www.joim.com

Journal Of Investment Management, Vol. 20, No. 4, (2022), pp. 87–101

INSIGHTS

“Insights” features the thoughts and views of the top authorities from academia and the profession. This section offers unique perspectives from the leading minds in investment management.

BIAS AND NOISE IN HUMANS & AI: WHEN TO TRUST HUMANS & MACHINES IN DECISION-MAKING

Vasant Dhar[a]

When should we trust machine-based and human decisions in finance? In this article I answer this question by drawing on two sets of insights about decision error. I first draw on research of leading theorists on human decision-making and prediction, summarized through a set of articles and conversations with them about the two sources of decision error, namely, bias and noise. I also draw on two decades of experience operating a machine-learning based trading platform, where algorithmic bias and noise also manifest themselves, but very differently than in human decision-making. This two-pronged analysis of the properties of humans and algorithmic decision-making provides a backdrop against which the challenges and opportunities for creating trustable decision-making systems in finance come into sharp focus.

1 Introduction

In 2016, I published an article in the Harvard Business Review titled “*When To Trust Robots with Decisions and When Not To*” (Dhar, 2016). In that article, I departed from much of the previous literature on automation, which has historically viewed *programmability*[1] as a proxy for automation. Instead, I tried to demonstrate why programmability is actually, the beginning, not the terminal point in considering the application of machine-based algorithms.

In financial domains especially, it turns out that errors and uncertainty play a very significant role in trust. Simply put, the higher the uncertainty associated with the magnitude and impact of mistakes, the less we tend to trust a system.

Since my original article on the subject, I have had the chance to continue to explore the impact of uncertainty on humans and machines

[a] Professor, Stern School of Business & Center for Data Science, New York University, USA.

in general, and on prediction, specifically. This exploration includes conversations with some leading thinkers on decision-making on my *Brave New World* podcast, including *Daniel Kahneman*, *Philip Tetlock*, *Terry Odean* and *Aswath Damodaran*.

Although my podcast deals with a broad range of issues with guests from diverse domains, the subject of how humans deal with uncertainty in their judgments and decisions has come up in a surprising number of conversations. These discussions have led me to a deeper understanding of the sources and nature of errors in human and machine-based decision-making, and more broadly, about the properties of models used by humans and machines for prediction.

Artificial Intelligence has also progressed rapidly in the six years since I first published the 2016 article. Machine learning has seen significant progress in the area of perception, particularly in vision and language. For example, the first release of GPT, the game-changing large-language model, was in 2018. GPT displays an impressive level of language understanding—at a level that might have been unimaginable in 2016. Even though it fails spectacularly in some settings, its intelligence is palpable, causing some researchers to consider whether it is displaying initial signs of sentience. Vision systems have also become incredibly powerful, aided by massive amounts of sensor and video data, and increased computing horsepower that enables us to build very complex and accurate models for applications such as facial recognition and self-driving vehicles.

To what extent should we expect these advances to translate into finance applications? How can systems based on similar technologies augment or replace human decision-making?

2 The Nature of Errors in Finance

The book, *Noise* (Kahneman *et al.*, 2021), systematically examines errors in decision-making, starting from the human perspective. The authors argue that, at the most basic level, individuals tend to ignore their own high degree of "objective ignorance" about problems—the things that make an outcome inherently highly unpredictable.

Consider an extreme case, where the underlying nature of a process is very noisy: predicting the outcome of a coin flip. It is difficult to predict an outcome with better "accuracy" than a random guess. Should we really consider mismatches between predictions and actual outcomes prediction errors? Or, are these mismatches simply an artifact of the "ground truth" itself being very shaky? Kahneman explains:

> *An error prediction is not necessarily a mistake, because many events that we try to predict or forecast are simply not completely predictable, like the stock market. I use the term objective ignorance for exactly that, where there is an objective limit to what we can accomplish. And, we'd better be aware of it, because we tend to blame people for having failed to forecast events. We blame them in hindsight, because in hindsight, we understand the events. We have causal interpretations after the fact that make the outcome seem obvious. But in fact, because of objective ignorance, they could not have predicted the event. It gives us the sense that we are in in touch with reality, that we understand the world. And that is to a significant degree an illusion.* (Brave New World, Episode 21.)

