Stanford SOCIAL INNOVATION Review

Feature

The Case for Causal AI

By Sema K. Sgaier, Vincent Huang & Grace Charles

Stanford Social Innovation Review
Summer 2020

Stanford Social Innovation Review
www.ssir.org
Email: editor@ssir.org

Using artificial intelligence to predict behavior can lead to devastating policy mistakes. Health and development programs must learn to apply causal models that better explain why people behave the way they do to help identify the most effective levers for change.

The Case for Causal AI

BY SEMA K. SGAIER, VINCENT HUANG & GRACE CHARLES

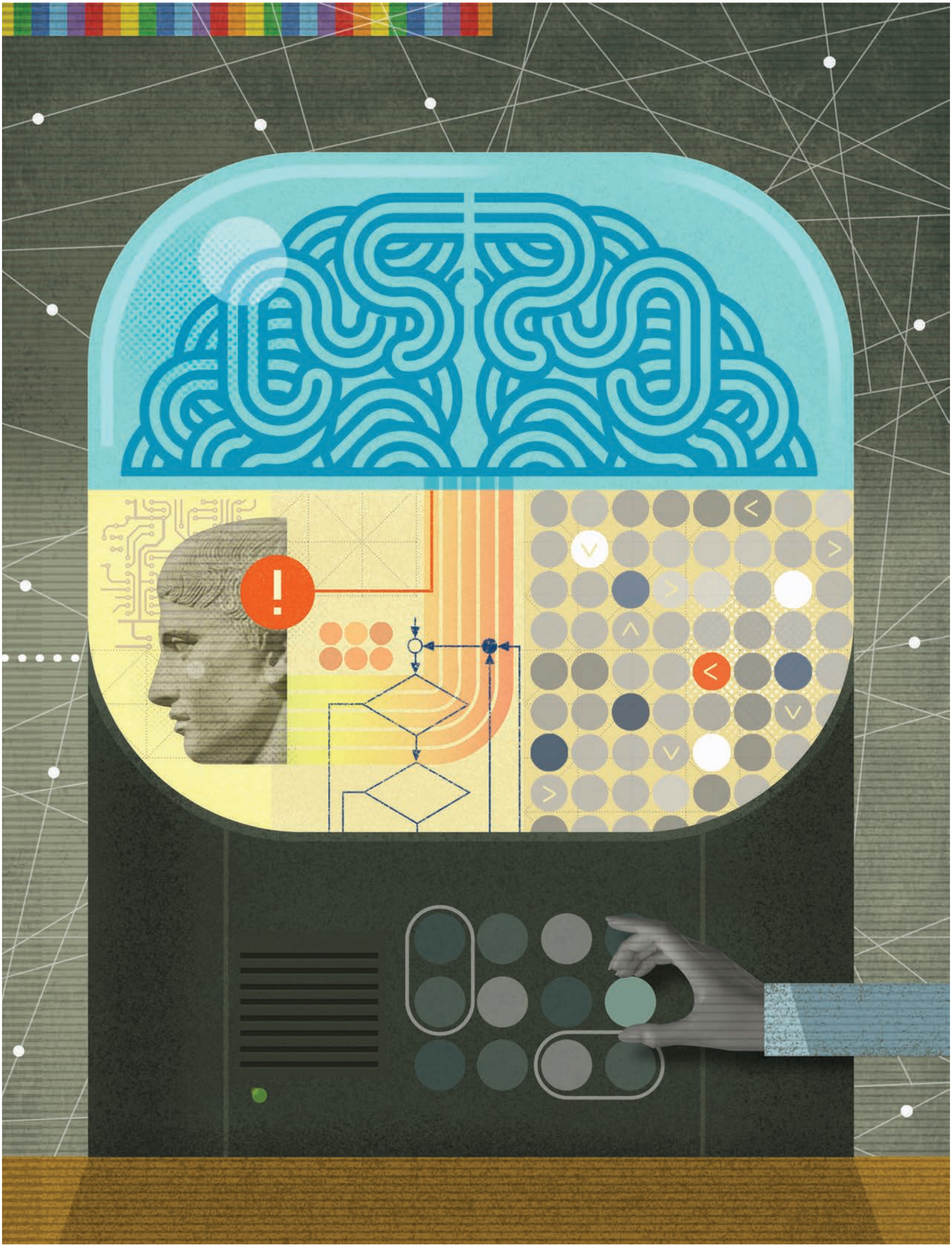
Illustration by Gordon Studer

Much of artificial intelligence (AI) in common use is dedicated to predicting people's behavior. It tries to anticipate your next purchase, your next mouse-click, your next job move. But such techniques can run into problems when they are used to analyze data for health and development programs. If we do not know the root causes of behavior, we could easily make poor decisions and support ineffective and prejudicial policies.

