

29. Opioids, COVID-19, and Inequality

Everything is complicated; if that were not so, life and poetry and everything else would be a bore.

—Wallace Stevens

In this final chapter, we apply many-model thinking to three salient policy issues: the opioid epidemic, the COVID pandemic, and economic inequality. We show how by engaging multiple models, we can better reason through these issues and better communicate the challenges that each presented. We can also see how, particularly in the case of COVID, that had the public and policymakers engaged multiple models, the costs of the pandemic would have been less severe. That said, we should not oversell the potential for models to avoid disaster.

Our treatment of the three cases differs. When analyzing the opioid epidemic, we include only the barest of details. We do not gather data or calibrate any models. Rather, we apply the models qualitatively to gain insights. This more superficial analysis serves as a template for how to apply many models when reasoning through policies or actions. On the COVID pandemic, we take a much deeper dive into the models applied and developed by researchers around the world. That analysis highlights the seven uses of models introduced early in the book. Finally, our analysis of income inequality, on the other hand, connects more tightly to the academic literature than to current policy concerns. Whereas many of the COVID models were built on similar foundations, our foray into inequality reveals a variety of models with unique sets of assumptions and different types of actors.

In all three cases, thinking with many models adds to our knowledge base and makes us wiser. That wisdom has limits. Given the complexity of these systems, even if guided by many models, as individuals and collectives, we will not reach full understandings, always take the appropriate actions, or design optimal policies. We will make mistakes. All models are wrong. The map is not the territory. We can learn from our models' shortcomings to build more and better models and become even wiser. The path toward

wisdom is paved with humility and a willingness to see our world through new and different lenses.

Many Models and the Opioid Epidemic

We begin with a many-model analysis of the opioid epidemic. To give some sense of its scale, according to one estimate in 2015, 4% of the population of the state of Massachusetts above age eleven had an opioid use disorder. Nationwide, in 2016, doctors wrote more than 200 million prescriptions for opioids, between 10 and 12 million people misused opioids, over 2 million people were classified as having an opioid use disorder, and more than 30,000 people died from opioid-related causes.

The primary reason that so many opioids were prescribed was that they work: they reduce pain. Given the 100 million Americans with chronic pain, opioids had an enormous potential market. The danger with opioids was, of course, the potential for people to become addicted. To make sense of how opioids received approval and how the epidemic arose, we apply four models to generate some core intuitions as to how the crisis came to be.

The first model, the multi-armed bandit model, explains why opioids were approved for use. When seeking drug approval, a pharmaceutical company runs clinical trials to demonstrate drug efficacy and a lack of deleterious side effects. We can model a clinical trial as a multi-armed bandit problem where one arm corresponds to prescribing the new drug and the other arm corresponds to a placebo or the existing drug.

A Model of Opioid Approval

Multi-Armed Bandit Model

To demonstrate their efficacy, opioids were tested against placebos. In clinical trials, patients were randomly assigned to take either opioids or a placebo. The assignment of the opioid can be modeled as one arm of a two-armed bandit and the placebo as the other arm. At the end of treatment,

each trial is classified as a success or a failure. Clinical tests found that patients who received opioids experienced (statistically) significantly less pain. Tests on patients who had hip replacements, dental surgery, and cancer treatments all found that opioids outperform placebos.

With any drug, the potential for addiction is a concern. Tests showed that a small percentage of patients, fewer than 1%, became addicted, allowing the drug to be approved. Those tests did not take into account the possibility that doctors would write longer prescriptions, in some cases a month's supply. The longer an individual takes opioids, the more likely that person becomes addicted. Empirical addiction rates exceeded 2.5% for patients with longer prescriptions. The Markov model shown in the box below shows how an increase in those rates from 1% to 2.5% can increase the equilibrium number of addicts 5-fold.

Transition-to-Addiction Model

Markov Model

A three-state Markov model reveals a nonlinear relationship between transitions to addiction and overall addiction rates. The model's states represent people not in pain, people using opioids, and addicts. We estimate the transition probabilities between those states, which we represent as arrows. The model on the left assumes that 1% of people who use opioids become addicted and that 10% of addicts revert to the no-pain state. It also assumes that 20% of the people in the no-pain state become opioid users. In equilibrium, only 2.2% of the population are addicts. To account for longer prescriptions, the model on the left assumes that 2.5% of people who use opioids become addicted and that 5% of addicts revert to the no-pain state. It also assumes that 20% of the people in the no-pain state become opioid users. Now, in equilibrium, 10% of the population are addicts.^{[1](#)}

image

The model's transition probabilities are only loosely calibrated to data. We are using the model to build intuition for how a relatively small rate of addiction can lead to a large number of addicts. By experimenting with the

model, we find that if we lower the probability of leaving the addict state and increase the probability of moving from the no-pain state to the opioid state, then the proportion of addicts can increase dramatically. If, for example, we lower the transition from addict to no pain to 1% in the second model, the proportion of addicts increases to 35%. One implication of this type of model thinking has been that some health care providers, such as Blue Cross, now limit the number of pills a doctor can prescribe. In addition, some states, including the State of Michigan, have passed laws restricting the number of pills in a single prescription.

Our third model relies on systems dynamics. This model, like the Markov model, assumes that there are people in pain, people who use opioids, and people no longer in pain. Rather than write transition probabilities between these states, however, it imagines a flow from people in pain to opioid use to people not in pain. More elaborate systems dynamics models can include sources for other drug providers, and allow for movements between opioid and heroin users. In addition, a richer model could include heterogeneous types of potential users. The fact that people who suffer from anxiety and depression are more likely to become addicted could therefore be included in the model.²

Paths to Heroin Addiction

Systems Dynamics Model

A population of people in pain produces opioid users and addicts. People on opioids flow into the no-pain state and also flow into the addict state. Addicts, in turn, can become heroin users. One reason that people use heroin is that they can no longer get opioids. Thus, as the flow of opioids increases, so does the number of heroin users.

image

A possible final model, which we do not write down formally, relies on social networks to explain why maps of per capita opioid use show clustering in rural counties. From our analysis of the square root rules, we know that smaller populations should have higher variation. (Recall the

example of the best and worst performing schools being small.) Higher use in rural areas could also be due to doctors who write prescriptions for more pills to accommodate rural patients who live farther from pharmacies. Those explanations aside, the clustering exceeds what would occur randomly. Clustering could arise if people provide or sell opioids to neighbors. Unlike used furniture, which people sell by placing ads, opioids cannot be sold in the open market. Most often, people sell them through trusted personal connections. A network model of opioid selling in which people distribute opioids through family and friends would likely produce local clusters of opioid abusers. The extent to which those clusters resembled the clusters seen in the data could be used to draw inferences about who sells to whom. And, that model could then be used to create interventions.

Other network and social influence models delve into what drives the choice to take or prescribe opioids or the potential for social influences to help people overcome their addictions, and even the interactions between the local economy and opioid addiction. Does, for example, the need to show up to work increase the prevalence of opioid use or reduce the likelihood of addiction? Or, perhaps it does both? Such questions, of course, require gathering data, but what data to gather and how to interpret it would be informed by models.³

Many Models and the COVID Pandemic

Models played prominent roles in the world's response to the COVID pandemic. Fatality rate models communicated the potential impact of the pandemic. Curve fitting and SIR models forecasted numbers of cases and locations of outbreaks. Empirical measures derived from models such as the *effective reproduction number* guided policy choices and individual behaviors. Elaborate microsimulation (agent-based) models informed decisions about what parts of society to close and what to keep open, and calibrated models of specific sectors and industries informed how re-opening occurred. The use of models to frame decisions became so prevalent during the COVID pandemic, that the terminology of epidemiological models became part of our daily conversations. We

worked to *flatten the curve*, to reduce the *probability of transmission* and prevent *superspreaders*.