In other words, the ground truth in domains is characterized by levels of noise, or objective ignorance, is itself not entirely reliable. This type of inherent "aleatory" uncertainty (Hüllermeier *et al.*, 2021) makes it difficult for humans or machines to learn patterns and relationships by observing the history of such processes. Financial prediction, such as in capital markets, is a domain with high objective ignorance. While one might argue that more and better data will make finance more predictable, much like vision and language, this will be difficult for several reasons.

Objective ignorance is common to finance problems, and doesn't always get better with more data. One reason is non-stationarity: there's typically no stable fixed point, and you rarely have enough date (Israel *et al.*, 2020). Secondly, major shocks can throw off entire valuations overnight, as happened during COVID. Being on the wrong side of the shock can be a purely random thing. Finally, finance is also adversarial, not cooperative, where "acting" on a decision invariably entails some sort of friction or cost. Adversaries can also exploit your behavior. For all these reasons, prediction problems in Finance fall towards the higher end of objective ignorance. This makes it challenging, both for humans and algorithms.

In my original paper I referred to objective ignorance as "predictability," which varies between zero and one, as can objective ignorance. Table 1 shows a "heatmap" with predictability on the X axis. The other axis, cost of error, grew out of the observation that for the automation of decision-making, the cost of errors can matter much more than their frequency, and that there is a relationship between the two that determines our level of trust in a decision-making algorithm.

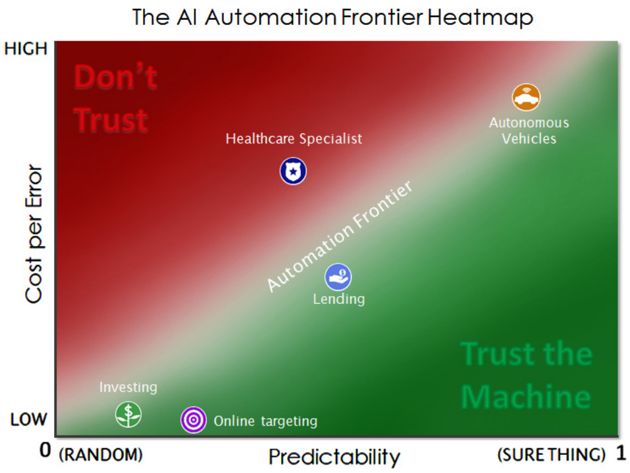

Figure 1

Specifically, our trust level is highest on the lower right corner and lowest on the upper left. The automation frontier represents a combination of the two factors when automation becomes possible.

I shall return to this figure to discuss the error properties of machine learning based models in Section 4. But first let us consider the sources of error in human decision-making.

3 Manifestations of Error in Humans: Noise and Bias

Kahneman *et al.* (2021) delineate two components of human error which they term *noise* and *bias*. They make the case that bias receives the bulk of our attention, but noise is usually the much bigger culprit.

What do the authors mean by "noise?" Fundamentally, they consider noise to be ***undesirable*** variability in decision-making. Sometimes variability is desirable, for example, when we desire multiple perspectives on something. But when we make ***opposite decisions*** based on the same data, that is undesirable variance.

Noise can arise at the individual level as well as across individuals. The latter arises due to differing bias levels of individuals, such as different levels of leniency among judges, which lead to different decisions for the same datum. Kahneman *et al.* call this "level noise." It measures the variability in the average level of different judges' judgements.

A second source of noise, that arises within individuals, is called "pattern noise." It reflects deviations from their typical pattern of decision-making that arise when they place a heavier weight than usual on some piece of data in a case that runs counter to their overall decision-making pattern. For example, a very lenient judge might be extremely severe towards repeat offenders, whereas another lenient judge might be particularly harsh towards those who prey on the elderly. The same is true for portfolio managers or insurance underwriters when they deviate from their general pattern of decision-making. This kind of deviation from their pattern of decision-making adds uncertainty, exhibited as noise, to the decision outcome.