AI, for example, has made it possible for health-care systems to predict which patients are likely to have the most complex medical needs. In the United States, risk-prediction software is being applied to roughly 200 million people to anticipate which patients would benefit from extra medical care now, based on how much they are likely to cost the health-care system in the future. It employs predictive machine learning, a class of self-adaptive algorithms that improve their accuracy as they are provided new data. But as health researcher Ziad Obermeyer and his colleagues showed in a recent article in *Science* magazine, this particular tool had an unintended consequence: black patients who had more chronic illnesses than white patients were not flagged as needing extra care.

What went wrong? The algorithm used insurance claims data to predict patients' future health needs based on their recent health costs. But the algorithm's designers had not taken into account that health-care spending on black Americans is typically lower than on white Americans with similar health conditions, for reasons unrelated to how sick they are—such as barriers to health-care access, inadequate health care, or lack of insurance. Using health-care costs as a proxy for illness led the predictive algorithm to make recommendations that were accurate for white patients—lower health-care spending was the consequence of fewer health conditions—but perpetuated racial biases in care for black patients. The researchers notified the manufacturer, which ran tests using its own data, confirmed the problem, and collaborated with the researchers to remove the bias from the algorithm.

This story illustrates one of the perils of certain types of AI. No matter how sophisticated, predictive algorithms and their users can fall into the trap of equating correlation with causation—in other words, of thinking that because event X precedes event Y, X must be the cause of Y. A predictive model is useful for establishing the correlation between an event and an outcome. It says, "When we observe X, we can predict that Y will occur." But this is not the same as showing that Y occurs *because of* X. In the case of the health-care algorithm, higher rates of illness (X) were correctly correlated with higher health-care costs (Y) for white patients. X caused Y, and it was therefore accurate to use health-care costs as a predictor of future illness and health-care needs. But for black patients, higher rates of illness did not in general lead to higher costs, and the algorithm would not accurately predict their future health-care needs. There was correlation but not causation.



This matters as the world increasingly turns to AI to help solve pressing health and development challenges. Relying solely on predictive models of AI in areas as diverse as health care, justice, and agriculture risks devastating consequences when correlations are mistaken for causation. Therefore, it is imperative that decision makers also consider another AI approach—*causal AI*, which can help identify the precise relationships of cause and effect. Identifying the root causes of outcomes is not causal AI's only advantage; it also makes it possible to model interventions that can change those outcomes, by using causal AI algorithms to ask what-if questions. For example, if a specific training program is implemented to improve teacher competency, by how much should we expect student math test scores to improve? Simulating scenarios to evaluate and compare the potential effect of an intervention (or group of interventions) on an outcome avoids the time and expense of lengthy tests in the field.

Certainly, predictive AI algorithms have an important role to play if applied and used correctly. A good example is precision agriculture, which uses predictive AI to process data from satellite imagery and sensors to help farmers predict crop yields, detect disease and weeds, and recognize different species of plants. But being able to predict an outcome is not the same as understanding what actually causes it. Predicting that a farmer's crop yield will be lower this year is one thing; understanding why makes it possible to take steps to increase the harvest.

Another challenge with using only predictive models is a fundamental lack of knowledge about why they make particular predictions in the first place. This is a problem with deep learning—the kind of predictive AI that's at work in precision agriculture. Deep learning was inspired by how human brain cells are organized (in "layers") and how they communicate with each other (taking input signals from cells of one layer, transforming the signals, and outputting the transformed signals to cells of another layer). Unlike commonly used methods for predicting outcomes—such as regression, a traditional statistical technique that maps the relationships between variables to the predicted outcome with a single best mathematical formula—deep learning can map variables to outcomes with much more complex relationships between them. By combining multiple layers between the input variables and outcomes, deep learning algorithms can learn input-output relationships far more complex than a single mathematical formula and use them to predict outcomes. However, the links and intermediaries between variables are "black boxed," meaning that the users—and even the creators—of the algorithms cannot easily discern how the variables relate to the outcome and to each other. This means it is often impossible to know which input features deep learning models have used to make their predictions.