A comprehensive analysis of the many models that were put to use during the pandemic—their successes, failures, and impact—could, and surely will, fill multiple volumes. Here, we focus on some of the more prominent models and their contributions. To frame how the models were used, some basic facts about COVID, though widely circulated, bear repeating. In December of 2019, a deadly strain of coronavirus, creating a disease known as COVID, was identified in the Wuhan province of China. Lethal variants of a coronavirus can cause severe acute respiratory syndrome (SARS).

By the end of January, the World Health Organization had declared it a global pandemic. Coronaviruses, named for their crown-shaped appearance when viewed under a microscope, are not rare. Along with rhinoviruses, they are a leading cause of the common cold. These can spread from person to person when someone coughs or sneezes. Droplets from sneezes can then linger in the air or land on surfaces and transfer the virus.

While the media focused much of their attention on the use of models to make forecasts of cases, hospitalizations, and fatalities, by far the more important contribution of those models was to inform interventions. By structuring our reasoning, models guided actions and informed the time and duration of interventions (Jewell, Lewnard, and Jewell 2020).

This dialogue between data and models happened rapidly. Empirical analyses of interventions based on models helped us understand the spread of COVID in Italy (Gatto et al. 2020) and on the potential effects of global travel restrictions (Chinazzi et al. 2020). These models were invaluable in helping countries and regions make wise policy choices. For example, almost all models showed that without stringent efforts to slow the virus's spread, upward of tens of millions of people would die. As a result, governments around the world restricted air travel, closed schools, and, in some cases, quarantined entire populations. Interventions were enacted broadly, even in countries and regions that had few cases because leaders trusted science and were persuaded by the evidence that had been structured by models. Nothing was spared. The 2020 Olympic games,

music festivals and cultural events, and professional conferences were all canceled or moved online.

In early March 2020, the Centers for Disease Control and Prevention announced that their models suggested that up to 1.7 million people in the United States could die from COVID and that the number of hospitalizations could exceed 20 million. To put those hospitalization numbers in perspective, the United States has approximately one million hospital beds (Fink 2020). Other projections were even more dire. A model developed by researchers from Imperial College in London estimated that without interventions, 2.2 million people in the United States and over 500,000 people in Britain would die from coronavirus (Ferguson et al. 2005, Ferguson et al. 2020).

These models played two important roles. Calibrated versions of these models helped to guide the allocation of respirators. Graphical versions that showed how flattening the curve would prevent hospitals from being overcrowded educated the public as to the need for drastic measures. To be effective, laws and policies require public support. Models can help build public support and drive behavioral changes. Unfortunately, not everyone believed what the models predicted. As would have been expected, people who trust science and understand models were more supportive of interventions than those who do not.

Models were also used to explore questions about the potential effects of viral mutation and how the efficacy of a vaccine will depend in large part on whether fear of the vaccine overpowers fear of the disease (Epstein 2020). If an insufficient number of people get vaccinated, and if groups of people who fear vaccines live near one another, they will be at risk of creating conditions for outbreaks.

Before discussing specific models, we should take a moment to discuss the difficulties associated with making accurate predictions about the extent of a pandemic whether measured in infections, hospitalizations, or fatalities. The central difficulty is that credible models change the world they are forecasting. A model that policymakers trust, one that predicts millions of deaths, often turns out to be a *self-unfulfilling prophecy*. We want predictions of dire consequences to be selfunfulfilling! We want

governments and people to take actions to prevent those deaths from occurring.⁴

To be accurate in the long run, a forecasting model must therefore predict the actions that governments will take and people's responses. Neither of those is easy. Government actions are notoriously hard to predict, and we spent an entire chapter on the difficulty of modeling people. Add to these difficulties the fact that a virus spreads largely through individual contacts, and as a pandemic begins its spread, our contact networks change. Predicting how people will change their behavior in a novel situation is difficult. How many of us will quarantine? Wear masks? These are hard questions to answer. We should not expect any model to be accurate, and even if one were, we probably could not identify that model before the fact.

The smarter approach, the one advocated in this book and followed by the Centers for Disease Control and Prevention, relies on many diverse models.⁵ Anthony Fauci, the director of the National Institute of Allergy and Infectious Diseases, often noted the CDC's reliance on multiple models.

To encourage the development of many models, the CDC sponsors annual contests among models to predict flu incidence. In those contests, the most accurate models are themselves ensembles of models. In the 2017–2018 influenza season, which by the way was the worst in the past decade with approximately 45 million people infected, a consortium of four research teams created an ensemble that combined twenty-one models. They weighted the models so as to maximize the ensemble's accuracy (Reich et al. 2019).⁶ That ensemble model outperformed all twenty-one of its component models.

The CDC contests embody the many-to-one principle. They also, as it turns out, satisfied the one-to-many principle. Detailed microsimulation (agent-based) seasonal flu models could be rewritten as COVID models. The ability of modelers to retool existing models endowed the CDC with multiple models to interrogate when making projections about cases, hospitalizations, and fatalities from COVID. All of those models, of course, would be wrong. But collectively, they would prove useful, and far more

accurate than claims made without models that the virus would disappear in summer or lead to a few thousand fatalities at most.

We begin our more formal analysis of COVID models by describing how *fatality rate models* were used to estimate approximately 2 million fatalities without interventions.⁷ A fatality rate model expresses the expected number of fatalities as the product of three terms: the population size, the proportion of people infected, and the fatality rate. To make a prediction, we need to assign values to each of the three terms on the right side of the equation.

Fatality Rate Model

The number of fatalities can be expressed as the product of the population size, the percentage of people infected, and the population level fatality rate.

Fatalities = Population Size x % Infected x Fatality Rate

No Intervention Estimate:

2,500, 000 = 333, 000, 000 x 30% x 2.5%

Best Case Scenario:

100, 000 = 333, 000, 000 x 3% x 1%

One of those terms is known. The population of the United States is approximately 330 million people. We must make predictions or forecasts of the other two terms. Data from China, Italy, and Spain gathered in the spring of 2020 estimated the fatality rate from COVID as being between 1% and 2.5%. The actual fatality rate for COVID depends on who catches the virus, which depends on behavior. Would people more at-risk take greater precautions? Recall that our task here is to predict fatalities assuming no extraordinary response. We might then assume that people would respond to the coronavirus in latter stages of the pandemic just as they had at the beginning. If so, we can conservatively estimate the fatality

rate to be 2.5%. In later, more sophisticated models, we will lower that estimate.

To estimate the proportion infected, we can look to past pandemics. The 1918 flu had a similar basic reproduction number to the novel coronavirus and infected approximately one-third of the world's population. In 1913, people traveled less and fewer people lived in cities. Medical care was less advanced, and information flowed much more slowly. These societal changes point in opposite directions. For more virulent flu viruses, such as the 2017–2018 seasonal flu, upward of 15% of people show symptoms. That year 40% of people received flu shots, which reduces the likelihood of illness by up to one-half. Combining all of this information, the one-third figure seems a reasonable initial benchmark.