Finally, there is "occasion noise" within individuals, where the decision is impacted by the environment, which should have had no impact on the decision. For example, extraneous factors such as the weather or the outcome of a football game appear to reliably alter *judgments* and sentencing, which adds noise to the process (Eren and Mocan, 2018). Portfolio managers are not immune to such phenomena either.

We should not be surprised by noise in human judgment. After all, that's what judgment is all about. We would expect judges to be impacted by specific factors about a case. But how much noise is "reasonable?" As Kahneman points out, it is often worse than we imagine:

> *Even if you take the crime for which the average sentence is seven years in prison, that difference between two randomly selected judges is going to be four years. So that's the average difference. If you pick two judges at random, 50% of the time, the difference between them will be bigger than four years. This is really shocking. That's noise.*

While we should expect inherent differences in the bias levels among judges, shouldn't we expect them to ***rank*** cases similarly, regardless of the sentences they impose?

It turns out that they do not. Kahneman argues that this happens because we tend to see the world differently, and have a hard time understanding why others don't see it the same way:

> *The main source of noise is that people look at the world and they see the world differently. They see every case different from the way others see it. And that would mean that if you took the judges, and gave them, say, 15 different crimes, and you ask them to rank the crimes, from the most severe to the least severe, they would not agree on their ranking, it's not only that one of them would set higher sentences than the other, the ranking would be different. And that is it turns out the biggest source of noise, and it's the most mysterious. We know that people are different from each other. But we're not quite prepared to recognize the extent to which people see the same situation differently from one another.* (Brave New World, Episode 21.)

Terry Odean and other behavioral finance researchers have built on the work of Kahneman and Tversky (1979), studying the various kinds of bias that arise in financial decision-making. Barber and Odean (2005) find that when building and managing their personal investment portfolios, most individuals limit their *attention* to a handful of stocks from a very large universe of candidates. Prior to the widespread use of the internet, individuals became aware of markets through weekly shows or publications on investing. In the Internet era, so-called "influencers" create attention-directing news through their comments and posts in discussion forums like Reddit. Tweets and comments by celebrities like Elon Musk also appear to have a large impact on peoples' attention. These dynamics produce a large class of investors who are net *buyers of attention-grabbing stocks*. Data from more recent trading platforms such as Robin Hood confirm this view. For example, on Robin Hood, a trading platform that targets retail investors, 35% *of the buying activity during a typical day is concentrated in 10 stocks* (Barber *et al.*, 2020). Interestingly, these stocks tend to underperform the market over the next month.

How about selling decisions? How do retail investors decide whether to sell positions or hold onto them? Odean observes several kinds of bias in this process that tend to impact performance negatively. A common observation is the *disposition effect*, the tendency to sell winners early and hold onto losers for too long. One explanation of this effect is individuals' natural tendency to avoid pain: it feels bad selling at a loss, whereas it feels good to sell at a profit. This tendency has been well documented, which Tversky and Kahneman explain in terms of *Prospect Theory* (Kahneman and Tversky, 1979). Odean notes:

> *You like to postpone feeling bad. And you tell yourself, it's coming back. It's just a paper loss, you know, it's going to come back. It just feels better.* (Brave New World, Episode 23)

People also tend to have other dispositions, such as trading more frequently than they should:

> *We don't know exactly how much people* **should** *trade. One measure would be if you sell one stock and buy another, on average, you'd like the one you buy to outperform the one you sold by enough to cover the transactions costs. I expected to find that the stocks people bought would do slightly better than the ones they sold. But it wouldn't be enough to cover the transactions costs, because that's sort of what my theory said. To my surprise, I found that on average, the stocks they bought went on to do worse than the ones they had just sold. And that was before factoring in the transaction costs. So, they were definitely trading too much. I teamed up with Brad Barber and was able to get a second dataset from the same brokerage firm. We followed up with a few more papers about overtrading and overconfidence in a paper called '*Trading is hazardous to your wealth,*" which showed that people who traded more did worse. (Barber and Odean, 2000)*

Does all of this evidence auger poorly for human investors? Should market participants just accept the efficient market hypothesis, and with it the notion that humans cannot consistently achieve better performance than the market overall? This is the predominant view in Finance despite the discoveries of some "anomalies" that contradict the hypothesis.