This opacity is unacceptable when dealing with the trajectory of people's lives, such as in the US criminal justice system. In 2016, 2.3 million American adults, or one in 111, were in prison, housed at great cost to federal and state governments. Courts throughout the United States have introduced "recidivism scores" in an attempt to lower incarceration costs by reducing the number of inmates without increasing crime. The recidivism score is a single number reached through a predictive algorithm that estimates the likelihood that a person convicted of a crime will reoffend. In theory, the score makes it possible for a judge to focus on incarcerating those more likely to commit additional crimes, and it should even help to remove potential bias in sentencing. But recidivism scores are inherently faulty because they are based on risk-assessment tools that pick up statistical correlations rather than causations. For example, low income is correlated with crime, but that does not mean it causes crime. Yet people from low-income households may automatically be assigned a high recidivism score, and as a result they are more likely to be sentenced to prison. Fixing the criminal justice system requires a focus on understanding the causes of crime, not merely its correlates.

SEMA K. SGAIER is cofounder and executive director of Surgo Foundation, adjunct assistant professor at the Harvard T. H. Chan School of Public Health, and affiliate assistant professor of global health at the University of Washington.

VINCENT HUANG is senior research scientist at Surgo Foundation.

GRACE CHARLES is research scientist at Surgo Foundation.

A closer look at causal AI will show how it can open up the black box within which purely predictive models of AI operate. Causal AI can move beyond correlation to highlight the precise relationships between causes and effects.

RANDOMIZED CONTROLLED TRIALS

The importance of testing causality is not new in either the health or development sectors. A straightforward way to do it is to conduct an intervention in people randomly assigned to one population group, known as the treatment group, and conduct no intervention in an otherwise identical group, known as the control group. By comparing the results between the two groups, it's possible to isolate the effect of the intervention. In clinical studies this is known as a *randomized controlled trial*, and in marketing research it's called *A/B testing*.

Development economists Michael Kremer, Abhijit Banerjee, and Esther Duflo were awarded the Nobel Prize in Economics in 2019 for spearheading the application of randomized controlled trials to identify root causes of development issues and to design solutions. Such trials have overturned some conventional wisdom about causality. For example, numerous observational studies had identified associations between vitamin D deficiency and increased risks of diabetes, hypertension, cardiovascular disease, and cancer. But randomized controlled trials demonstrated that vitamin D supplements do not reduce the risks of these conditions—they have not found a causal link between vitamin D supplements and health outcomes.

Randomized controlled trials, however, have limitations. Large groups of individuals are required to ensure that the results aren't biased or affected by coincidental, outlier characteristics such as age, sex, health status, or educational level. This tends to make such trials extremely expensive (in the millions of dollars) and time-consuming (they can take years to conduct). Furthermore, randomized controlled trials can test the effect of only one or at most a few bundled interventions at a time, despite the fact that health and social outcomes are complex, with many underlying drivers. Finally, they can predict only whether an intervention will cause an effect on a typical member of the treatment group, not on a specific individual.

This is where causal AI comes in. It offers new opportunities to test causality in individuals and population groups faster and more efficiently, along with the ability to unravel the underlying complexity. It allows researchers and program designers to simulate an intervention and infer causality by relying on already available data.

TWO APPROACHES TO DISCOVERING CAUSALITY

There are two approaches to causal AI that are based on long-known principles: the potential outcomes framework and causal graph models. Both approaches make it possible to test the effects of a potential intervention using real-world data. What makes them AI are the powerful underlying algorithms used to reveal the causal patterns in large data sets. But they differ in the number of potential causes that they can test for.

To understand the two methods and how they work—as well as their differences—consider the following hypothetical scenario: Researchers wanted to discover if an antismoking advertising campaign persuaded people to quit, but there was no control group because the ads were released nationally. They only had a data set showing whether individuals were exposed to the ads, whether they gave up smoking, and information on their demographics and other health behaviors. Even without a control group, causal AI provides ways to infer causality.

The potential outcomes framework, proposed by statisticians Paul Rosenbaum and Donald Rubin in 1983, compares the outcome (quitting smoking) of an individual who has been exposed to the cause of interest (the antismoking ad) with an inferred "potential outcome" of the same individual had he/she not been exposed. The challenge is of course that no data exists on nonexposure outcomes for a person who was in fact exposed to the campaign. So, for each individual who was exposed to the ad, the AI algorithms instead find an individual in the data set who was not exposed to the ad but who is identical in other significant respects (such as age, race, and education). In other words, an artificial control group is reverse engineered to mimic a randomized controlled trial. The limitation is that while it is able to solve the problem of having no control group, the potential outcomes framework can test the effect of only one prespecified intervention at a time—in this case, did the ad campaign lead to that person's decision to quit smoking?