If we multiply those three numbers together, we arrive at 2.5 million fatalities. To put that number in perspective, COVID loomed a larger threat to human life than the ten leading causes of mortality—heart disease, cancer, accidents, respiratory failure, stroke, Alzheimer's, influenza and pneumonia, nephritis, and suicide—*combined*. Actions were necessary.

We can also apply the fatality rate model to construct a best case scenario. To do so, we again use flu virus data to benchmark estimates. CDC data show that the flu season with the fewest cases in the past decade was 2011–2012 in which approximately 3% of the population was infected. That season's flu vaccine had an estimated effectiveness of 47%, which is relatively high. While there would be no available vaccine for COVID in the first six months, the proposed interventions for COVID could, in a best case scenario, reduce the incidence of the disease to that of a mildly virulent flu. For a fatality rate, we might optimistically assume that medical treatment and isolation of the most vulnerable could reduce the fatality rate to 1%. In a moment, we will give that assumption more justification. Multiplying those three numbers together, we arrive at a best case scenario of 100,000 fatalities, a 20-fold reduction in the number of fatalities.

We can now return to the question of how to estimate the realized fatality rate, which will depend on how people react to the virus. Recall the three types of models of human behavior: *rational choice*, *sociological*, and *psychological*. To make a robust prediction of the fatality rate, we should

consider all three. A rational-choice approach would predict that people balance the risk of death against the enjoyment of going out to dinner. If so, high-risk people will be far more likely to quarantine, lowering the realized fatality rate. That behavioral prediction requires that people know their risks. General health issues—obesity, diabetes, heart damage, respiratory problems, and the like—were known comorbidities.

Age functions as a reasonable proxy for these comorbidities. Early data showed that older people were more likely to suffer fatalities. Figure 29.1 shows the approximate population size of the United States broken up into ten-year age groups along with rough estimates of the COVID fatality rates for each group. The youngest two categories each contain approximately 40 million people and have fatality rates of less than one-fifth of a percent. The oldest two categories together contain 40 million people. These groups have fatality rates of 8% and 15%—more than forty times the fatality rate of the youngest categories.⁸

image

Figure 29.1: Population Sizes and Fatality Rates for US Population by Age

A rational-actor model would assume that people over fifty recognize the risk and quarantine themselves or, at least, more actively engage in social distancing. Given that people under fifty suffer few fatalities, a purely self-interested rational-actor model might assume that they reduce their social distancing. A rational-actor model that assumes people care about the health of others (which most people do) would assume that younger people who do not sequester would choose to avoid older people.

Sociological models of behavior would assume that highly educated scientifically minded people would place themselves at less risk than people who are skeptical of science, or who feel immune to ill effects of COVID. This last group would consist of mostly young people. That model might even assume that some young people and religious people would ignore the advice to socially distance and continue to meet in large groups. Thus, the sociological model would assume, contrary to a rational choice model, that some high-risk people would choose risky behaviors and pay the consequences.

Last, psychological models would assume that more affable, conscientious, and introverted people would be better at sequestering, and that people who are more neurotic, or who lack self-control or are more extroverted would be more likely to socialize. These personality characteristics could be captured in a model by adding personality type to a rational choice framework, and having personality enter into decisions. The result would be that some low-risk people would be more likely to sequester and some high-risk people would not.

If we combine all three of these models, we might expect (optimistically) that only 1% of people over fifty become infected, but that say 4% of those under fifty do. The large proportion for the younger people would result from both rational calculations and sociological conformity. Those assumptions produce an estimate just shy of 80,000 fatalities, a number close to our previous estimate of 100,000.⁹ More sophisticated models such as the ones we cover later produced similar projections.

The central takeaway is that interventions, if successful, could save millions of lives. However, there was no guarantee that interventions would produce numbers as low as the best case scenario. Estimates of how well interventions might succeed varied. One that informed policy in the United States was developed by the Institute for Health Metrics, and Evaluation (IHME). Early on, this model predicted that with successful interventions, the United States would have only 60,000 fatalities. The model, developed by Christopher Murray and colleagues, received an enormous amount of attention in the media and from politicians. Media coverage gave the impression that the IHME model alone was being used to predict outcomes and inform policies. As already noted, that was not the case; the CDC relied on multiple models. Even had the IHME model been known to be by far the best model (it was not), the CDC would have relied on multiple models.

The popularity of the IHME model stemmed from its simplicity, its brashness (the model presented narrow ranges of likely outcomes), and its exceptional graphical interface. Each of these three attributes positioned the model for media amplification. Ironically, use of the IHME model spread

through a contagion process. When leading media sources promoted the model, other media sources did the same.

While the features of the IHME model that drove its contagion—simplicity, confidence, and graphics—should not be the primary reasons for choosing among models, each can be justified. All else equal, simpler models are often better. They reveal logic more clearly and they can be communicated with less effort. Higher statistical confidence, the narrower outcome ranges, are also preferred. Who would not want a more accurate model? However, the error ranges reported by the model were conditional on assumptions that could not be expected to hold—as we learn when we unpack the model. Finally, choosing a model based on graphics may sound ridiculous, but the quality of graphics may signal quality of programmers or level of funding. The IHME model had deep-pocketed funders, including the Gates Foundation. As we learned from signaling models, people who lack full information rely on signals to make choices. In this instance, the Gates Foundation signal combined with fabulous graphics, functioned as the equivalent of a peacock's feathers.

To understand how the original IHME model made its predictions, we need to recall the SIR model from our study of contagious processes. That model shows that viruses will produce an S-shaped pattern of the number of infected people. The initial upward slope results from the increasing number of infected people. As the number of susceptible people decreases, the slope of the curve attenuates.

The initial IHME model was not based on an underlying SIR model. Instead, the model assumed that the number of cases and fatalities could be approximated by a *logistic function*. A logistic function is an S-shaped curve with time as the independent variable.

The shape of the S depends on three parameters: a *growth rate*, a *maximum value*, and a *center*. To make their predictions, IHME researchers “fit” data on the number of fatalities to a logistic curve. They used fatalities rather than the number of cases because early in a pandemic when testing is not widespread, fatality data are more accurate.

By fitting the data to the logistic function, they could estimate the growth rate, center, and maximum value, which in this case corresponds to the number of fatalities. The initial data that they fit to the logistic curve came from people largely unaware of the virus and an almost complete absence of government policies to reduce the virus's spread. Policies to quarantine, social distance, and close schools and businesses were in the process of being enacted.

The IHME model had to confront the question of how to model people's responses to the interventions. Here, the IHME team ran into a difficulty: *they had no explicit model of behavior, so how could they model a behavioral response?* They had constructed their estimate by leveraging the fact that diseases produce S-shaped curves and used data to fit such a curve. That approach is less a “model” than an exercise in fitting data to a curve derived from a model.

To account for the effects of government policies, they fit a second curve. Using data from regions in China, Spain, and Italy that had imposed strong stay-at-home orders and reduced the spread of the virus, they constructed a second curve. IHME now had two fitted curves. The first was a United States curve, which was the sum of a collection of state-level curves estimated from data. That curve had a high maximum number of fatalities. Call this the *no-intervention curve*. The second curve, which had a much lower maximum number of fatalities, corresponded to an *effective interventions curve*.

The IHME Model

The IHME model assumes that the empirical number of fatalities (cases) closely fits a **logistic function**.



This function produces an S-shaped curve as a function of time, t , and three parameters: a *maximum value*, m , a *growth rate*, g , and a *center*, c .