Over 20 years of trading algorithmic systems based on machine-learning, I have for the most part come to subscribe to this view as well. The evidence suggests that predictability can exist in the short-term, and decays rapidly. As a result, I tend to favor algorithms over humans when it comes to short-term prediction. But my recent interactions with both Philip Tetlock, about his research on "*superforecasting*" (Tetlock, 2015), and with Aswath Damodaran's emphasis on credible narratives to explain the numbers (Damodaran, 2017), suggest that there exists an element of skill in long term prediction. While the overwhelming bulk of the theory and empirical evidence suggests that that few professionals, if any, achieve better risk adjusted performance than a benchmark index like the S&P500, a few humans behave in ways that suggest that they might possess superior forecasting skills. The question is, what are these skills and can they be learned?

Tetlock's research on superforecasting focuses on the *prediction* of long-term phenomena like climate change, political outcomes, and economic shocks, where objective ignorance is inherently high. His main findings, which he reports in his book, Superforecasting (Tetlock, 2015), are based on the results from several forecasting tournaments, in which participants are incented to make earnest predictions about phenomena that are extremely challenging to forecast.

Tetlock's research funds that a select few humans consistently predict better than the vast majority of us. Tetlock notes that such individuals make accurate predictions over relatively long horizons of between one and three years, and occasionally, as far as five years into the future. How do they do it?

Tetlock attributes better prediction to *more eclectic self-critical cognitive styles*. In his words, these characteristics, which might be regarded as "necessary" conditions to becoming a superforecaster, break down into demonstrating the following attributes and behaviors:

- *They are reasonably intelligent*, *open minded*, *numerous*, *with intuitive appreciation for the* *rules of probability. If you're violating basic axioms of probability, there will be logical inconsistencies that will make it impossible to be all that accurate.*

- *They commit unnatural cognitive acts. There are two categories of unnatural cognitive acts. One of one of them is using base rates, and the other is an unusual amount of tolerance for cognitive dissonance and arguments and counter arguments. They're more likely to say, however, than moreover, so they have a higher "however to moreover ratio".* (*Brave New World, Episode 31.*)

My interpretation of Tetlock's results is in terms of something Kahneman has termed "System 1 and System 2 thinking," which he described in "*Thinking Fast and Slow*" (Kahneman, 2011). *System 1 thinking* is fast and intuitive. *System 2 thinking* is slow and deliberative.

It seems likely that Tetlock' superforecasters perform the deep deliberative System 2 thinking so frequently and facilely, that it becomes their System 1 thinking, effectively making the deliberative process second nature. It anchors their estimates in the right ballpark.

These individuals begin their decision process by first taking "the outside view," by considering things like base rates *before* the details of the problem at hand. For example, if asked to predict whether there will be a recession within the next year, the outside view would ask, "How many recessions have there been in the past, say 100 years?" This establishes a base rate, which they can then adjust up or down depending on the specific details of the problem at hand. In contrast, most people are led astray by the details of the problem that anchors them in the wrong ballpark.

Finally, Tetlock observes that skilled forecasters make the adjustments to their estimates more frequently and at a more granular level.

The valuation guru Aswath Damodaran, who has been publishing his analyses and investment recommendations for decades, perhaps best exemplifies superforecasters in Finance. Damodaran's numerous "Musings on Markets" posts have received hundreds of millions of views over the past 20 years. Damodaran uses "credible narratives," as an inherent part of his valuation work, in addition to his traditional analysis of "the numbers." These narratives provide an explanation for a decision in qualitative terms.

For example, consider Damodaran's analysis of the *IPO of Zomato* (Damodaran, 2021), a new food delivery company in India. What is the "outside view" in such a case? A "similar situation" that has already played out? What could this be? For Zomato, he points to things like market penetration and margins in a comparable market like China, where the industry has already matured. Adjustments to this base rate involve things like differences in eating habits and income levels between India and China. These form a story. Damodaran makes the case that people understand stories. The trick is to come up with stories that are honest.