Causal graph models, by contrast, can do more than test a single pair of variables for their cause-and-effect relationship. They can be used as exploratory tools to map all the different causal pathways to an outcome of interest and show how different variables relate to each other. Applying a causal graph to our antismoking campaign might show that exposure to the ad in a pharmacy caused some people to stop smoking directly but others to buy nicotine patches, which in turn caused them to quit.

There are several causal graph models. One widely used method is the structural equation model, in which researchers specify the variables that may interact and how they might do so, and the model then analyzes the data to reveal whether they actually do. While this model can test many such relationships in the data, the whole network of interaction between different variables needs to be specified using existing knowledge. The limitation of this model is that it tests only the linkages between the hypothesized variables: If the variables that actually cause the effect are not included among the specified ones, they won't be evaluated against the other options.

Another causal graph method is the causal Bayesian network, a term coined in the 1980s by computer scientist and philosopher Judea Pearl and named for 18th-century English statistician Thomas Bayes. This method estimates the relationships between all variables in a data set. It results in an intuitive visual map showing which variables influence each other, as well as the extent of their influence. The advantage of this approach is that, unlike in a structural equation model, these interactions do not need to be specified ahead of the test, making it a true discovery method.

Although causal Bayesian networks require an abundance of data to capture the universe of possible variables, the potential of this approach is exciting for several reasons. It enables the data-driven discovery of multiple causal relationships at the same time. In the example of the antismoking ad campaign, a causal Bayesian network might show how advertising and the availability of different quit-smoking aids each affected people's behavior, or it might reveal how personal aspirations played a role. Equally important, unlike the black box of predictive AI, in the causal AI approach the relationships between the variables (exposure to ads, the availability of nicotine patches) and the outcome (stopping smoking) become visible to researchers, program implementers, and policy makers.

Causal graphic models also make it possible to simulate many possible interventions simultaneously. For example, what if different antismoking ads targeted different age groups or combined a general campaign with outreach by peer educators? They also allow for the incorporation of expert knowledge to counter the possible limitations of a purely data-driven approach. Experts can, for instance, help to determine which variables should go into the model, they can place conditions on the model to improve its accuracy, and they can help understand results that are counterintuitive.

EFFECTIVE APPLICATION

The field of causal AI is evolving rapidly. As its potential becomes more apparent, researchers are putting it to work in fields as diverse as climate change and health, demonstrating its broad potential.

Climate change | Causal AI techniques have been applied to climate change to understand whether and how humans are one of its contributing causes and what drives people's beliefs about it. To investigate this question, British scientists used a causal AI technique called *counterfactual event attribution* in the potential outcomes framework to determine whether human-produced greenhouse gas emissions were an underlying cause of the deadly European heatwave of 2003, which by some estimates was responsible for more than 70,000 deaths. Using historical data, solar data, information on volcanic eruptions, and atmospheric data on greenhouse gases, aerosols, and ozone, the researchers simulated summer temperatures across Europe in 2003, with and without the impact of humans. They found that the heatwave was much more likely to occur when the model included activities such as air travel or electricity production than when those effects were excluded. Published in 2004, this was one of the first studies linking an extreme weather event to human activity, and it provided a powerful argument for reducing the greenhouse gases generated by such activity. The research has been cited by the United Nations' Intergovernmental Panel on Climate Change.

Causal AI has also identified the factors that lead people to become more polarized in their beliefs about climate change. Researchers surveyed participants from the United States and Australia and used Bayesian networks to model how different people responded to a range of messaging about climate change. They found that when presented with consensus information about climate change in an online survey, Americans who actively distrusted climate scientists responded by updating their beliefs in the *opposite* direction of the information they were given. This causal framework provided a new way to estimate the interconnected relationships between worldviews, scientific beliefs, and trust in scientists. Insights like this are important for shaping public perceptions of the need for action to combat climate change. Such results provide a framework for designing interventional messaging that takes into account how participants might react to information, based on their beliefs and backgrounds.

Childhood diarrhea | Causal AI offers opportunities to address widespread and complex health problems where other approaches have not been successful. Childhood diarrhea is one example. This illness is the second biggest cause of death globally among children under 5 years of age. Many factors are associated with diarrhea, but it is extremely challenging to disentangle the causal pathways, both biological and structural, of diarrheal disease. This makes designing effective interventions difficult.