Using data on the number of fatalities, IHME estimated the parameters for each state in the United States. Using post-intervention data from China, Italy, and Spain, they also estimated parameters assuming interventions, a *post-intervention estimate*. If states adopted intervention policies, their estimated logistic function was predicted to shift to the post-intervention estimate.

The IHME team assumed that interventions would move the United States curve toward the effective interventions curve. To make their estimates, they assumed that each of four policies, school closings, closings of nonessential businesses, stay-at-home orders, and travel restrictions, would shift the US curve one-third of the distance to the China-Italy curve, with a maximum shift of the full amount. In other words, they assumed that once three of the policies had been adopted, that adding a fourth policy had no effect. That assumption could not be correct, but it built in conservatism on the effect of interventions.

Their assumptions may seem ad hoc. To an extent, they are. Why should each policy have the same effect? Why would the fourth policy have no effect? And, perhaps most troubling, why would the effects be additive? It is not clear that any one policy would be that effective alone. Those critiques aside, these were not unreasonable assumptions about policy efficacy. They just turned out to be wrong. People did not behave as (implicitly) predicted.

A deeper problem with the initial IHME model, or any curve-fitting exercise, is that it presumes that the spread of the virus will follow an S-curve. The model cannot predict a *second wave* in which a second S-shaped curve, a second wave, occurs. Second waves arise when few people are infected initially, and society then opens up, allowing the virus to spread anew. This occurred with the 1918 flu virus, and second waves of the coronavirus also appeared in late summer of 2020.

The “cannot get a second wave if you fit a curve” criticism may be a bit harsh. Think of the IHME model as a short-term model to predict where and approximately how fast the virus was likely to spread. For those purposes, the model was useful. (Yes, in the long run, the criticism holds.) To predict a second wave, the model would have to make explicit

assumptions about numbers of infected and susceptible people and rates of transmission. The IHME team knew this, and in later versions of their model, they included an SIR model as part of their calibration.

Before describing how to build a coronavirus-specific SIR model, we first return to the basic SIR model to show how it was used in communicating risk and in evaluating policy efficacy. Recall that with the SIR model, we can calculate the basic reproduction number, R_0 , which equals the product of the probability of spreading the virus and the probability of contact divided by the probability of recovery. If R_0 exceeds one, the virus will spread, and if it is less than one, the virus will dissipate. Early estimates of R_0 for the coronavirus were around 2.5, a number that when combined with the fatality rate caused sufficient concern for actions to be taken.

Those interventions and the resulting changes in behaviors changed the probability of spreading and the probability of contact. Epidemiologists, therefore, calculate the *effective reproduction number*, R_t , which is a function of time, t .

One quick and powerful method for gauging the success of interventions is to track R_t by region. Estimated effective reproduction numbers can identify hot spots and guide policy at regional and state levels. If estimates of R_t far exceeded one in a region, policymakers knew to take more aggressive actions to slow the spread of the virus. The often heard phrase of “flattening the curve” refers to reducing the effective reproduction number.

According to the SIR model, the goal of policy interventions during a pandemic should be to reduce the effective reproduction number below one. The difference between an effective reproduction number of 1.1 and an effective reproduction number of 0.9 equates to the difference between the disease continuing to spread through the population and the spread slowing.

Flattening (and Unflattening) the Curve

In the SIR model, the number of newly infected depends on the *effective reproduction number*. Reducing that number through more social distancing and fewer contacts *flattens the curve*. It reduces the maximal number of new cases, which occurs at a later date.

image

Flattening the curve reduces the peak number of cases. In doing so, it improves the capability of a health care system to respond. We do not need a model to know that. What the model can tell us, when calibrated to data, is approximately how much lowering the effective reproduction number will reduce the maximum number of beds needed. The SIR model also tells us something less intuitive: *flattening the curve pushes back the peak date*. This gives health care systems more time to prepare. In brief, lowering the effective reproduction number offers a win-win. It reduces the maximal number of cases and delays when those occur.

The policies that closed businesses and schools and limited travel proved successful, at least initially. As shown in figure 29.2, in Michigan, the effective reproduction number fell well below one soon after interventions were imposed. Over time though, people became less vigilant. Quarantining became difficult. People wanted to socialize. As evident in the graph, when policies were relaxed, the effective reproduction number crept back above one, meaning that the virus had picked up steam. In Michigan, by August, the effective reproduction number would once again fall below one. Figures for most other states followed a similar pattern, with the effective reproduction number falling below one but then increasing. That pattern reveals the adaptive response to the threat. When people perceived risk, they acted in such a way to reduce the virus's spread. When they felt the virus was under control, people relaxed, allowing the virus to spread more quickly. We saw that same dynamic in our *ping-pong model* covered earlier in the book: when people take actions that oppose a trend, they reverse the trend.

image

Figure 29.2: Sequestration Policy Effects on R_t in Michigan (rt.live)

Many of the models that government officials and health administrators used for making short-term forecasts of the number of cases and hospitalizations relied on variants of the SIR model. These models were calibrated to COVID and included features like a latency stage, degrees of infection, and the possibility of hospitalizations. The latency period, the time in which a person has the virus but shows no symptoms, could range from two to five days depending on the individual. Someone who experienced symptoms might recover on their own or require hospitalization.

We will describe one such model built by Marissa Eisenberg, a mathematical epidemiologist. Rather than three states—*susceptible*, *infected*, and *recovered*—the model includes eight states connected by ten arrows. Each arrow corresponds to a probability or rate of moving from one state to the next. The values assigned to those arrows are calibrated by region to develop regional policies.

In deciding how large to make the regions, the modeler once again confronts the *bias-variance tradeoff*. Make the regions too large, say an entire state, and the model's predictions will be biased for regions that differ demographically or in their density of connections. Make the regions too small, and the lack of data leads to high variance in predictions.

SIR with Latency and Severity (Eisenberg)

At any moment in time, each member of the population can be classified in one of eight states. The circle on the far left represents people who are *susceptible*, S . Once infected, people enter a *latent* (L) state where they cannot infect other people. After exiting the latent state, individuals either *recover* (R), or become infected. Infections can be either *severe* (I_s) or *mild* (I_m). In either of the infected states, people could infect other people.

People with severe cases become hospitalized (H). After entering the hospital, they either recover or suffer a fatality (F). People with mild cases do not enter the hospital. They either recover or seek care (C) and then recover.



A model that bases its assumptions on a large population, such as the entire state of Michigan or New York, would underestimate the fatality rate for regions with more older people and might underestimate infection rates for more dense regions. For instance, people living in upstate New York and the Upper Peninsula of Michigan are older but interact with fewer people than people living in New York City and Detroit. So, you would not want to lump upstate New York with New York City or the Upper Peninsula with Detroit.

Models like this do more than make predictions. They help policymakers reason. The Eisenberg model included a dashboard that had two sliders. Adjustments to the first slider changed how much people reduced their contacts.¹⁰ Contact rates could be set to some percentage of normal. The second slider set a start date and an end date for when social distancing occurred. According to the model, if no social distancing occurred, Michigan would have had over 20,000 cases per day in mid-April. If social distancing were put in place from April 1 until May 31 *and* if people reduced their contacts to 25% of normal, then in mid-April, the peak number of cases would have been cut in half.

The model also revealed the potential for a second wave if the interventions were strict and imposed for too short a time period. According to the model, if people reduced their contacts to 10% of normal for two months and then society returned to normal levels of contact, a second wave would occur in which nearly as many people would catch the virus in July and August as in April and May. This potential for a second wave weighed heavily on the minds of policymakers.