Interestingly, superforecasters tend to be even better as a group than they are individually, suggesting that a group of Damodarans—hard as it is to find, would do even better than Damodaran himself, if the group has a chance to consider the estimates of others. As Tetlock points out, however, this "process gain" for a team (over individuals independently) is typically very difficult to achieve and hence quite rare. It requires a process that is able to derive a synergy from the group. Tetlock explains the success of one such group as follows:

They helped each other. They managed to divide the labor effectively. They asked each other challenging questions. And they avoided the pathologies that degrade group decision making in many workplaces. They avoided groupthink. They avoided free riding. So, they avoided the perils of *groupthink and free riding and factionalism. They also tend to be more curious, measured by the number of questions they ask. They are also more likely to gather news and opinion pieces and share them. They comment more on other peoples' queries.* (Brave New World, Episode 31.)

But it usually takes a long time to confirm investment skill. A pattern of consistently good outcomes often takes years to pan out, and very few people actually keep score. Which raises a tantalizing question: Is there a ***process*** one might follow to achieve superior investment performance?

One method that appears promising for this purpose is the method of "*reciprocal scoring*," developed by Ezra Karger and colleagues (Karger *et al.*, 2021). The central concept of reciprocal scoring is that rather than asking forecasters to make their own predictions about an outcome, they predict the predictions of *other superforecasters*. The method is particularly useful in forecasting things that do not have an objectively resolvable outcome, or where we cannot wait for an answer, such as:

- "What is the best way to keep the casualty count in Ukraine below 100,000?"
- "What is the size above which gatherings should be prohibited to curtail the spread of COVID19?"
- "What is the expected number of years for the closing price of the S&P500 index to double from its close on March 31, 2020?"
- "What is the probability that Google will beat earnings estimates for the next four quarters?"

The expectation is that by going through an exercise where participants try hard to predict the answers of other thoughtful individuals, they will arrive at a more thoughtful answer themselves. Tetlock explains the thinking behind the method as follows:

We're asking you to become your better cognitive self to look at the world as you would if you were really a very thoughtful person trying to reach the correct answer, as opposed to trying to generate a convenient answer. The research shows that when people are trying to predict the predictions of smart people, it is as though they are doing the same things they do when they're trying to predict the objective truth. And that's the mindset you want them to carry over when they start trying to answer controversial policy questions that don't have objectively resolvable answers. You want to be sure that that mindset perseveres into domains that are easily distorted by wishful thinking, ideological prejudices, and so forth. It produces the kind of epistemic accountability, where you're trying to get to the truth. And it I suppose, it rests on the optimistic assumption that inside each of us, there is some little voice that says, think again. (Brave New World, Episode 31.)

If this is true, it suggests that "clubs" of highly motivated skilled individuals might do well at prediction if they try and predict the predictions of other such individuals. As Tetlock cautions, however, the success of this method would depend not only on the individual abilities of the team members, but their collective ability to eke out a process gain by avoiding the usual perils of group decision-making.

4 Machine-Learning-Based Investing

The ability to learn predictive models from data arises when data are abundant and the target is well defined. In finance, high-frequency trading has been machine-based for many years. Short-term price fluctuations are often predictable. At lower frequencies, however, involving holding positions over multiple days, predictability declines rapidly. What does machine learning bring to the table for short-term prediction (Dhar, 2013)?

It is useful to contrast machine learning models with traditional quantitative models. For starters, the latter impose a strong structural and theoretical prior on the nature of the underlying data generating process. Said differently, model specification is generally presumed to reflect some sort of prior theory. The objective of such models is to *explain and (sometimes) predict* (Shmueli, 2010) some economic or financial behavior. The explanation part can be understood in terms of the sign and magnitude of the model parameters (sometimes evaluated locally for non-linear models). At a surface level, this contrasts with machine learning models which offer far more flexible nonlinear model structures, and require few or no assumptions about the data and data-generating process.

Leo Breiman contrasts the "*two cultures*" (Breiman, 2001). He notes that by freeing themselves from a top-down specification of traditional models, machine learning models gain in accuracy, albeit at the expense of some transparency. Equally significantly, an ensemble of simple nonlinear models tends to provide better predictive power than single monolithic models. Ensembles also reduce performance variance.

