The Review of Austrian Economics
https://doi.org/10.1007/s11138-023-00625-z

The use of algorithms in society

Cass R. Sunstein[1]

Accepted: 27 April 2023

Abstract

The judgments of human beings can be biased; they can also be noisy. Across a wide range of settings, use of algorithms is likely to improve accuracy, because algorithms will reduce both bias and noise. Indeed, algorithms can help identify the role of human biases; they might even identify biases that have not been named before. As compared to algorithms, for example, human judges, deciding whether to give bail to criminal defendants, show Current Offense Bias and Mugshot Bias; as compared to algorithms, human doctors, deciding whether to test people for heart attacks, show Current Symptom Bias and Demographic Bias. These are cases in which large data sets are able to associate certain inputs with specific outcomes. But in important cases, algorithms struggle to make accurate predictions, not because they are algorithms but because they do not have enough data to answer the question at hand. Those cases often, though not always, involve complex systems. (1) Algorithms might not be able to foresee the effects of social interactions, which can depend on a large number of random or serendipitous factors, and which can lead in unanticipated and unpredictable directions. (2) Algorithms might not be able to foresee the effects of context, timing, or mood. (3) Algorithms might not be able to identify people's preferences, which might be concealed or falsified, and which might be revealed at an unexpected time. (4) Algorithms might not be able to anticipate sudden or unprecedented leaps or shocks (a technological breakthrough, a successful terrorist attack, a pandemic, a black swan). (5) Algorithms might not have "local knowledge," or private information, which human beings might have. Predictions about romantic attraction, about the success of cultural products, and

Robert Walmsley University Professor, Harvard University. This is a revised text of a lecture given at King's College in March 2023. I am grateful to Mark Pennington for superb comments on a previous draft; to Daniel Kahneman and Olivier Sibony for instructive discussions of bias and noise; and to Jon Kleinberg and Sendhil Mullainathan for instructive discussions of the limits of algorithms and data. I am also grateful to the audience at King's College and to participants in a workshop at Harvard Law School for valuable help. None of the foregoing people is responsible for my errors.

Extended author information available on the last page of the article

Published online: 04 May 2023

about coming revolutions are cases in point. The limitations of algorithms are analogous to the limitations of planners, emphasized by Hayek in his famous critique of central planning. It is an unresolved question whether and to what extent some of the limitations of algorithms might be reduced or overcome over time, with more data or various improvements; calculations are improving in extraordinary ways, but some of the relevant challenges cannot be solved with ex ante calculations.

Keywords Algorithms · Cognitive bias · Local knowledge · Complexity · Hayek

JEL Codes: B31 · D80 · D81 · D83 · D90 · D91

1 Two claims

If doctors are unrealistically optimistic, their judgments will be wrong, and in a predictable direction. Optimistic bias produces *systematic error*. If doctors are too optimistic in the morning and too pessimistic in the afternoon, their judgments will be noisy, in the sense that they will show *unwanted variability*. Human beings, including doctors, often aim to solve prediction problems, where they may be biased, noisy, or both.

I offer two claims here. The first is that in important domains, algorithms can reduce or eliminate bias while also eliminating noise. In particular, algorithms can overcome the harmful effects of cognitive biases, which can have a strong hold on people whose job it is to avoid them, and whose training and experience might be expected to allow them to do so. Even more, *algorithms might help us to learn what biases are leading to human error*; they might even identify new or unnamed biases. And much of the time, algorithms are not noisy; they can be designed so as to give the same answer every time. In short, the first claim is that *algorithms can make better predictions than human beings do, because they are less biased and less noisy*.

My second claim is not in conflict with the first, but it is in a very different spirit. It is that no less than human beings, algorithms have great difficulty in solving (some) prediction problems. One clue is provided by the data in the very domains in which algorithms outperform human beings: Even when algorithms are superior, they are usually not spectacularly superior. They do better than human beings do across large populations, but they cannot say what will happen in individual cases.

Consider five challenges: (1) Algorithms might not be able to foresee the effects of social interactions, which can lead in directions that are exceedingly hard to predict ex ante. (Consider the question whether a song will become a big hit.) (2) Algorithms might not be able to foresee the potentially large effects of context, timing, serendipity, and mood. (Consider the question whether two people will fall in love.) (3) Algorithms might not be able to identify people's preferences, which might be concealed or falsified, and which might be revealed at an unexpected time. (Consider the question whether a social movement will arise in a specified month or year.) (4) Algorithms might not be able to anticipate change, including rapid change, which might be a product of unexpected shocks (a technological breakthrough, a successful terrorist attack, a pandemic, a black swan). (5) Algorithms might not have local

knowledge, or knowledge about what is currently happening or likely to happen on the ground. In all of these cases, the problem is not algorithms as such. It is a lack of necessary data.

I should confess that I have more confidence in my first claim, about the ability of algorithms to reduce bias and noise, than I do in my second claim, about the limits of algorithms in solving prediction problems. The five challenges are different from one another, and some might prove more tractable than others. Still, I am going to press the second claim as vigorously as I can.

My title is of course a play on Friedrich Hayek's great essay, *The Use of Knowledge in Society*.[1] Hayek drew attention not to the motivations of planners, but to what he saw as their inevitable lack of information. Hayek began: "*If* we possess all the relevant information, *if* we can start out from a given system of preferences, and *if* we command complete knowledge of available means, the problem which remains is purely one of logic." He emphasized that the "peculiar character of the problem of a rational economic order is determined precisely by the fact that the knowledge of the circumstances of which we must make use never exists in concentrated or integrated form but solely as the dispersed bits of incomplete and frequently contradictory knowledge which all the separate individuals possess." Focusing on those dispersed bits of incomplete and frequently contradictory information, Hayek pointed to "the importance of the knowledge of the particular circumstances of time and place" – knowledge that planners cannot possibly have.

Hayek also pointed to a separate problem: change. At Time 2, things might be very different from what they were at Time 1, and planners might struggle to understand that. What is true at Time 1 might not be true at Time 2. The knowledge that people have in markets shifts rapidly over time. As Hayek had it, the price system is a "marvel," because it can incorporate knowledge that is not only dispersed but also fleeting. In identifiable circumstances, I suggest, algorithms are akin to planners. In those circumstances, the limits of prediction cannot be overcome; they are built into the human condition[2] – now and (let us put it boldly) forever.

It should be clear that my two arguments bear directly on mounting debates over "techno-optimism" with respect to the knowledge problem.[3] Consistent with that form of optimism, I shall be emphasizing that broadly speaking, algorithms can do better than people do, at least for certain kinds of prediction problems (my first argument). Still, algorithms face serious knowledge problems too, which means that some such problems will be impossible to solve and perhaps even to dent (my second argument).

[1] Friedrich Hayek, *The Use of Knowledge in Society*, 35 Am. Econ. Rev 519 (1945).

