:

$$P(y \mid do(x)) \triangleq P(Y_x = y) = \sum_{u: Y_x(u)=y} P(u)$$

$$P(Y_{x'} = y' \mid x, y) = \sum_{u: Y_{x'}(u)=y'} P(u \mid x, y)$$

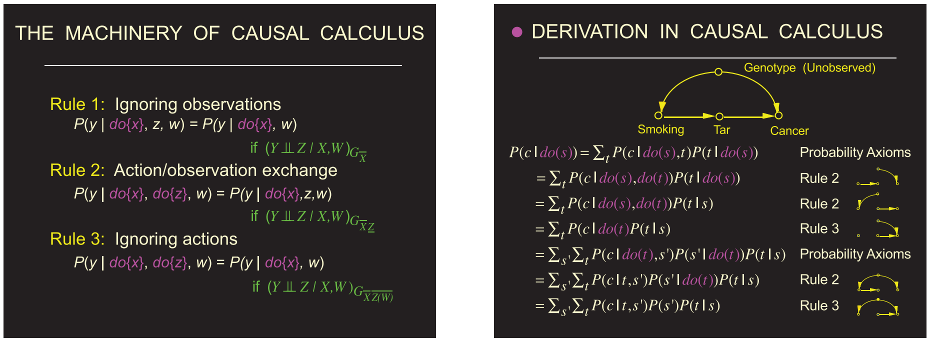
Figure 2.7 Counterfactuals are simple.

you can compute the answer. Everybody knows how to solve equations, right? The semantics is extremely simple.

And Joe Halpern and David Galles came up with a complete axiomatization of that. Why do we need an axiomatization? So that if anybody says, "You can do counterfactuals differently," you can compare the axioms and evaluate if they are equivalent or not. The workhorse is a composition axiom that tells you that if you do something that would have occurred anyhow, you have not done a thing. This sentence says, essentially, that our world is closer to our world than any other possible world, if you go to the possible worlds interpretation of it.

I'll give you now an example of what you can do with it. You have a collection of equations and you think that Nature works like that. The first questions that you have to ask yourself are "Is this model testable?" or "Does the model have any testable implications?" As I said before, if it does not have testable implications, you cannot learn or verify the model. And the idea for the verification is very simple. Everything that we did with Bayesian nets translates now into Causal Bayesian nets, and the criterion of d-separation gives you a finite set of testable implications. Just look at the missing arrows: every one carries the promise of a test. If the test fails, the model is wrong.

What else can these models do for you? They can handle interventions; they are, indeed, an oracle for interventions. So, if you have questions like "What is the average causal effect of X on Y, given that you can measure variables W and Z" or "Can you do this without manipulation, just by observation?", you can produce answers like "Yes, if you can measure variables like age or ethnicity." Namely, you are guaranteed that you can answer the query without bias by simple adjustment (regression). Of course, these results are built on the assumptions encoded in the causal graph. Each missing link in the graph is an assumption of a causal nature, not of a statistical nature.



THE MACHINERY OF CAUSAL CALCULUS

Rule 1: Ignoring observations

$P(y \mid do\{x\}, z, w) = P(y \mid do\{x\}, w)$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}$

Rule 2: Action/observation exchange

$P(y \mid do\{x\}, do\{z\}, w) = P(y \mid do\{x\}, z, w)$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$

Rule 3: Ignoring actions

$P(y \mid do\{x\}, do\{z\}, w) = P(y \mid do\{x\}, w)$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\,\overline{Z(W)}}}$

DERIVATION IN CAUSAL CALCULUS

Genotype (Unobserved)
Smoking
Tar
Cancer

$P(c \mid do(s)) = \sum_t P(c \mid do(s), t) P(t \mid do(s))$ Probability Axioms

$= \sum_t P(c \mid do(s), do(t)) P(t \mid do(s))$ Rule 2

$= \sum_t P(c \mid do(s), do(t)) P(t \mid s)$ Rule 2

$= \sum_t P(c \mid do(t) P(t \mid s)$ Rule 3

$= \sum_{s'} \sum_t P(c \mid do(t), s') P(s' \mid do(t)) P(t \mid s)$ Probability Axioms

$= \sum_{s'} \sum_t P(c \mid t, s') P(s' \mid do(t)) P(t \mid s)$ Rule 2

$= \sum_{s'} \sum_t P(c \mid t, s') P(s') P(t \mid s)$ Rule 3

Figure 2.8 Causal calculus in action.