Perhaps even more significantly, machine learning models are designed to handle multiple types of data naturally, such as text and images that are difficult for traditional methods to deal with. It is relatively easy for an AI system to analyze language data for estimating things like the "sentiment" expressed in text, or figure out the condition of something from an image. Traditional econometric methods find this virtually impossible without a lot of effort.

Finally, a recent key development in AI is the concept of "transfer learning," where models learned in one domain are transferred to another without modification. For example, current-day natural language models have been trained on vast amounts of publicly available data, and learn the implicit relationships among things expressed in such data. These models can be used "*out of the box" for all kinds of general-purpose reasoning*. As the data grow, the algorithms get better.

In summary, machine learning models eek out accuracy at the expense of complexity. How do these advances transfer to prediction in finance? In particular, now are bias and noise embedded in such algorithms?

Machines and bias

Decision-making models created from machine learning algorithms tend to reflect the bias in data (Barocas and Selbst, 2016). This phenomenon has been observed in domains such as the justice system and credit decisions. For example, researchers have pointed to bias in the COMPAS system that is widely used by *U.S. courts* to assess the likelihood of a defendant becoming a recidivist (Larson *et al.*, 2016).

In finance, the nature of bias is very different, and is driven by the base rate of the phenomenon of interest, which is typically the rate of occurrence of the minority class. In other words, we are interested in "failures," which occur a minority of the time, not the majority "normal" outcome, which is the default prediction.

The interesting questions here are whether machine learning algorithms **amplify** bias in data, and how amplification is impacted by the nature of the problem. Is it related to the degree of objective ignorance of problems? Is it impacted by the base rates? The answers to both questions are yes, but quantifying the impact is the interesting part.

Figure 2, from my prior research, shows the predictive bias of machine learning models applied to problems with varying degrees of predictability and base rates. For details of the experiment and the dataset construction, see Dhar and Yu (2020).

The figure shows that for problems with high objective ignorance (low predictability) and low base rates, the machine has a hard time predicting the minority class often enough. This makes sense. If false positives are frequent, which is to be expected for high objective ignorance and low base rates, the algorithm will be biased towards

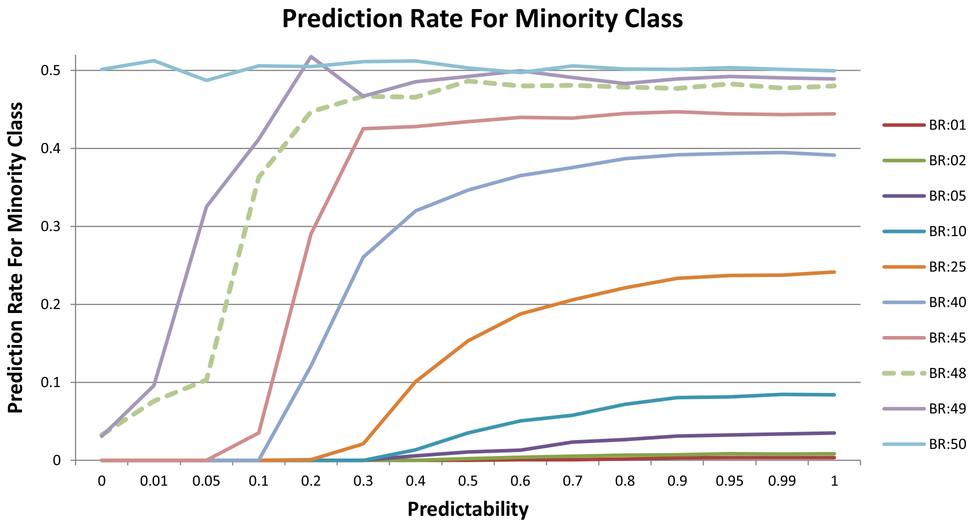

Figure 2

predicting the majority class in order to minimize loss.

Interestingly, as Figure 2 shows, if the base rate is low, the algorithm is only able to start predicting the minority class for higher levels of predictability, that exceed some minimum threshold. As we might expect, the prediction rate of the minority class converges asymptotically towards its base rate as the problem approaches determinism. In this extreme case, the predicted and actual outcome distributions should be identical. At the other extreme, with total objective ignorance, the predicted distribution will shrink to a single point, namely, the mean outcome.