[2] See Daniel Kahneman et al., Noise ch. 11 (2021).

[3] For a skeptical view, see Peter Boettke and Rosolino Candela, *On the Feasibility of Technosocialism*, 205 J. Economic Behavior & Organization 44 (2023).

2 Jail and Bail

Some of the oldest and most influential work in behavioral science shows that statistical prediction often outperforms clinical prediction; one reason involves cognitive biases on the part of clinicians, and another reason involves noise.[4] Algorithms can be seen as a modern form of statistical prediction, and if they avoid biases and noise, no one should be amazed. What I hope to add here is a concrete demonstration of this point in some important contexts, with some general remarks about both bias and noise.

Before we begin, we need to define the word "algorithm." According to a standard definition, an algorithm is "a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer." According to another, an algorithm is "a procedure used for solving a problem or performing a computation." Consider a procedure for deciding whether to drink alcohol: one drink every week, on Saturday night. Is that an algorithm? Consider a procedure for deciding whether to exercise: once a day, late in the afternoon. Is that an algorithm? Consider a procedure for deciding whether to exceed the speed limit: never. We can think of a rule, or a set of rules, as an algorithm, and a rule, or a set of rules, might greatly simplify decisions. In ordinary language, however, the term is usually reserved for computers, machine learning, and artificial intelligence, as in this account: "Algorithms act as an exact list of instructions that conduct specified actions step by step in either hardware- or software-based routines." I will be adopting that ordinary usage here.

Consider some research from Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, who explore judges' decision whether to release criminal defendants pending trial.[5] Their goal is to compare the performance of an algorithm with that of actual human judges, with particular emphasis on the solution to prediction problems. It should be obvious that the decision whether to release defendants has large consequences. If defendants are incarcerated, the long-term consequences can be very severe. Their lives can be ruined. But if defendants are released, they might flee the jurisdiction or commit crimes. People might be assaulted, raped, or killed. And while the decision whether to release criminal defendants pending trial is highly unusual in many ways, my goal here is to draw some general lessons, applicable to ordinary life, about the choice between decisions by human beings and decisions by algorithms.

In some jurisdictions in the United States, the decision whether to allow pretrial release turns on a single question: flight risk. It follows that judges have to solve a prediction problem: *what is the likelihood that a defendant will flee the jurisdiction*? In other jurisdictions, the likelihood of crime also matters, and it too presents a prediction problem: w*hat is the likelihood that a defendant will commit a crime*? (As it turns out, flight risk and crime are closely correlated, so that if one accurately predicts the first, one will accurately predict the second as well.) Kleinberg and his colleagues built an algorithm that uses, as inputs, the same data available to judges at the time of the bail hearing, such as prior criminal history and current offense. Their central

[4] *See* Paul E. Meehl, Clinical Versus Statistical Prediction (2013 ed.; originally published 1953).

[5] Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q.J. Econ. 237 (2017).

finding is that *along every dimension that matters, the algorithm does much better than real-world judges*. Among other things:

1. Use of the algorithm could maintain the same detention rate now produced by human judges and reduce crime by up to 24.7%. Alternatively, use of the algorithm could maintain the current level of crime reduction and reduce jail rates by as much as 41.9%. That means that if the algorithm were used instead of judges, thousands of crimes could be prevented without jailing even one additional person. Alternatively, thousands of people could be released, pending trial, without adding to the crime rate. It should be clear that use of the algorithm would allow any number of political choices about how to balance decreases in the crime rate against decreases in the detention rate.
2. A major mistake made by human judges is that they release many people identified by the algorithm as especially high-risk (meaning likely to flee or to commit crimes). More specifically, judges release 48.5% of the defendants judged by the algorithm to fall in the riskiest 1%. Those defendants fail to reappear in court 56.3% of the time. They are rearrested at a rate of 62.7%. Judges show leniency to a population that is likely to commit crimes.
3. Some judges are especially strict, in the sense that they are especially reluctant to allow bail—but their strictness is not limited to the riskiest defendants. If it were, the strictest judges could jail as many people as they now do, but with a 75.8% increase in reduction of crime. Alternatively, they could keep the current crime reduction, and jail only 48.2% as many people as they now do.

3 Two biases

Why does the algorithm outperform judges? The most general answer is that it is less biased, and it is not at all noisy. A more specific answer is suggested by point (3) above: judges do poorly with the highest-risk cases. (This point holds for the whole population of judges, not merely for those who are most strict.) The reason is an identifiable bias; call it *Current Offense Bias*.[6] Kleinberg and his colleagues restrict their analysis to two brief sentences, but those sentences have immense importance.[7] As it turns out, judges make two fundamental mistakes. *First*, they treat high-risk defendants as if they are low-risk *when their current charge is relatively minor* (for example, it may be a misdemeanor). *Second*, they treat low-risk people as if they are high-risk *when their current charge is especially serious*. The algorithm makes neither mistake. It gives the current charge something closer to its appropriate weight. It takes that charge in the context of other relevant features of the defendant's background, neither overweighting nor underweighting it. The fact that judges release a number of the high-risk defendants is attributable, in large part, to overweighting the current charge (above all, when it is not especially serious).

[6] *Id.* at 284.

[7] *Id.*

Intriguing and ingenious work by Ludwig and Mullainathan has suggested another reason that algorithms do better than human judges.[8] Even after controlling for race, skin color, and demographics, judges give more weight than do algorithms to the defendant's mugshot! As Ludwig and Mullainathan put it, "the mugshot predicts judge behavior: how the defendant looks correlates strongly with whether the judge chooses to jail them or not."[9] Perhaps unsurprisingly, judges are responsive to whether the mugshot shows the defendant as "well-groomed": judges are more likely to release defendants whose faces are clean and tidy as opposed to unkempt, dishevelled, and messy. Perhaps surprisingly, judges are more likely to release defendants whose mugshots show them as "heavy-faced" (with a wider or puffier face). Call it *Mugshot Bias*. We would not know that judges show Current Offense Bias, or Mugshot Bias, without the help of the algorithm.

4 Biased doctors

The bail study has a sibling, which involves doctors.[10] The central question has to do with diagnosis of heart attacks (or acute coronary episodes). Whom do doctors test for heart attacks, and when do they test them? Would an algorithm do better? In a close parallel to the bail study, it turns out that doctors test numerous people who should not be tested, and fail to test numerous people who should be tested. More specifically, doctors order a number of tests that are unlikely to find anything of interest—and thus waste a good deal of money. It also turns out that doctors do not test many patients that the algorithm rightly predicts will be "high-yield," in the sense that they have indeed had acute coronary episodes. The central results are precisely parallel to those in the bail study. Use of the algorithm could save a great deal of money (by reducing unnecessary and unhelpful tests), could prevent a number of deaths, or both.