Here is another example, one which is highly applicable. You are in the sports medicine business, and you wonder whether warm-up is a cause of injury or prevents injuries in the game (Figure 2.6). It's an extremely important question for our society, for our culture, right? You can take measurements of previous injuries, team aggressiveness, and so on. Which one would you measure? Each one takes a lot of dollars to measure. The answer is given to you automatically: "Thou shalt measure this, and you're okay; thou shalt not measure that because you would get bias; thou shalt measure that—fine, and here there is another alternative." Indeed, you can pick the measurements according to their cost and their reliability. There are three rules that drive this answering mechanism (Figure 2.8). The rules take the graph into account, are applied repeatedly, and produce the answer.

Another example: Does smoking cause cancer? The query given to you contains a causal symbol (Figure 2.8): the purple expression *do*(*s*) stands for doing the action of smoking. We do not have the data for the effect of this action: we cannot conduct randomized experiments on smokers. So we have to answer the query analytically. We apply the rules one after the other until we get rid of all the purple expressions. Once we do this, it means that you can answer the query from data obtained by hands-off, passive observations. And you can answer the question quantitatively: this is the extent to which smoking causes cancer.

What else can this calculus do for you? Find equivalent models, identify counterfactual queries, mediation, which is about the distinction between direct and indirect effects, explanation, which is about finding the causes of observed effects, and transportability, which is about generalizing what you learn in one domain into another domain in which you cannot conduct any experiments.

Counterfactuals are very interesting because philosophers have gone through a great deal of pain to understand why we are able to agree on their truth value. Here is a typical example: "If Oswald didn't kill Kennedy, someone else did" and

"If Oswald hadn't killed Kennedy, someone else would have." If I give you this pair of sentences, you'll tell me "Yes" on the first one and "No" on the second. How are we able to agree on this? This was a puzzle for philosophers.

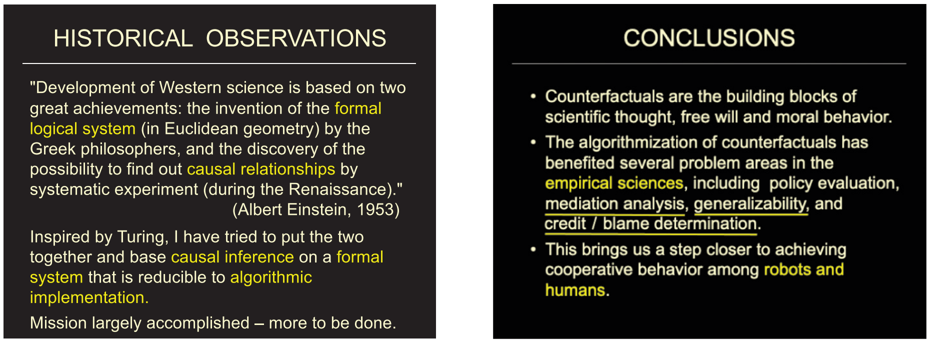
Hume tried to explain causes in terms of counterfactuals, and David Lewis tried to explain causes in those terms too. The puzzle that I faced was different. Why don't we try to define counterfactuals in terms of causes, rather than the other way around? Are counterfactuals less problematic? Apparently so, because we do form consensus on counterfactuals. And these two pillars of philosophy tried indeed to define causes in terms of counterfactuals. To me it means that we do count on a counterfactual engine in our mind that is swift and reliable, and we form consensus because we share the architecture of this engine. So, this is an AI problem, not a philosophy problem.

Indeed, what Lewis came up with in his possible-worlds semantics for counterfactuals does not solve the consensus puzzle as it relies on assessing, for example, how close is a world in which we are all dead after Nixon presses the button, relative to a world in which Nixon presses the button but somebody disconnected the wires. That is a typical question in philosophy—assessment of how similar worlds are. In our structural world, you do not rely on similarity among worlds; you rely on equations which are common equations of physics, and mutilating those equations.