Keep in mind that given the difficulty of making accurate long-term predictions of a first wave, we should be skeptical of any model's quantitative estimates of the likelihood of a second wave or its size. And yet, every ounce of skepticism about numerical estimates should be balanced by a pound of engagement with the logic the model reveals: *too much flattening of the curve could leave so many people susceptible and invite a second wave*. We would expect that the first signs of a second wave

would lead to more shutdowns. That in fact happened in many states: policies that shut down parts of the society were reinstated. From our ping-pong model analogy, we should have expected interventions to reappear, and, more generally, should have known beforehand that managing the pandemic would require constant vigilance.

More detailed SIR models like the one just covered do not take full advantage of the types of data now available. The next model we cover, a microsimulation (agent-based) model developed by researchers at Imperial College in London, does. It can be best understood as a giant simulation of the real world (Ferguson et al. 2020). Agents in the computer model represent individual people who live in communities, belong to families, and go to work and school. Using census data, the model matches household sizes and age distributions. The model includes data on workplace sizes (not firm sizes), lengths of commutes, and school populations. The model distributes schools based on population density. Denser regions have more schools.

The Imperial College model was not constructed anew. It was a repurposing of an earlier model of the flu virus (Ferguson et al. 2005). To apply that model to COVID, the researchers performed the same type of calibration as in the previous cases. They estimated latency and varying levels of infectiousness. Because the model differentiates between where someone catches the virus, it can (and does) assume different rates of social contact. Based on past flu data, schools were assumed to have double the per capita contacts as the workplace, home, or community. That calibration implied that without interventions, infections would be equally likely to occur at work or school, within households, or in neighborhoods.

Imperial College Microsimulation Model

The model assumes four locations in which the virus could spread: **neighborhoods, households, schools, and workplaces**. Agents representing individuals in the population live in households (H_I) that are parts of neighborhoods (N_i). Household sizes and demographics and neighborhood densities and demographics are calibrated to data. Younger

agents within a household attend schools (S_k) in their neighborhoods with school and classroom sizes calibrated to data, meaning that neighborhoods with higher population density have more schools. Where people work (W_j) is assigned to calibrate with commute times.

image

The arrows in this diagram correspond to agents from neighborhoods going to school and work. Within each household and location, the model assumes an SIR model with latency and severity of infectiousness and a basic reproduction number of 2.4.

The Imperial College model includes far more detail than the other models. By including so much detail, the model enables us to explore what happens if schools close but workplaces are left open, or to estimate the consequences of keeping schools open while shutting down workplaces and limiting community interactions.

Models like this one that allow for many types of people and locations can also include heterogeneous behaviors. Some people may elect to follow the rules and maintain social distance, wear masks, and not gather in large groups. Other people may not. By altering behavioral assumptions within the model and by exploring the effects of policy alternatives, policymakers can identify crucial communities, behaviors, and policies. The estimates will not be perfectly accurate. If they are off by less than a factor of two or three, they will be of great value. Without a model, we would be left to conjecture the effects. Our intuitions about dynamical processes are so poor that we might make decisions that cost many lives.

The models we have covered so far have considered regions or entire countries. Industries, sectors, and organizations also built models to inform designs for reopening. We consider here one such model developed at Cornell University that used enrollment data to evaluate risk. Kim Weeden and Ben Cornwell constructed this model to calculate how many students co-enrolled in classes. They considered the entire university, undergraduates, and liberal arts majors. All three co-enrollment networks could be described as *small world networks*. Such networks have short path

lengths. Short path lengths facilitate sharing ideas, so universities like short path lengths. Unfortunately, so does a virus.

The numerical estimates from Cornell shocked college administrators. On average, each Cornell student took at least one class with 2.4% of the student body. Unfortunately, each student was also only one person removed from 60% of the population, and two students removed from over 95% of the population (Weeden and Cornwell 2020). In other words, start with any student at Cornell and you can get to any other student through just two other students. Viruses thrive on networks such as these.

To make sense of how these numbers could arise, we can perform the same sort of algebraic calculations used to explain six degrees of separation between any two people. Cornell University has approximately 15,000 undergraduates. Assume a student enrolls in one large class with 200 students, two medium-sized classes with 60 students, and two small classes with 20 students. That student would be co-enrolled with 360 students. With no overlap in students, that would amount to 2.4% of Cornell's student body. If, on average, those 360 students each took classes with only 25 other unique students, then the number of unique students who are connected to our initial student through just one other student would be 9,000, or 60% of the student population.¹¹ Each of those 9,000 students would need, on average, fewer than one unique connection to reach the 95% proportion for students one more connection away.

Based on the Cornell findings, other universities built models of their entire student population to explore different procedures for opening their campuses. Those exploratory models argued strongly against holding large classes. They led to design ideas for creating cohorts or bubbles of students. Similar applications of models can be found in every sector. As should be evident from this cursory overview, in helping society respond to the coronavirus, models performed all seven uses. They helped us reason and explain patterns. They were used to communicate concepts. They made predictions that guided actions and informed the design of interventions. The models helped administrators explore the implications of reopening plans.

On a final note, a great deal of ink has been spent characterizing how wrong the models have been. The IHME model with its estimates of 60,000 and then 80,000 fatalities has been a particular focus of criticism. Those projections implicitly assumed that Americans would lock down as well as people in other countries. We did not. The model proved wrong. Other models though, including simple fatality rate models, were nearly accurate. Imagine for a moment that we had not had models. How much worse might the responses of countries around the world have been? How many resources would have been misallocated? How many lives would have been unnecessarily lost? Or, imagine if *everyone* had believed the models and acted accordingly. How many lives might have been saved?

Many Models of Inequality

Our final many-model exercise delves relatively deep and wide into the causes of economic inequality. We undertake this effort for three reasons. First, inequality is one of the most important policy issues of our time. Income and wealth correlate with human flourishing. Higher-income individuals enjoy better health, longer life expectancy, and higher life satisfaction and happiness. Those at the bottom of the income distribution experience higher rates of homicide, divorce, mental illness, and anxiety.¹² We must be careful not to confuse correlation with causation: a substantial part of this correlation can be explained by the fact that healthier, happier people earn more money. Nevertheless, almost all studies show a connection between income and flourishing. No one prefers to be poorer. Second, we have a plethora of models of inequality written by a diversely tooled collection of economists, sociologists, political scientists, and even physicists and biologists. Third, we have abundant data on income and wealth within and across countries. We have both current data and time series stretching back hundreds of years.

We start by summarizing some empirical regularities. First, in all countries at all times, the distribution of income has an elongated tail, with many low-income people and a small percentage of people who earn large incomes. Historically, income distributions were calibrated to lognormal distribution or Pareto distributions. Recently, more granular data reveal the

tail to be longer than lognormal, though not quite that described by a power law. Wealth distribution is similarly skewed.



Figure 29.3: Income Shares of the Top 0.1%, 1916–2010. Source: Piketty 2014.

Second, within most developed countries, income and wealth inequality, however measured, have been rising in recent decades. Current levels of income and wealth inequality in the United States approach those of the Gilded Age. Shifts in entire distributions can be hard to discern, so, following convention, we describe those shifts with respect to the share of income that goes to the upper tail of the distribution. Figure 29.3 shows how the top 0.1% has increased its share of income. The share of income to the top 0.1% of families fell steadily through 1950, and remained stable at less than 4% until around 1980, when it began to climb. In 2018, the proportion of total wealth owned by these super rich was around 10%.