Why do doctors err, compared to algorithms? As in the bail study, much of the answer lies in cognitive biases. Doctors give excessive weight to highly salient symptoms, such as chest pain, especially when those symptoms fit the stereotype of a heart attack. It is true, of course, that chest pains can be associated with heart attacks. The problem is that doctors give them more weight than they should. By contrast, the algorithm gives such symptoms something closer to the appropriate weight. Call it *Current Symptoms Bias*. Doctors also give undue weight to demographics; call it *Demographic Bias*. For example, doctors over-test older patients relative to their actual risks. If doctors had relied on algorithms in deciding whom to test, they could have avoided those biases, again saving money, lives, or both.

[8] Jens Ludwig & Sendhil Mullainathan, *Algorithmic Behavioral Science: Machine Learning as a Tool for Scientific Discovery* (Chicago Booth, Working Paper No. 22 – 15, 2022).

[9] *Id.* at 2 (emphasis omitted).

[10] Sendhil Mullainathan & Ziad Obermeyer, *Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care*, 137 Q.J. Econ. 679 (2022).

5 Biases and algorithms

When human beings suffer from a cognitive bias, a well-designed algorithm, attempting to solve a prediction problem, can do much better. Here is a simple illustration. When babies are born, the nurse or doctor might well give them an Apgar score, developed in 1952 by Virginia Apgar, an obstetric anesthesiologist. The evaluator measures the baby's color, heart rate, reflexes, muscle tone, and respiratory effort, sometimes described as a "backronym" for Apgar's name. The Apgar score refers to *appearance* (skin color), *pulse* (heart rate), *grimace* (reflexes), *activity* (muscle tone), and *respiration* (breathing rate and effort). In the Apgar test, each of these five measures is given a score of 0, 1, or 2. The highest possible total score is 10, which is rare. A score of 7 or above is considered indicative of good health. You can think of an Apgar score as a simple algorithm (in line with the definitions above), even though it does not involve a computer or artificial intelligence. And the Apgar test works; a central reason is that it greatly reduces the potential effects of biases in human judgment.

It is worth emphasizing that in both law and medicine, we are dealing not with novices, but with human beings who are both trained and experienced. They are experts. Nonetheless, they suffer from cognitive biases that produce severe and systematic errors. Current Offense Bias is best understood as a close cousin of *availability bias*: when we make judgments about probability, we often ask whether relevant examples are easily brought to mind.[11] In general, doctors are subject to availability bias;[12] for example, their decisions about whether to test patients for pulmonary embolism are affected by whether they have recently had a patient diagnosed with pulmonary embolism.[13] Mugshot Bias, Current Symptom Bias, and Demographic Bias are best understood as a form of *representativeness bias*: individual judgments about probability are frequently based on whether the known feature of a person or situation is representative of, or similar to, some unknown fact or condition.

Cognitive biases typically involve *attribute substitution*.[14] Availability bias is product of the availability heuristic, which people use to solve prediction problems. We substitute a relatively easy question ("does an example come to mind?") for a difficult one ("what is the statistical fact?"). Current Offense Bias reflects what we might call the Current Offense Heuristic, which also involves a relatively easy question ("how bad was the current offense?"), substituted for a harder one ("what is the flight risk?"). Representativeness bias is a product of the representativeness heuristic,

[11] *See* Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, *in* Judgment Under Uncertainty: Heuristics and Biases 3 (Daniel Kahneman et al. eds., 1982).

[12] *See* Ping Li et al., *Availability Bias Causes Misdiagnoses by Physicians: Direct Evidence from a Randomized Control Trial*, 59 Internal Med. 3141 (2020).

[13] *See* Dan P. Ly, *The Influence of the Availability Heuristic on Physicians in the Emergency Department*, 78 Analysis Emergency Med. 650 (2021); *see also* Carmen Fernández-Aguilar et al., *Use of Heuristics During the Clinical Decision Process from Family Care Physicians in Real Conditions*, 28 J. Evaluation Clinical Prac. 135 (2022).

[14] *See* Daniel Kahneman & Shane Frederick, *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*, *in* Heuristics and Biases: The Psychology of Intuitive Judgment 49, 49–81 (Thomas Gilovich et al. eds., 2002); Daniel Kahneman, Thinking, Fast and Slow (2011).

which people also use to solve prediction problems. They substitute a relatively easy question ("is the feature of the case representative of or similar to some fact?") for a difficult one ("what is the statistical fact?"). Apparently judges fall prey to the Mugshot Heuristic, and doctors use the Current Symptom Heuristic and the Demography Heuristic.

Because of the availability heuristic, people are likely to think that more words, on a random page, end with the letters "ing" than have "n" as their next to last letter[15]—even though a moment's reflection will show that this could not possibly be the case. Furthermore, "a class whose instances are easily retrieved will appear more numerous than a class of equal frequency whose instances are less retrievable."[16] Consider a simple study showing people a list of well-known people of both sexes, and asking them whether the list contains more names of women or more names of men. In lists in which the men were especially famous, people thought that there were more names of men, whereas in lists in which the women were the more famous, people thought that there were more names of women.[17]

This is a point about how *familiarity* can affect the availability of instances, and thus produce mistaken solutions to prediction problems. A risk that is familiar, like that associated with smoking, will be seen as more serious than a risk that is less familiar, like that associated with sunbathing. But *salience* is important as well. For example, "the impact of seeing a house burning on the subjective probability of such accidents is probably greater than the impact of reading about a fire in the local paper."[18] Current Symptom Bias reflects the power of salience. *Recency* matters as well. Because recent events tend to be more easily recalled, they will have a disproportionate effect on probability judgments. Availability bias thus helps account for "recency bias."[19] Current Offense Bias can be understood as a sibling to recency bias.

In many domains, availability bias and representativeness bias can lead to damaging and costly mistakes. Whether people will buy insurance for natural disasters is greatly affected by recent experiences.[20] If floods have not occurred in the immediate past, people who live on flood plains are far less likely to purchase insurance. In the aftermath of an earthquake, insurance for earthquakes rises sharply—but it declines steadily from that point, as vivid memories recede. Note that the use of the availability heuristic, in these contexts, is hardly irrational. Both insurance and precautionary measures can be expensive, and what has happened before seems, much of the time, to be the best available guide to what will happen again. The problem is that the availability heuristic can lead to serious errors, in terms of both excessive fear and neglect.

If the goal is to make accurate predictions, use of algorithms can be a great boon. For individuals, and for both private and public institutions (including governments all over the world), it can reduce or eliminate the effects of cognitive biases. Suppose

[15] *See* Tversky & Kahneman, *supra* note 11.

[16] *Id.* at 11.

[17] *Id.*

[18] *Id.*

[19] *See* Robert H. Ashton & Jane Kennedy, *Eliminating Recency with Self-Review: The Case of Auditors' 'Going Concern' Judgments*, 15 J. Behav. Decision Making 221 (2002).