A study in Pakistan used data from a national survey of more than 110,000 individuals from more than 15,000 households. The survey included household, social, environmental, and economic variables. When using multivariate regression, a traditional statistical technique, the researchers found 12 household variables that were significantly associated with diarrhea. However, these were not easy to interpret: For example, one variable was the number of rooms in the household. By contrast, analyzing the same data set with a causal Bayesian network produced a network map revealing three variables that directly influenced diarrheal disease in children: the use of dry-pit latrines rather than toilets connected to drains; reliance on a water source other than piped, river, or stream water; and lack of formal trash collection. If incorporated societally or by national policy, these insights could lead to effective interventions to reduce childhood diarrheal disease.

Maternal and newborn mortality rates | Mortality rates remain stubbornly high in many low-income countries for mothers and their newborns. Women delivering their babies at health-care facilities is critical for the survival and well-being of both mother and infant. Through a national incentive scheme that pays families to deliver their babies at facilities (300 Indian rupees [around $4] for the hospital delivery itself, and a further 300 Indian rupees if the mother has also made use of antenatal care), the Indian government has been able to rapidly improve the rate of institutional delivery. However, in many Indian states this trend has plateaued at about 80 percent.

At Surgo Foundation, we tried to understand why women were not choosing institutional delivery and what kinds of additional interventions were needed in order to get them to do so. Our work has used a variety of techniques, including causal AI, to identify why some families still decide to deliver at home. In the state of Uttar Pradesh, with a population of more than 230 million people, we conducted several large-scale quantitative surveys to measure a large number of potential drivers of institutional delivery. We then used a causal Bayesian network to discover the variables driving this behavior and identify which were the most promising targets for a public health intervention.

A broad set of variables was correlated with delivering in a health-care facility, but causal AI identified the direct causes. To our surprise, and counter to common belief, the mother's proximity to a health-care facility was not one of them—but access to transportation was. This suggested that the government should solve transportation issues rather than building more health facilities closer to families. We were also surprised to find that a belief about whether hospital deliveries were safer than home deliveries was far more important than beliefs about hospital cleanliness, staff competencies, and staff biases. Having a delivery plan also increased the likelihood of institutional delivery; so did the mother's awareness of financial incentives, validating the impact of the government's incentive scheme. Findings from this study are currently being used to model hypothetical scenarios and pilot an intervention in which frontline health workers help mothers in Uttar Pradesh develop detailed plans ahead of time for their delivery, such as where they will give birth, how they will reach the facility, and how they will pay for extra costs.

Causal AI indentifies the underlying web of causes of a behavior or event and furnishes critical insights that predictive models fail to provide.

SEVEN RECOMMENDATIONS TO SCALE

AI is being adopted by businesses and governments eager to improve processes, solve problems, and create efficiencies. It is equally important that people working on health and development issues study and scale up the use of causal AI. It offers a way forward with distinct advantages over purely predictive AI. Predictive models can provide powerful and often accurate information, such as identifying whether the result of a mammogram reading is likely to be a case of breast cancer. But causal AI can help by identifying the underlying web of causes of a behavior or event and furnishing critical insights that predictive models fail to provide, which can lead to more effective interventions that drive positive outcomes. Moreover, causal AI doesn't operate within a black box, allowing researchers to check the model's reasoning and reducing the risk of biases like those described earlier.

Three converging factors indicate that causal AI's time has come. First, advances in the field of AI are highlighting the many applications of causal approaches, and as models are refined, scaled up, and applied to novel situations, more is learned about their value and limitations. Second, large-scale data sets are becoming more readily available. Like a 4K ultra-high-definition TV that packs more pixels per square inch of screen than a standard-definition

TV of old, more data makes predictions clearer and more accurate, and boosts confidence in the insights gleaned from causal networks. Finally, the health and development sectors are placing an increasing emphasis on precision policy—that is, coming up with interventions that have the strongest results, in order to deploy limited resources where they can have the greatest effect. Causal AI is ideally positioned to meet this challenge.

The path toward successful uptake of these approaches will require some work. Below are seven recommendations that can facilitate the adoption and use of causal AI.