I will not have time to talk about the counterfactual triumph, which is the ability to distinguish between direct and indirect effects. It is an important distinction because we send people to prison if they are directly responsible for murder, and fine them if they are only indirectly responsible. So, it is a key notion in law, in ethics, and in understanding how the world works. However, it requires the ability to answer questions about different kinds of interventions—interventions where you enable and disable certain mechanisms, rather than fixing variables, as I mentioned before.

Direct and indirect effects is a booming field now in statistical epidemiology, called "mediation analysis." And the impetus for that was counterfactuals. We were able to express the idea of indirect effects by counterfactuals, as you see here. What is the definition of "indirect effect?" It is the expected change in output when we keep the input constant but change the mediator. "What would you have gotten had the input changed?" is a nested counterfactual that is not about fixing the value of variables. It is now the accepted definition when you have indirect effects. That's why I consider this account a triumph.

I'll now talk about the next triumph: transportability. And I say it's a triumph because here the *do*-calculus appeared out of the blue. We didn't expect it to reveal its potency in an area like that, which has very little to do with interventions.



HISTORICAL OBSERVATIONS

"Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance)."
(Albert Einstein, 1953)

Inspired by Turing, I have tried to put the two together and base causal inference on a formal system that is reducible to algorithmic implementation.

Mission largely accomplished – more to be done.

CONCLUSIONS

- Counterfactuals are the building blocks of scientific thought, free will and moral behavior.
- The algorithmization of counterfactuals has benefited several problem areas in the empirical sciences, including policy evaluation, mediation analysis, generalizability, and credit / blame determination.
- This brings us a step closer to achieving cooperative behavior among robots and humans.

Figure 2.9 Logic and experiment for a science of cause and effect.

Imagine that we want to transfer relationships that we learn from experiments in one environment to a different environment in which no experiments can be conducted. So, we can think about training a robot in the cockpit and moving him or her to another environment where only observations are allowed, but no interventions.

How much of the causal knowledge that the robot acquired in the cockpit is transferable? We typically want a crisp logical answer, yes or no, regarding whether a certain relationship is or is not transferable given what we know about the two environments. And this has surprisingly a complete answer; that is, an answer that cannot be improved. When the method says that the information cannot be transferred, we also get an explanation for why, in terms of the assumptions about the disparities and commonalities between the two environments.

I think I'm close to the end of the talk. I have five seconds. [LAUGHTER]

I didn't talk about our new game, which is meta-analysis, in which big data comes to play. Imagine that you have data coming from 1,000 hospitals in the United States or worldwide, each one conducted under different conditions with different populations. You want to use all this data to come up with an answer to a query in another environment, where no measurements are allowed. All you know is the structure. Can you do it or not? We look for a crisp, yes or no answer. And if you can, how? So, I go through the "how" over many slides here, which I'll have to skip. Believe me, there is a method here, and there is a lot of work to be done in terms of decomposing the relationships into sub-relationships for picking up from every study the commonalities, and for putting them together to come up with an unbiased estimate.

It is time to move to the conclusions (Figure 2.9). Counterfactuals are the building blocks of scientific thought, free will, and moral behavior. The algorithmization of counterfactuals has benefited several problems in the empirical sciences, and brings us a step closer to achieving cooperative behavior between humans and robots.

Historically—I have to play the sage at this point—Einstein noticed that there have been two major advances in Western science. One is the development of logic by the Greeks. The other is the recognition by Galileo that you can find cause–effect relationships from experiments. I'm following these paths, trying to combine the two: the logic of the Greeks with the experiments of Galileo, to come up with logically sound theories of causes and counterfactuals. Our mission is largely accomplished, but more remains to be done. Thank you. [APPLAUSE]

References

A. Balke and J. Pearl. 1994. Probabilistic Evaluation of Counterfactual Queries. In *Proceedings of the AAAI-94*, Seattle, WA, Volume I, 230–237.

A. Darwiche and J. Pearl. 1994. Symbolic Causal Networks for Reasoning about Actions and Plans. In *Proceedings of the AAAI-94*, Seattle, WA, Volume I, 238–244.

J. Pearl. 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the 2nd AAAI Conferences on Artificial Intelligence.* 1982, 133–136.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Burlington, MA.

J. Pearl. 2000. *Causality*: *Models, Reasoning, and Inference.* Cambridge University Press, Cambridge.