[20] *See* Paul Slovic, The Perception of Risk 40 (2000).

that the question is whether to open an office in a new city; whether a project will be completed within six months; whether a particular intervention will help a patient who suffers from diabetes and cancer. In all of these cases, some kind of cognitive bias may well distort human decisions. There is a good chance that availability bias, representative bias, or one of their cousins will play a large role, and unrealistic optimism, embodied in the planning fallacy, may aggravate the problem. Algorithms have extraordinary promise. They can save both money and lives.

6 The best human judges (on local knowledge)

There is an important qualification, one with a close connection with Hayek's argument about the knowledge problem. We might easily imagine that in some contexts, algorithms *generally* perform better than human beings do – but also that in those very contexts, algorithms do not perform better than *all* human beings do. In other words, the best doctors might do better than algorithms, and the best human judges might do better than algorithms. What about the top 5% of human beings? Do they do better than algorithms do? If so, why?

Some important work suggests that while algorithms outperform 90% of human judges in the context of bail decisions, *the top 10% of judges outperform algorithms*.[21] The reason appears to be that the best judges have and use private information to make better decisions. They consider factors that algorithms do not. They appear to have something like local knowledge – an understanding of the defendant or the circumstances that algorithms lack. We could easily imagine a similar finding for doctors. It is possible that the best doctors know whom to test for heart disease, because they see something, or intuit something, that algorithms do not consider.

What might that something be? It would be extremely valuable to know. One possibility is that the local knowledge that they have is available to algorithms in principle, and that in the fullness of time, algorithms will be able to obtain it. Another possibility is that the relevant information is not knowable ex ante. It requires fine-grained understandings that can only be obtained on-the-spot. Perhaps those understandings are a product of interactions between judges and defendants.

Whatever the precise reason, there is a potentially large lesson here: Algorithms may lack information that human beings have, and for that reason, some human beings might be able to outperform algorithms. It may or may not be challenging to identify the best judges, with the relevant information, in advance. Market competition might be the best way to identify them.

In these circumstances, and in a Hayekian spirit, we might identify two kinds of actors. Some should rely on the algorithm, just as some should rely on the price. In both cases, it is unnecessary to know why an algorithm does well, or why a price as is at is. But some should see the algorithm's prediction and ask if they can do better, just as some people can see market prices and take them as imperfect signals to

[21] *See* Victoria Angelova et al., Algorithmic Recommendations and Human Discretion (Oct. 25, 2022; unpublished manuscript).

be challenged. The latter can produce a better signal. The challenge, of course, is to know in which category one falls.

In this connection, and to underscore that challenge, note that it is immensely important not to celebrate local knowledge as such; it might produce an unhelpful or misleading steer. Judicial judgments might be a product of distortions or bias. And indeed, *the low-performing judges are not using local or private knowledge for the better*. For example, they show an increased likelihood of detaining low-risk defendants *if they have recently heard a case in which another defendant, unrelated to the current one, committed a violent offense during release*. This patent overreaction appears to reflect a behavioral bias, closely akin to or perhaps a form of availability bias.

As I have noted, algorithms do better than people do, but they do not do spectacularly better. The impressive aggregate figures, in terms of welfare gains, come from the fact that very large populations are involved. If algorithms show a modest percentage increase in accuracy as compared with human beings, we might find seemingly major improvements. If an algorithm can produce a slight increase in the accuracy of screening for heart disease, we might see a significant reduction in deaths. But a slight increase in accuracy remains slight. I have said, for example, that formulas do better than clinicians. But in the median study, formulas are right 73% of the time, while clinicians are right 68% of the time.[22] That is not exactly amazing. Or turn to the heart attack study described above. People whom the algorithm placed in the middle of the risk distribution had had a heart attack 9.3% of the time, while people whom the algorithm placed in the highest decile had a heart attack 30% of the time. That is good, but it is very far from perfect. I will return to these points.

7 Noise

Recall that people are not merely biased; they are also noisy.[23] To see the difference between bias and noise, imagine two bathroom scales. The first scale is cruel: every day, it shows you as ten pounds heavier than you actually are. The second scale is capricious: on some days, it shows you as ten pounds heavier than you actually are; on other days, it shows you as ten pounds lighter than you actually are. The cruel scale is biased, in the sense that it is systematically wrong, and in a predictable direction. The capricious scale is noisy, in the sense that it shows unwanted variability. Note that the capricious scale is terrible even if it is right on average. On some days, it will give you unwelcome news, and on other days, it might delight you, but on all days, it is not telling you the truth.

Human judgment can be biased, noisy, or both. An obvious advantage of a good algorithm is that it can avoid bias. If you rely on it, you will not make a systematic error. A less obvious advantage of a good algorithm is that it can avoid noise. It can be designed so as to yield the same answer every time. It need not show unwanted variability. To be sure, a biased but noise-free algorithm is nothing to celebrate; it

[22] Daniel Kahneman et al., Noise: A Flaw in Human Judgment 143 (2021).

[23] *See* id.

will go systematically wrong in every case. But the elimination of noise is a great gain in itself.[24]

To see why noise can be a problem, return to the medical context. Suppose that doctors order a large number of tests in the morning, but that in the afternoon, they ask patients to go home and take aspirin. Or suppose that when doctors are in a good mood, they make very different decisions from those they make when they are grumpy. If so, doctors might not show a systemic bias of any kind. But they will be noisy, and noise will be responsible for plenty of mistakes. They will be like the capricious scale. For all of us, algorithms can eliminate the caprice. And indeed, judges are noisy when they are making bail decisions, and doctors are noisy when they are deciding whom to test for heart attacks. The noiselessness of the relevant algorithms, and not just their freedom from cognitive biases, helps account for their superiority over human beings.

The focus here is on individual decisions, where both bias and noise can be problems. But across institutions or systems, the problem can be even worse. A group might amplify the bias of individual members, ensuring that it is even more biased than its median member.[25] Systems are often noisy. In a hospital, patients might find themselves in a lottery: which doctor do they draw? One doctor might recommend a very different treatment from another. A large advantage of algorithms is that they can eliminate the lottery.[26]

8 Algorithm Aversion and Algorithm Appreciation

To say the least, many people do not love the idea of making decisions by algorithm. One reason appears to be a *general preference for agency*. Sometimes people decide to decide, because they like being the ones who decide.[27] Indeed, many people seem to want to retain agency even if they know that if they delegated the decision to another (including an algorithm), they would end up with better outcomes. A general lesson is that agency has intrinsic value, which means that people would demand a significant premium to give it up.[28] At the same time, it is reasonable to think that if people find it difficult or unpleasant to exercise agency, they will not want to do so, and they might even be willing to pay something to have access to a delegate, including an algorithm.[29] When might that be? Suppose that the decision involves highly technical issues. Or suppose that people are facing a high level of stress in their lives, or multiple tasks and burdens. If so, algorithm aversion might be converted into algorithm appreciation.