Third, globally, the number of people living in poverty has dropped precipitously. We should see no logical contradiction in these opposing trends. Fast-rising incomes in poor countries reduce cross-country differences and more than offset within-country increases in inequality. Our model of group selection produced similar effects. The growth in the number of altruistic communities outpaced the trend toward selfishness within each community.

Inequality has multiple, interwoven causes. Economic forces, sociological trends, exercises of political power, and the weight of history all contribute to disparities. Thus, as Steven Durlauf points out, we should not try to explain the levels or trends in disparities with a single equation. Nor should we base policy on one.¹³ We must be nuanced in our thinking. The processes concentrating wealth and income in the top 1% or top 0.1% may be unrelated to the forces trapping the bottom 20% in a cycle of poverty. To understand the disparate causes of inequality requires a variety of models.

We start by describing models that explain the changing distribution of income. Income has four sources: wages and salaries, business income,

capital income, and capital gains. The relative sizes of those shares vary by income level. Low-income people earn few capital gains or capital income. Many of the highest earners receive substantial income from every category. They earn income from wages, businesses, and capital.

Our first model extends the Cobb-Douglas production model to include two types of labor: educated and uneducated. The wage paid to a type of labor depends on the relative supply of that type and on technology.¹⁴ This model explains the recent rise in inequality based on supply and demand.

Technology and Human Capital Model

Growth Model

Output depends on *physical capital* (K), *educated labor* (S), and *uneducated labor* (U) as follows:

$$\text{Output} = AK^\alpha S^\beta U^\gamma$$

The parameters A , α , β , and γ capture the technology and the relative value of the three types of labor. The relative market wage for high- and low-skilled workers is as follows:¹⁵

image

Cause of inequality: Technological changes that favor educated workers increase β and decrease γ . These changes, along with increases in the supply of low-skilled workers, increase inequality.

During the 1950s, the rise in manufacturing increased demand for uneducated workers. At the same time, increased college enrollments due in part to the GI Bill increased the supply of educated workers. In the 1980s, decreased incentives to attend college slowed the growth in the number of college graduates, and a subsequent inflow of immigrants with low education levels increased the supply of low-skilled workers. At the same time, technological changes—the rise of automated manufacturing and the transition to a more digital economy—increased the relative value

of educated workers, and their rising wages reflected this change in relative value.

Time series data on average incomes by education level fit this model reasonably well. For this reason, many economists rely on the model to guide policy. The model advocates increasing access to education, as that will depress the wages of educated workers and reduce inequality. This model explains broad trends well, but it cannot explain the increase in variation within each income class.

The next model, the *positive feedback model*, can explain the increased variation within professions. It focuses on the tail of the distribution and, in particular, on entrepreneurs. In 2011, entrepreneurs made up 70% of the 400 wealthiest individuals in the United States.¹⁶ The model assumes that technologies—the internet and smartphones in particular—have made us more connected and more influenced by the choices of others.¹⁷ A person buying wireless stereo speakers can read reviews online and select “the best” from among a dozen choices. In the past, that person might have had a single option at her local stereo store. Now, a person who twists her knee can search the web and learn the identity of her favorite athlete’s doctor. That behavior creates a positive feedback and more inequality. We model socially influenced economic choices by reframing the preferential attachment model as a model that links positive feedbacks to talent.

Positive Feedbacks to Talent

Preferential Attachment Model

There exist N producers, and each begins with zero sales. The first consumer buys from a random producer with zero sales, giving that producer positive sales. Each subsequent consumer with probability p buys from a producer with zero sales, and with probability $(1 - p)$ buys from a producer with positive sales. When buying from a producer with positive sales, a customer selects randomly, with the probability of choosing a particular producer that is proportional to that producer’s sales.

Cause of inequality: Increased connectedness increases social influences, creating a positive feedback.

Though the positive feedbacks model cannot be fitted to time series data with the same fidelity as the previous technology model, we can look to experiments to see how feedbacks contribute to inequality. Recall the music lab experiments described in [Chapter 6](#). College students sampled and downloaded music under two treatments. In the first treatment, subjects could not see what music others had downloaded. This treatment captures the pre-internet world. In the second treatment, subjects could see the download numbers for each song. In the treatment without social information, no song receives more than two hundred downloads and only one song receives fewer than thirty. When people can see downloads, one song receives more than three hundred downloads and over half receive fewer than thirty. Information and social influence amplify the Matthew effect. The rich get even richer, and the poor become relatively poorer.

We can apply that same logic to the economy writ large.¹⁸ The potential for positive feedbacks through social networks to contribute to inequality depends in part on the nature of what people buy. Weightless goods such as movie and music downloads, web applications, and some technologies can be scaled quickly, if not immediately. Tractors, cars, and washing machines cannot be duplicated by clicking on an icon. So, while a new smartphone application can scale up with little to no capital outlay, a best-selling car cannot. As a benchmark, in May 2015, Volvo announced that it would build its S60 sedan in South Carolina. The company broke ground on the plant three months later. In 2018, the first S60s rolled off the line. Due to changing demand, the plant will also produce sport utility vehicles, but not until 2022.

Our next model applies the spatial voting model to explain the rise in CEO pay, which is not determined by social forces. In 2012, the average income of a CEO at a Fortune 500 company exceeded \$10 million, or roughly 300 times the average pay of a worker. By comparison, in 1966, the CEO made only about 25 times the average worker's salary. CEOs in other countries earn much less. In Japan, CEOs earn about 10 times what the average

worker does. In Canada and throughout Europe, CEOs earn approximately 20 times the pay of the average worker.

At most large companies in the United States, the CEO's pay is set by a compensation committee consisting of members of the board of directors. That pay includes salary, bonuses, and stock options. The people who determine the pay of CEOs are often other CEOs. They have an incentive for the pay of other CEOs to be high in order to drive up their own pay. We can use the spatial model to represent the preferences of the compensation committee. According to the spatial model, the salary will be set at the median voter's preferences. The difference in CEO pay by country can be explained by the composition of boards and the compensation committees. In Germany, boards of directors include workers, who prefer that the CEO be paid less.

The model explains the rise in CEO pay based on board members preferences over CEO. The preferences of compensation committee members could be informed by models or data. Or, those preferences might also be socially constructed or even part of an elaborate log roll, in which CEOs vote to raise one another's pay.

CEO Political Capture

Spatial Voting Model

CEO pay is determined by a vote of a compensation committee. In the United States, compensation committees include many current and former CEOs, who prefer higher pay, as well as compensation experts (X). Other countries include workers (W) on compensation committees, resulting in a median voter who prefers much lower pay.



Cause of inequality: CEOs determine their own pay through capture. Increases in the pay of any one CEO shifts preferences toward higher pay for all CEOs.

Our next model of income inequality comes from Thomas Piketty's best-selling book *Capital in the Twenty-First Century*. Picketty's analysis relies on models that show that the rate of return on capital exceeds the growth rate of capital. When that holds, the portion of income that high-income individuals receive from returns on capital will increase over time. By constructing more elaborate versions of the growth models from [Chapter 8](#), it can be shown that the return to capital will generally exceed the rate of growth in the broader economy. Over the long haul, an economy might grow at less than 2% or 3%, but returns to capital will be greater.

It follows that in an economy that consists of workers who earn wages and capitalists who earn income from rents, the share of income going to capitalists will increase relative to the share that goes to income. To be a bit more formal, the rate at which capital increases will depend on three rates: the consumption rate, the tax rate, and the return on capital.