Make better use of data and improve their quality. Investments in several large-scale data-collection efforts have been made over the last decade. However, these data sets are often underused and could be mined further to extract more insights. While we are seeing growth in data, other challenges remain. Data sets often are fragmented and vary in quality. Linking different data sets is also a challenge—for example, when information in one data set is recorded at an individual level, and in another at a regional or national level. Designing common indicators to be used in all data-collection efforts in a country would help get the best from data sets once they're linked.

Collect more comprehensive data. Applying causal AI successfully requires understanding all the variables that may drive behaviors—structural factors like policies and laws as well as individual beliefs, motivations, biases, and influencers. If data collection is done with too many prior assumptions about what's important to collect, the causal variables that truly underlie behaviors or events may be missed and consequently lead to the wrong causal links being established.

Design scalable, high-performance open-source tools for applying causal AI algorithms. Proprietary algorithm platforms are costly, making them frequently inaccessible to the health and development sectors. Open-sourcing makes software free, more accessible, and of better quality in the long run since more people can examine the source codes and provide feedback. Some open-source algorithms (such as bnlearn) are available, but their accuracy and speed need improvement. Practitioners who are not experts in causal AI need to know what steps they should follow to apply this approach in their area. Surgo Foundation is developing tools to lower barriers to entry and help organizations new to causal AI to avoid process pitfalls. One example is an open-source tool that evaluates whether a given data set is amenable to the application of Bayesian networks, and which algorithms are best suited to use on it. Surgo is also developing a workflow guide to help causal AI make the leap from academic research to practical use in the field.

Mix artificial intelligence with human intelligence. A purely data-driven approach cannot solve development problems alone. Expert knowledge must be included throughout the process to make sure that researchers and program developers interpret causal networks correctly. Experts can improve the performance of causal AI by adding constraints that reflect practical knowledge of how systems work on the ground and identifying whether known confounding variables are missing from the data. And, as the use of causal AI increases, ethicists and policy experts will have important roles to play to ensure that the approach avoids the pitfalls of bias or inaccuracy that have sometimes dogged the application of predictive AI models.

Improve ways to evaluate algorithm performance. Computer scientists are researching ways to improve the accuracy and overall robustness of causal AI algorithms. A typical way to evaluate the accuracy of causal models is to compare results against known causal relationships. But what should a researcher do if there are no known causal relationships to validate a model? (After all, discovering those relationships is often the goal of performing causal AI in the first place.) Furthermore, what happens if the results of a causal AI model conflict with existing expert knowledge? One solution may be to generate artificial data sets with characteristics similar to a real data set, but with predetermined causal relationships between variables. Evaluating how well a causal AI model performs on an artificial data set can help researchers infer expected performance on a real data set with similar characteristics.

Demonstrate the value of causal AI in the development sector. The examples we have outlined above are powerful but few in number. Wider awareness of the work that is being done will help spur the uptake of causal approaches. Surgo Foundation is using causal AI to understand how to optimize the performance of frontline health workers, how to decide which interventions we should scale up to improve student learning, and how to improve uptake of modern family planning methods. As the foundation moves forward, we are looking to test the application of causal AI in areas such as agriculture and climate change.

Build the awareness and knowledge of key stakeholders. Causal AI is still a very novel concept for those outside the field. Work is required to explain its potential to policy makers and funders; program managers; and monitoring and evaluation experts in the many sectors where causal AI could be applied, so that they understand these approaches, at least conceptually.

THE NEXT LOGICAL STEP

In order to make sense of the world, humans take account of and analyze repeating patterns. We have come a long way from creating mythologies for explaining the weather to using rigorous data collection and mathematical modeling to predict the next rainfall or hurricane path. But we continually run up against the limits of what we are able to observe and the methods available to analyze our data.

Causal AI is the next logical step, made feasible by recent technological transformations and the increasing pervasiveness of data. Its advantage over some other disciplines in the social sciences—and indeed over predictive AI—is that it can help identify the precise causal factors that directly lead to particular behaviors or outcomes, and it can efficiently test different approaches to changing those behaviors or outcomes. This edge enables researchers and practitioners to focus on the best mix of interventions for addressing some of today's most critical issues, from climate change to health care. Better causal inferences will help programs do more with fewer resources and waste less time doing it. And by integrating causal AI with human expertise, programs can avoid the mistakes that arise when people—or the machines or software that they create—ignore crucial context or fall into the trap of mistaking correlation for causation.

Ultimately, knowing the "why" behind complex problems helps us to understand how the world really operates and, in turn, to identify the right actions to achieve desired outcomes. We may yet find that an ounce of causal AI is worth a pound of prediction. ■