[24] For more details, see *id.*

[25] *See* Cass R. Sunstein & Reid Hastie, Wiser: Getting Beyond Groupthink to Make Groups Smarter (2014).

[26] *See* Kahneman et al., *supra* note 22.

[27] *See* Roy Shoval et al., *Choosing to Choose or Not*, 17 Judgment & Decision Making 768 (2022); Sebastian Bobadilla-Suarez et al., *The Intrinsic Value of Choice: The Propensity to Under-Delegate in the Face of Potential Gains and Losses*, 54 J. Risk & Uncertainty 187 (2017).

[28] *See* Bobadilla-Suarez et al., *supra* note 27.

[29] *See* Shoval et al., *supra* note 27.

Some evidence identifies a particular source of algorithm aversion: people are far more willing to forgive mistakes by human beings than to forgive mistakes by algorithms.[30] If your investment adviser makes a terrible mistake, and if you lose money as a result, you might well think something like, "nobody is perfect," or "to err is human." If, by contrast, an algorithm makes a mistake, and if you lose money as a result, you might lose faith in it. Hence a key empirical finding: *people are especially averse to algorithmic forecasters after seeing them err, even if they do better than human forecasters*.[31] In short, people are less forgiving of algorithms than they are of human beings. This evidence strongly supports a speculation, which is that when making their own decisions, people will not want to rely on algorithmic forecasters that make mistakes, and will prefer to decide themselves, even if they know that algorithmic forecasters are better than they are.

Is that rational? If people want to make the correct decision, it is not. If their goal is to make money or to improve their health, they should rely on the better decider. But one more time: if people *enjoy* making decisions, a preference for making one's own decisions might be perfectly rational. Perhaps people find the relevant decisions fun to make. Perhaps they like learning. Perhaps decision-making is a kind of game. Perhaps they like the feeling of responsibility. Perhaps they like the actuality of responsibility. If so, algorithm aversion is no mistake at all.

There is another factor. People have been found not to trust algorithms, and not to want to use them, in part because they do not know how they work.[32] Suppose that you learn that an algorithm can predict what jokes your best friend will find funny, and indeed that an algorithm can make better predictions, on that count, than you will. Will you consult the algorithm in deciding what jokes to tell your best friend? For many reasons, you might not. You might want to tell her *your* jokes, not those recommended by an algorithm, even if it is more accurate. But research finds that people are more likely to trust algorithms, and to be willing to rely on them, if they are given a simple account of why they work.[33]

In the context of jokes, for example, algorithms can make good predictions about what jokes Erika or Paul will find funny if they obtain some data about what jokes Erika or Paul have found funny in the past. The reason is that algorithms have a great deal of data about what jokes people find funny, and they can "match" the answers of Erika and Paul to the answers of numerous other people. Having done that, they predict that if Erika and Paul find certain jokes funny, they will find other jokes funny, because people who like the jokes that Erika and Paul find funny find those other jokes funny as well. Once people learn that algorithms work for that reason, they tend to trust algorithms much more.[34] We can imagine analogies in many contexts. You

[30] *See* Berkeley J. Dietvorst et al., *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, 144 J. Experimental Psych. 114 (2015).

[31] *Id.*

[32] *See* Michael Yeomans et al., *Making Sense of Recommendations*, 32 J. Behav. Decision Making 403 (2019).

[33] *Id.*

[34] *See id.*

might be willing to make decisions by algorithms once you are given more clarity about why their predictions are accurate.

Indeed, findings of algorithm aversion are complemented by findings of algorithm appreciation.[35] In many contexts, people seem to prefer decision by algorithms to decision by human beings. In estimating the weight of people in a photograph, for example, people were more likely to update their judgments in response to the assessment of an algorithm than in response to an assessment from a human being. People showed a similar preference for an algorithm in predicting the rank of a song on Billboard's "Hot 100" and in predicting whether someone would enjoy a date with a particular person. (There is an irony here, and we will get to it shortly.) People were also more likely to update in response to the advice of an algorithm in response to these questions:

- "What is the probability that Tesla Motors will deliver more than 80,000 battery-powered electric vehicles (BEVs) to customers in the calendar year 2016?"
- "What is the probability that a North American country, the EU, or an EU member state will impose sanctions on another country in response to a cyber attack or cyber espionage before the end of 2016?"
- "What is the probability that the United Kingdom will invoke Article 50 of the Lisbon Treaty before July 1, 2017?"

Interestingly, national security experts discounted the advice of the algorithm; in fact they discounted advice from all sources. This finding fits well with work attempting to reconcile algorithm aversion and algorithm appreciation, and finding that people are highly attentive to whether there is good reason to trust the algorithm or the human alternative.[36] If, for example, the human being is described as a "human expert" or a "physician," we might find algorithm aversion; if the human being is described as "another participant" or "a randomly chosen participant from a pool of 314 participants who took a past study," we might find algorithm appreciation. People seem to make rational, intuitive judgments about comparative expertise.

9 Simple and complex phenomena

Algorithms tend to do better than human beings in simple cases, in which the question is whether the presence of certain factors A, B, and C are likely to be associated with some outcome X or Y. Suppose that we are dealing with a medical question. What is the likelihood that women with certain characteristics have breast cancer, or that children with certain characteristics have asthma? We have seen enough to know that armed with sufficient data, algorithms are likely to be able to improve on human

[35] See Jennifer Logg et al., *Algorithm Appreciation: People Prefer Algorithmic to Human Judgment,* 51 Organizational Behavior and Human Decision Processes 90 (2019).

[36] See Yoyo Hou and Malte Yung, *Who Is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making*, 5 Proceedings of the ACM on Human-Computer Interaction 1 (2021).

judgment (with relevant qualifications about the best judges). But consider complex phenomena, in which the question is not whether and to what extent certain identifiable factors are associated with certain outcomes, but in which relevant factors (and perhaps a large number of such factors) are interacting with one another, and in which the nature of the interactions, and their consequences, depend on the concrete circumstances, which are difficult or impossible to anticipate in advance.

Here is a simple example of a complex phenomenon. Suppose that we are asking whether a group of workers is going to go on strike, when the decision of each worker is dependent on the decision of other workers, and when different workers have different thresholds for deciding whether to participate in a strike. For familiar reasons, it is easy to imagine multiple equilibria.[37] Even if we know a great deal about each individual worker, and even if an algorithm can have access to that knowledge, it might not be possible to predict the outcome. Everything might turn on who does what at the relevant time, which might depend on random factors, and on the existence and consequences of interactions among workers, which might also depend on random factors. An algorithm might be able to say something about the probability of a strike – say, it is more than 10% and under 90% – but two questions remain. (a) How helpful is a wide range of that kind? (b) When we are speaking of a single event, does it really make sense to speak of probabilities?