Consumption depends on the level of capital. A person with little capital will consume a large percentage of her income. A person who owns a substantial amount of capital will consume a small percentage of her income. As shown formally in the box below, if we make the consumption level constant, the consumption rate will equal that amount divided by the level of capital. Thus, wealthier people will consume at a lower rate making it more likely that their net capital increases.

Rent-from-Capital Model (Piketty)

Rule of 72

The economy consists of **workers** and **capitalists**. The wages of workers increase at a rate g , the growth rate in the economy. The capitalists have wealth W_t at time t and earn return r (net of taxes) and consume a constant amount A . The income of capitalists will increase faster than that of workers if and only if

image

Cause of inequality: In a market economy, the rate of return on capital exceeds the overall rate of growth ($r > g$). Capitalists with large accumulations of wealth spend a small proportion of their income from capital on consumption, so their share of total income increases over time.

To see how the difference in rates produces inequality, we can apply the rule of 72. If initially the incomes of workers equal those of capitalists and wages grow at 2% while capital grows at 6%, then in thirty-six years wages double but income from capital increases 8-fold. Within seventy-two years, capitalists earn sixty-four times the income of workers.

Piketty applies this model to explain long-term trends in inequality of both income and wealth. The model calibrates remarkably well with three centuries of data from France and England. The model also sheds light on patterns of inequality over the past century in the United States and Europe, in which the two world wars destroyed capital stocks in Europe, evening out the income and capital distributions there. One reason the model fits the data as well as it does is that it omits two effects that cancel out. By excluding entrepreneurs, the model understates inequality. In assuming that all succeeding generations of capitalists invest wisely—not all do—the model overstates capital accumulation's contribution to inequality. The creation of a new class of rich individuals and the loss of an old class of rich individuals need not balance out. A more granular model would include movement in and out of the wealthy class.

That caveat aside, the model's implication is that so long as capital increases, capitalists earn an increasing portion of the economic pie. If we keep applying the rule of 72, we find that the income of the capitalists eventually dwarfs that of the workers. The problem of capital accumulation has a straightforward solution: impose a wealth tax. That may not be politically possible. As an alternative, we might wait for a war or revolution to redistribute wealth by force or for technological breakthroughs that produce a new set of wealthy capitalists.

Our next two models give priority to sociological forces. Both also have strong empirical support. The first explains rising inequality based on *assortative mating*. A family's income depends on the incomes of both partners. If a low-income person marries a high-income person, then that

marriage will contribute toward equalizing income distributions. If high-income people marry other high earners, then income disparities will increase. Most people marry at an age when a potential partner's lifetime income cannot be known with certainty. People do know the education level and general health of potential partners and get signals of their ambitions. Evidence shows that as men and women become more educated and earn higher incomes—refer back to the technology and human capital model—they choose life partners who also have higher education levels.

Assortative Mating

Sorting Model and Categories

Each individual has an education level: $\{1, 2, 3, 4, 5\}$ where 1 = dropout, 2 = high school diploma, 3 = some college, 4 = college degree, and 5 = postgraduate.

Let $P(m, j)$ and $P(w, j)$ denote the probability that a man and woman have education level j . Income (g, l) equals the (estimated) income of a person of gender g and income level l . Household income for a couple consisting of a man with education level l_M and a woman with education level l_W earns the following estimated household income:¹⁹

$$\text{Income}(M, l_M) + \text{Income}(W, l_W)$$


Cause of inequality: Increases in the number of educated women, increased pay for workers with higher levels of education, and assortative mating (the tendency for people to marry others of the same income level) result in an increase in household-level income inequality.

The increase in inequality results from the following factors. First, women increasingly earn college degrees. Second, relative income increases with education level. Third, educated men and women prefer educated partners. Therefore, families with two educated people will be more likely to have two high incomes contributing to household-level income inequality. The logic is airtight. The only question concerns the size of the effect.²⁰

Sociologists calibrate the model by categorizing people into five education levels: dropout, high school, some college, college degree, and postgraduate. They then calculate the average income for each education level and fit the data for the number of marriages between each pair of education levels, resulting in a crude approximation of the impact of assortative mating.

Had marriages been random rather than assortative, income inequality would be much less. One study finds that inequality as measured by the Gini coefficient, a common measure of inequality, would have decreased by 25%.^{[21](#)}

Our next model analyzes movements between income categories using a Markov model. It categorizes people (or households) by income level: high, upper middle, lower middle, and low. Each category contains one-fourth of the distribution. Given a time period—it could be a year, a decade, or a generation—we can then estimate the transition probabilities between income categories to capture mobility.

If there were no stickiness across generations, then the income of the child of a high-income parent would be equally likely to belong to any of the four income classes—all of the transition probabilities equal . In the most extreme case of no mobility, transition probabilities would consist of only 1s along the diagonal. Empirical estimates suggest that the reality lies between these extremes.

We can run experiments by taking 100 randomly selected low- or high-income families and computing the probability distribution of incomes in subsequent generations. Using the probabilities shown in the box, the children of high-income parents have a 60% chance of being high-income and only a 5% chance of being low-income. The grandchildren of the high-income parents have less than a 43% chance of being high-income and more than a 10% chance of being low-income.^{[22](#)}

The income dynamics model also serves as a baseline from which to evaluate the causes of income mobility. We might use a linear model to estimate a child's income as a function of parental wealth, parental income,

and parental ability levels (assuming we had data). The Piketty model would imply a positive coefficient on parental wealth. The ability-based model would imply a positive coefficient on parental ability given that there exists some correlation between parental ability and ability of offspring.

Intergeneration Income (Wealth) Dynamics

Markov Model

The population can be divided into four income (or wealth) categories with equal numbers of people. We can estimate the **transition probability** that an individual (or family) in one category moves to another category within a generation, as shown in the figure below. More equal transition probabilities correspond to greater **social mobility**.



Transition Probabilities Between Income Levels

Cause of inequality: Social skills, tacit knowledge, attitudes toward risk and education, and bequests reduce mobility between income classes.

Note that determining the coefficient on parental income requires data on the income of each child and each parent. Scholars have individual-level income data only for the past few decades. In *The Son Also Rises*, Gregory Clark (2014) found a novel solution to the problem of lack of data: he relies on surnames. He calculates the average income of everyone named, say, Thatcher, in 1888 and compares this to the average income of everyone named Thatcher in 1917. The thirty-year increment represents the length of a work life. He finds substantial correlation across surnames' average incomes, suggesting a lack of income mobility.

This type of model allows us to identify racial differences in intergenerational transfers. African Americans exhibit less persistence of wealth at the top of the income distribution and more persistence at the low end. A wealthy African American will be less likely to have wealthy

children, and a poor African American will be more likely to have poor children.²³

Our last model based on neighborhood effects, Durlauf's *persistent inequality model*, leverages the empirical regularity that people segregate by income category—that is, high-income people live in communities with other high-income people and low-income people live near low-income people. Segregation by income produces economic, sociological, and psychological externalities that reduce mobility. In the model, an individual's income depends on ability, educational spending, and spillovers.