These comments raise many questions. In a paper published in 1964,[38] Hayek attempted to engage some of those questions, with reference to the theory of evolution. Hayek emphasizes that Darwin's theory points to a process or mechanism that need not have produced the same organisms that we observe on earth. The theory of evolution describes "a range of possibilities," one that is extremely wide. And even if we knew (as we do not) everything about (1) the mechanism of mutation, (2) the circumstances in which particular mutations would appear, and (3) the precise advantages that any mutation would confer, we still would not be able "to explain why the existing species or organisms have the particular structures which they possess, nor to predict what new forms will spring from them." This is a striking claim, and it is not intuitive.

Hayek explains that the reason for our ignorance is "the actual impossibility of ascertaining the particular circumstances which, in the course of two billion years, have decided the emergence of the existing forms, or even those which, during the next few hundred years, will determine the selection of the types which will survive." The number of relevant facts is simply too large. It is not possible to insert them into some formula that could then spit out some predictions.

In Hayek's account, complex social phenomena have the same characteristics. In the social domain, "individual events regularly depend on so many concrete circumstances that we shall never in fact be in a position to ascertain them all; and that in consequence the ideal of prediction and control must largely remain beyond our reach." To drive the point home, Hayek observes that "almost any event in the course of a man's life may have some effect on almost any of his future actions," which

[37] For one account, see Cass R. Sunstein, How Change Happens (2019).

[38] See Friedrich Hayek, *The Theory of Complex Phenomena: In Honor of Karl R. Popper*, in The Critical Approach to Science and Philosophy 332 – 59 (Mario Bunge ed. 1964).

"makes it impossible" for us to "translate our theoretical knowledge into predictions of specific events." Hayek acknowledges that the advances of science have produced a great deal of exuberance, but he is exuberant in his own way in offering this conclusion: "It is high time, however, that we take our ignorance more seriously."

Is this a prescient argument, fully applicable to an era of artificial intelligence and machine learning? I am going to argue that it is. But let us keep in mind two questions. The first is whether and when algorithms might get close to understanding, in advance, the likely concrete circumstances. The second is whether algorithms might be able to make probability judgments that have a sufficiently narrow band.

10 Life trajectories

In 2020, a large team of researchers – 112, to be exact -- engaged in an unusually ambitious project. They wanted to see if life trajectories could be predicted. To do that, they challenged the world. Their challenge had a simple name: The Fragile Families Challenge.[39]

The challenge began with an extraordinary data set, known as the Fragile Families and Child Wellbeing Study, which was specifically created in order to enable social science research. That study, which is ongoing, offers massive amounts of data about thousands of families, all with unmarried parents. Each of the mothers gave birth to a child in a large city in the United States around 2000. The data was collected in six "waves," at birth and at the ages of 1, 3, 5, 9, and 15. Each collection produced a great deal of information, involving child health and development, demographic characteristics, education, income, employment, relationships with extended kin, father-mother relationships, and much more. Some of the data was collected by asking a battery of questions to both the mother and the father. Some of it came from an in-home assessment (at ages 3, 5, and 9) that included measurements of height and weight, observations of the neighborhood and home, and various tests of vocabulary and reading comprehension. The Fragile Families Challenge was initially launched when data had been collected from the first five waves (from birth to the age of nine years), but when complete data from the sixth wave (year 15) were not yet available.

That was a terrific advantage, because it allowed the researchers to create the Challenge, which was to predict the following outcomes:

1) Child grade point average.
2) Child grit (determined by a self-reported measure that includes perseverance).
3) Household eviction.
4) Household material hardship.
5) Layoff of the primary caregiver.
6) Participation in job training by the primary caregiver.

[39] Matthew Salganik et al., *Measuring The Predictability of Life Outcomes With A Scientific Mass Collaboration*, 117 PNAS no. 15 (2020).

Those who took the challenge were given access to background material from the first five waves, and also to data on one-half of the families from the sixth wave. The material contained data on a total of 4,262 families, with a whopping 12,942 variables about each family. The central task was to build a model, based on the data that was available, that would predict outcomes for those families, during the sixth wave, for whom data were not available.

The researchers sought to recruit a large number of participants in the Fragile Families Challenge. They succeeded. In the end, they received 457 initial applications, which were winnowed down to 160 teams. Many of the teams used state-of-the-art machine-learning methods, explicitly designed to increase accuracy. The central question was simple: Which of the 160 teams would make good predictions?

The answer is: None of them. True, the machine-learning algorithms were better than random; they were not horrible. But they were not a lot better than random, and for single-event outcomes – such as whether the primary caregiver had been laid off or had been in job training – they were only *slightly* better than random. The researchers conclude that "low predictive accuracy cannot easily be attributed to the limitations of any particular researcher or approach; hundreds of researchers attempted the task, and none could predict accurately."

Notwithstanding their diverse methods, the 160 teams produced predictions that were pretty close to one another – and not so good. As the researchers put it, "the submissions were much better at predicting each other than at predicting the truth." A reasonable lesson is that we really do not understand the relationship between where families are in one year and where they will be a few years hence. Seeming to draw that lesson, the authors of the Fragile Families Challenge suggest that their results "raise questions about the absolute level of predictive performance that is possible for some life outcomes, even with a rich data set." You can learn a great deal about where someone now is in life, and still, you might not be able to say very much at all about specific outcomes in the future.

Here is a way to understand that point. Take a girl who is ten years old and learn everything you can about her: her family, her demographics, her neighborhood, her schooling, her sports. Now predict various things about her life at the age of twenty-one. Do you have much confidence in your prediction? You shouldn't. The number of variables that can move a life in one direction or another is very high, and it is not possible to foresee them in advance. Someone might break a leg at a crucial moment, meet an amazing music teacher, find a new friend, hear a song on the radio on Sunday morning, or see something online or on the news that changes everything.

11 Love and romance

Can algorithms predict whether you will fall in love with a stranger? Can they actually help people to find romantic partners? Thus far, the results on such counts are not promising. Samantha Joel and colleagues find that algorithms struggle to predict "the compatibility elements of human mating. . before two people meet," even if one has a very large number of "self-report measures about traits and preferences that

past researchers have identified as being relevant to mate selection."[40] Joel and her colleagues suggest that romantic attraction may well be less like a chemical reaction with predictable elements than "like an earthquake, such that the dynamic and chaos-like processes that cause its occurrence require considerable additional scientific inquiry before prediction is realistic."

What are "dynamic and chaos-like processes"? It is worth pondering exactly what this means. Most modestly, it might mean that algorithms need far more data in order to make accurate predictions – far more, at least, than is provided by self-report measures about traits and preferences. Such measures might tell us far too little about whether one person will be attracted to another. Perhaps we need more data about the relevant people, and perhaps we should focus on something other than such measures. It is possible that algorithms cannot make good predictions if they learn (for example) that Jane is an extrovert and that she likes football and Chinese food. It is possible that algorithms would do better if they learn that Jane fell for John, who had certain characteristics that drew her to him, and also for Tom and Frank, who had the same characteristics. If so, perhaps she is most unlikely to fall for Fred, who has none of those characteristics, but quite likely to fall for Eric, who shares those characteristics with John, Tom, and Frank.

On this view, the right way to predict romantic attraction is to say, "if you like X and Y and Z, you will also like A and B, but not C and D." Or perhaps we should ask whether people who are like Jane, in the relevant respects, are also drawn to Eric -- an approach that is not unrelated to that described above in connection with humor. Of course it would be necessary to identify the relevant respects in which people are like Jane, and that might be exceedingly challenging.

More radically, we might read the findings by Joel and her colleagues to suggest that romantic attraction is not predictable by algorithms for a different reason: It depends on so many diverse factors, and so many features of the particular context and the particular moment, that algorithms will not be able to do very well in specifying the probability that Jane will fall for Eric. The reference to "dynamic and chaos-like processes" might be a shorthand way of capturing mood, weather, location, time of day, and an assortment of other factors that help produce a sense of romantic connection or its absence. Jane might smile at a certain moment at lunch, and Eric's heart might flutter, or Jane might not smile at that moment, because she is distracted by something that happened in the morning, Eric might say something witty as sandwiches come to the table, because of something he read in the paper that morning, and that might initiate a chain of events that culminates in marriage and children. For romance, ultimate outcomes may depend on factors that cannot be identified in advance. This is the sense in which algorithms are sometimes like centralized planners: They do not have relevant information about time and place. (Again, there does not seem to be anything like the price system to replace them with.)

We do have to be careful here. An algorithm might be able to say that there is essentially no chance that Jane will like Carl, because there are things about Carl that we know, in advance, to be deal-breakers for Jane. Jane might not be drawn to

[40] *See* Samantha Joel et al., *Is Romantic Desire Predicable? Machine Learning Applied to Initial Romantic Attraction*, 28 Psych. Science 1478 (2017).

short men or to tall men; she might not be attracted to much older men or to much younger men; she might not be attracted to men. An algorithm might be able to say that there is some chance that Jane will like Bruce; there is nothing about Bruce that is a deal-breaker for her, and there are some clear positives for her. Perhaps an algorithm can specify a range of probability for Jane and Bruce; perhaps the probability of a romantic connection (suitably defined) is more than 10% but less than 70%. So too, an algorithm might be able to say that Eric is within the category of "it might well happen" for Jane, because Eric is in some sense "her type." Perhaps an algorithm can specify a range of probability for Jane and Eric; perhaps the probability of a romantic connection (suitably defined) is more than 30% but less than 80%. The real question is whether and to what extent algorithms will eventually be able to do much better than that. We might speculate that the importance of particular factors – the concrete circumstances – is such that there are real limits on their predictive power (even if they might be able to outperform human beings, whose own predictive power is sharply limited in this context).

The topic of romantic attraction is intriguing in itself, and it can be seen as overlapping with an assortment of other prediction problems: whether you will enjoy living in Paris; whether you will become friends with a coworker; whether you will like a new job; whether a pandemic will occur in the next five years; whether a recession will occur in the next six months; whether a new movie will be make a specified amount of money; whether a new book will hit the bestseller list; whether there will be a revolution in a specific nation by a date certain. It is generally agreed that in stable environments with fixed rules, algorithms, armed with a great deal of data, are able to make pretty good predictions. But if the future is unlikely to be like the past, there is a real question whether, where, and when algorithms will do well, or even outperform human beings.[41] One problem might be the sheer number of possible events, not knowable in advance, that might produce one or another outcome; this is why the case of romantic attraction has general lessons.[42] Another problem might be an external shock or unexpected event, which might turn everything around (a technological innovation, a terrorist attack, a pandemic). We are speaking here of the essentially unpredictable nature of many events, because of the role of randomness.

12 Revolutions

In work that predated the rise of algorithms, the economist Timur Kuran urged that revolutions are unpredictable by their very nature.[43] Kuran argued that an underlying problem lies in "preference falsification": People do not disclose their preferences, which means that we cannot know whether they will, in fact, be receptive to a revolutionary movement. If we do not know what people's preferences are, we will not know whether they might be willing to participate in a rebellion once the circumstances become propitious. Kuran added that we cannot observe people's *thresholds*

[41] *See* Gerd Gigerenzer, How to Stay Smart in a Smart World (2022).

[42] See Kahneman et al., supra note 22.

[43] *See* Timur Kuran, Private Truths, Public Lies (1995).

for joining such a movement. How many people would be willing to join when a movement is at its early stages? Who will require something like strong minority support before joining it? Kuran also noted that social interactions are critical, and they too cannot be anticipated in advance. For a revolution to occur, people must see other people saying and doing certain things at certain times. How can we know, before the fact, who will see whom, and when, and doing what? The answer might well be that we cannot possibly do that.

Kuran was not writing about algorithms, but they are unlikely to be able to do that, either. Algorithms will find it challenging or impossible to learn what people's preferences are, and they might not be able to learn about thresholds. Even if they could do both, they would not (to say the least) have an easy time obtaining the data that would enable them to predict social interactions, and they might not even be able to identify their probability. In some ways, the challenge of predicting a revolution is not so different from the challenge of predicting a romantic spark.

Kuran did not deny that we might be able to learn something about (1) when a revolution is improbable in the extreme and also (2) when a revolution is at least possible. For one thing, we might be able to make at least some progress in identifying private preferences – for example, by helping people feel safe to say that they dislike the status quo, perhaps by showing sympathy with the view that the status quo is bad, or perhaps by guaranteeing anonymity. Algorithms might be able to help on that count. Kuran wrote before the emergence of social media platforms, which give us unprecedented opportunities to observe hitherto unobservable preferences (for example, via google searches, which might reveal widespread dissatisfaction with the current government). Perhaps algorithms can say something about probabilities, based on data of this kind. But if Kuran is right, they will not be able to say a lot, because their knowledge of preferences and thresholds will be limited, and because they will not be able to foresee social interactions. The general analysis should not be limited to revolutions. Preference falsification, diverse thresholds, and social interactions – one or more of these are in play in many domains.

13 Hits

Consider the question whether books, movies, or musical albums are likely to succeed. Of course we might know that a new album by Taylor Swift is likely to do well, and that a new album by a singer who is both terrible and unknown is likely to fail. But across a wide range, a great deal depends on serendipity, and on who says or does what exactly when.

This point clearly emerges from research from a number of years ago, when Matthew Salganik, Duncan Watts, and Peter Dodds investigated the sources of cultural success and failure.[44] Their starting point was that those who sell books, movies, television shows, and songs often have a great deal of trouble predicting what will succeed. Even experts make serious mistakes. Some products are far more successful

[44] See Matthew Salganik et al., *Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market*, 311 Science 854 (2006).

than anticipated, whereas some are far less so. This seems to suggest, very simply, that those that succeed must be far better than those that do not. But if they are so much better, why are predictions so difficult?