Persistent Inequality (Durlauf)

Schelling Segregation Model + Local Majority Model

Individuals belong to income classes and segregate residentially by income. Individuals allocate a portion of their income to education, resulting in positive spillovers that increase with community income level. The future income of a child living in community C depends on her innate ability, spending on education, and spillovers. The contributions of education and spillovers depend on the level of income within the community, I_C .

$$\text{Income}_C = F(\text{ability}, \text{education}(I_C), \text{spillover}(I_C))$$

Cause of inequality: Children who grow up in low-income neighborhoods receive fewer educational opportunities and economic spillovers.

The educational attribute captures public spending on education, which empirically correlates with average income: high-income locations spend more on education than low-income locations, resulting in better educational outcomes and higher incomes for children in high-income neighborhoods.

The spillover term can be interpreted as socially transmitted knowledge of appropriate tools to acquire. Here we can link Durlauf's model to how people who live in high-income communities gain awareness of appropriate

tools. We can also link the model to our network model and the *strength of weak ties* phenomenon: people who live in high-income communities will be connected indirectly to more people with access to economically valuable information. This will produce a positive feedback on income.

We can also interpret the spillover as socially transmitted behaviors, such as the number of hours spent studying or working. If income includes a random component, then a person in a low-income community will see (correctly) a low return to time spent on self-improvement. Relatedly, the spillover could include psychological attributes—a positive or negative outlook on life, a feeling of safety, or a belief in oneself.

In the complete model, Durlauf solves for equilibrium levels of educational spending and derives conditions in which persistent inequality arises. That inequality results from what he calls *poverty traps*. Individuals living in low-income communities lack the educational resources and levels of spillovers necessary to earn high incomes regardless of their ability levels. Durlauf's model can help to explain the enormous racial gaps in income levels. African Americans disproportionately live in poor neighborhoods, and as a result, they may become trapped in low-income trajectories owing to a lack of resources on multiple dimensions.

Sociologists refer to this as *compounded disadvantage* (Sampson 2019). The persistent inequality model calls into question the efficacy of what we earlier defined as big coefficient thinking to reduce racial inequality. African Americans attend weaker schools, have access to fewer family resources—both in terms of financial capital and human capital. They live in neighborhoods with more crime and less healthy water and air. Identifying which one of these impediments has the largest effect and fixing it will have little impact if the other forces remain in play. Such efforts would be like small steps up a slippery hill, quickly erased by a larger force. This model suggests that to be effective, a policy must work to create a new reality, which would require multiple simultaneous large actions.

We have now covered a number of models that each describe a distinct cause of income inequality. In a sense, each is correct, but, as we know, each model is also wrong. This can be seen by examining their explanatory

contributions. The models vary in how much and what part of the variation in income they can explain. For the upper end of the income distribution, the empirical evidence most strongly supports the models that rely on technological change.²⁴ For over twenty years, the IRS has tracked the highest 400 incomes. Those at the top of the distribution come from technology, mass retail, and finance, three industries that can scale quickly. That high growth rate could stem from winner-take-all markets for search engines (Google) or social networking sites (Facebook). These models tell us little about the lower end of the income distribution. Nor do they say much about income mobility, or explain why CEO pay in the United States far exceeds that in other countries.

To explain these other features of the data, we need the other models, such as the income mobility model, Durlauf's persistent inequality model, and the spatial voting model. By constructing a dialogue between multiple models and data, we come away with a deep, multifaceted understanding of the causes of inequality. We identify multiple processes that produce and maintain inequality and see how they overlap and intersect. Our understanding of the complexity of inequality and the self-reinforcing causal forces that sustain it should make us dubious of quick fixes. As discussed in the context of the persistent inequality model, real change will require concentrated efforts on multiple fronts.

Into the World

We have just learned how by applying many models as an ensemble we can explicate the multiple causes of the opioid epidemic and income equality and reveal the limits of any one model to predict outcomes of a pandemic. Were we policymakers, we could fit some of these models to data to gauge effect sizes. We could then run experiments to help us guide policy choices based on what we have learned. Findings from experiments of limited duration and scope, though informative, should be interpreted through multiple models. For example, systemic effects or feedbacks may reverse or mitigate the effects found were the experiment to be expanded.

The core theme of this book, that we need many models, to reason, act, design, and so on, can be applied to any number of our present challenges.

Reversing trends in obesity, improving school performance, mitigating climate change, managing water resources, and improving international relations can all benefit from a many-model approach.

Adding even a single new model can have enormous consequences. Take for example the problem of predicting financial collapses. The United States Federal Reserve relies on traditional economic models using national accounting data on inflation, unemployment, and inventories. Those data suffer from lags. They are released weekly, quarterly, or annually. Those data also come from surveys, that is, samples of the entire economy.

Complexity scholar J. Doyne Farmer argues for creating a second class of models based on real-time data scraped from the web. These new models would rely on more granular, instantaneous data, and therefore differ from traditional models. Farmer argues that such models could prove much better than existing models. He may be right. Yet, these new models need not be more accurate to be of use in predicting and preventing financial disasters. Given the new models would use different data and rely on different assumptions, so long as they are not far less accurate, when combined with existing models, these new models would improve predictions. Policymakers, to use Farmer's turn of phrase, would become more collectively aware.^{[25](#)}

Business leaders and policymakers might engage in a similar exercise when making important decisions. They could apply multiple models informed by data to decide on product or policy attributes, time product or program launches, design compensation plans, construct supply chains, and forecast sales or impact. Each of these actions occurs within a complex system. No single model will suffice to excel at any of those tasks. Many models will be needed.

That logic applies when confronted with any choice—when asked to make a prediction, or when faced with a design challenge, we should take a many-model approach. Many-model thinking produces better performance than taking actions based on hunches and gut instincts. That said, we have no guarantee of success. Even with many models, we may not identify the most relevant logical chain. The domain of interest might be so complex

that even ensembles of models can only explain a small portion of the variation.

The same holds when applying models to aid in design; we may find ourselves unable to construct useful abstractions. The simplicity of models may, in those cases, be their undoing. In the face of complexity, it is possible that we find models not up to the tasks of communicating ideas, making accurate predictions, or pointing us toward the best actions.

Our explorations with models may often lead us down rabbit holes. We might not be able to identify a new policy or a behavioral intervention likely to improve the world. Nevertheless, even in those cases, we benefit from contemplating and applying models. And, we benefit all the more from applying many models as they will clarify our thinking by uncovering interdependencies.

Complex processes, such as those that result in epidemics and inequalities, often frustrate our attempts to understand, explain, or communicate. That is not a reason to give up on models and to rely on our intuition, but a reason to press on, to continue to develop new and better models.

As we learn to apply those models, we must maintain a degree of humility. Even when using many models, our abilities to reason have limits. By definition, complex phenomena are difficult to predict, explain, and understand. We must keep our minds open. We must continue to build new models and to improve upon existing ones. If a model leaves out key features of the world—such as social influences, positive feedbacks, or cognitive biases—then we should build other models that include those features. By doing so, we can begin to discern when those attributes matter and how much. The fact that all models are wrong should not take the wind out of our sails but instead motivate us to develop more models. With many models, we have the possibility of wisdom.

We should also seek joy in our modeling efforts. Throughout the book, much of the focus has been on pragmatic ends—to become better thinkers, to be more effective at work, and to operate as more informed citizens of the world. We have, in a few spots, taken notice of the beauty and elegance of models. When we see model building as an art as well as a science, we

begin to recognize the beauty of the exercise. We make the assumptions, write the rules, and then allow our minds to play within those rules. The laws of logic outline the boundaries and we explore within those limits. Through logical play, we improve ourselves and become wiser. May we take that wisdom out into the world and help to change it in positive ways.