To explore the sources of cultural success and failure, Salganik and his coauthors created an artificial music market on a preexisting website. The site offered people an opportunity to hear forty-eight real but unknown songs by real but unknown bands. One song, for example, by a band called Calefaction, was 'Trapped in an Orange Peel.' Another, by Hydraulic Sandwich, was 'Separation Anxiety.' The experimenters randomly sorted half of about 14,000 site visitors into an 'independent judgment' group, in which they were invited to listen to brief excerpts, to rate songs, and to decide whether to download them. From those 7,000 visitors, Salganik and his coauthors could obtain a clear sense of what people liked best. The other 7,000 visitors were sorted into a 'social influence' group, which was exactly the same except in just one respect: the social influence group could see how many times each song had been downloaded by other participants.

Those in the social influence group were also randomly assigned to one of eight subgroups, in which they could see only the number of downloads in their own subgroup. In those different subgroups, it was inevitable that different songs would attract different initial numbers of downloads as a result of serendipitous or random factors. For example, 'Trapped in an Orange Peel' might attract strong support from the first listeners in one subgroup, whereas it might attract no such support in another. 'Separation Anxiety' might be unpopular in its first hours in one subgroup but attract a great deal of favorable attention in another.

The research questions were simple: would the initial numbers affect where songs would end up in terms of total number of downloads? Would the initial numbers affect the ultimate rankings of the forty-eight songs? Would the eight subgroups differ in those rankings? You might hypothesize that after a period, quality would always prevail – that in this relatively simple setting, where various extraneous factors (such as reviews) were highly unlikely to be at work, the popularity of the songs, as measured by their download rankings, would be roughly the same in the independent group and in all eight of the social influence groups. (Recall that for purposes of the experiment, quality is being measured solely by reference to what happened within the control group.)

It is a tempting hypothesis, but that is not at all what happened. 'Trapped in an Orange Peel' could be a major hit or a miserable flop, depending on whether a lot of other people initially downloaded it and were seen to have done so. To a significant degree, everything turned on initial popularity. Almost any song could end up popular or not, depending on whether or not the first visitors liked it. Importantly, there is one qualification: the songs that did the very best in the independent judgment group rarely did very badly, and the songs that did the very worst in the independent judgment group rarely did spectacularly well. But otherwise, almost anything could happen. The apparent lesson is that success and failure are exceedingly hard to predict, whether the prediction is being attempted by algorithms or human beings. There are many reasons. Here is one: it is difficult to know, in advance, whether a cultural product will benefit from the equivalent of early downloads.

Early popularity might be crucial, and early popularity can turn on luck. Because of the sheer number of variables that can produce success or failure, algorithms might well struggle to make successful predictions at early stages (though they can do better if they are given data on an ongoing basis). And in the case of financial markets, there is a special problem: Once it is made, a prediction by a terrific algorithm will automatically be priced into the market, which will immediately make that prediction less reliable, and possibly not reliable at all.

Consider, most broadly, these remarks from Keynes[45]:

By "uncertain" knowledge, let me explain, I do not mean merely to distinguish what is known for certain from what is only probable. The game of roulette is not subject, in this sense, to uncertainty; nor is the prospect of a Victory bond being drawn. Or, again, the expectation of life is only slightly uncertain. Even the weather is only moderately uncertain. The sense in which I am using the term is that in which the prospect of a European war is uncertain, or the price of copper and the rate of interest twenty years hence, or the obsolescence of a new invention, or the position of private wealth-owners in the social system in 1970. About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know.

Keynes is pointing to cases in which we cannot assign probabilities to outcomes – cases of uncertainty rather than risk. He acknowledged that people have strategies for handling such situations. For example: "We assume that the present is a much more serviceable guide to the future than a candid examination of past experience would show it to have been hitherto. In other words, we largely ignore the prospect of future changes about the actual character of which we know nothing." Keynes did not mean to celebrate those strategies. He thought that they were ridiculous. "All these pretty, polite techniques, made for a well-panelled Board Room and a nicely regulated market, are liable to collapse," because "we know very little about the future." If "we" do not know, because we lack relevant data, perhaps algorithms cannot know, either.

14 Back to the future

I have made two claims here. The first is that in many domains, algorithms outperform human beings, because they reduce or eliminate both bias and noise. As Current Offense Bias and Mugshot Bias make clear, experienced judges (in the literal sense) can do significantly worse than algorithms. The same is true of Current Symptom Bias and Demographic Bias. It will be noticed that the four biases did not even have names in advance of the relevant research; algorithms can help not only to counteract human biases but also to identify them.

At the same time, there are some prediction problems on which algorithms will not do well; the reason lies in an absence of adequate data, and in a sense in what we might see as the intrinsic unpredictability of human affairs. (1) Algorithms might not be able to foresee the effects of social interactions, which can lead in all sorts of unanticipated directions. (2) Algorithms might not be able to foresee the effects of context,

[45] John Maynard Keynes, The General Theory of Employment, Interest and Money 113–14 (1936).

timing, serendipity, or mood (as in the case of romantic attraction or friendship). (3) Algorithms might not have local knowledge about relevant particulars, or knowledge about what is currently happening or likely to happen on the ground. (4) Algorithms might not be able to identify people's preferences, which might be concealed or falsified, but which might be revealed at an unexpected time (perhaps because of a kind of social permission slip, which is itself hard to anticipate). (5) Algorithms might not be able to anticipate breakthroughs or shocks (a technological discovery, a successful terrorist attack, a pandemic).

These are disparate challenges, but all of them are closely connected to the knowledge problem, and in particular to Hayek's claims about the dispersed nature of knowledge in society, the importance of local knowledge, and the difficulty of making predictions when we are dealing with complex phenomena. For current purposes, his claims about complex phenomena deserve particular attention. They help to explain some of the difficulties that algorithms face; consider the Fragile Families Challenge in particular.

In some cases (category (3) is the obvious example), some human beings might be able to do better than algorithms can do, because they have knowledge of those particulars. In other cases (category (4) is the most obvious example, and category (3) might turn out to be an example as well), algorithms should be able to make progress over time. But in important cases (defined above all by category (1)), we are dealing with complex phenomena, and the real problem is that the relevant data are simply not available in advance, which is why accurate predictions are not possible – not now, and not in the future, either.

Funding No funding to disclose.

Authors and Affiliations

Cass R. Sunstein[1]

✉ Cass R. Sunstein
csunstei@law.harvard.edu

[1] Robert Walmsley University Professor, Harvard Law School, 1585 Massachusetts Avenue, 02138 Cambridge, MA, United States