The novel result is that *machine learning algorithms amplify bias in the data in inverse proportion to predictability* (Dhar and Yu, 2020), and the model's bias increases as the base rate gets lower.

Removing bias from data algorithmically is an intense area of current research (Hall *et al.*, 2022). In my current research on this subject in finance, I direct the learning algorithms to search for models that exhibit "behavior" that we desire, such as predicting the minority class sufficiently often and having low correlation with the benchmark. By sufficiently often, I mean that its prediction rate for the minority class should approach the base rate, all else being equal. In capital markets, for example, this would require a system to be short the market index almost half the time and still be profitable.

It is worth noting that the amplification of bias result in Figure 2 generalizes beyond finance. It is likely to occur in all domains—medical, law, lending etc., where historical data exhibit some sort of bias in terms of gender, race, and other factors that could have impacted prior decisions. My research quantifies the extent to which bias depends on the specifics of the problem, in terms of objective ignorance and base rates. It indicates when we should be wary about the bias-driven error in learning algorithms.

This results above are important in establishing expectations about the behavior of machine learning algorithms problems on problems with different base rates and levels of predictability. For problems such as credit default, which occur infrequently and have very low base rates, the predictability must be high in order to predict default, otherwise the model will always predict no-default. Such a model may be accurate, but it would be useless. The good news, however, is that for relatively balanced problems, such as predicting direction of equity markets, the algorithm becomes capable of predicting the minority class even with a relatively small level of inherent predictability.

Machines and Noise

Consider Kahneman's definition of noise, namely, undesirable variance in decision-making. Kahneman contrasts noisy humans with noise-free machines in that the latter produce the same output given an input, regardless of irrelevant factors such as the outcome of the football game the night before. In other words, when a machine makes predictions using a deterministic function that it has learned from data, it will always produce the same output when given a specific input.[2]

But, this observation is more tautological than constructive, and assumes stationarity. It reminds me of the Yogi Berra quote "in theory there's no difference between theory and practice, in practice there is." The unstated implication is that the machine is consistent, *assuming that the "true" process* is known. In reality we don't know the true process, and must make decisions about how to construct the training set. What history should be used? In non-stationary domains, the trained model, and hence its decisions on unseen data, will vary depending on the choice of the training set. In other words, machines are not immune to noise when there is variance in decisions in response to variance in the training set. I refer to this phenomenon as "model variance."

One way to think about model variance is to consider the "flip rate" of the predictions as a function of both the predictability and the base rate of a phenomenon. The flip rate represents the percentage of ***decisions*** that change as we perturb the training data along one of these two dimensions.

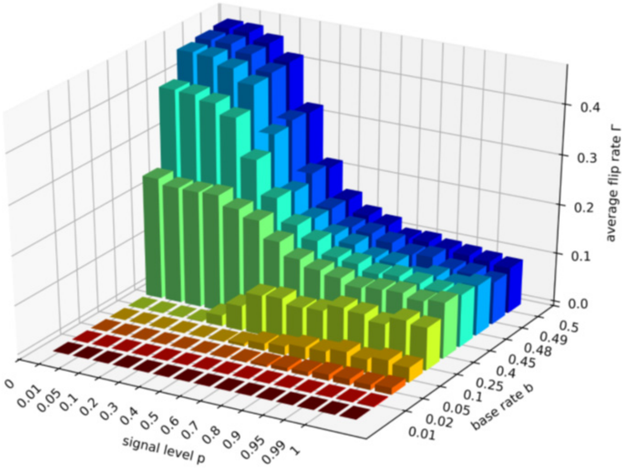

Figure 3

Figure 3 shows the results of simulation experiments that colleagues and I did to explore this phenomenon. The figure shows the variance, in the form of the flip-rate, along the vertical axis, with predictability and base rate represented on the horizontal axes. (Here I have reversed the axis for predictability, relative to the earlier format to make it easier to see the relationships.)

Model variance is a measure of the ***stability*** of decision-making of the learned model. Figure 3 has two distinctive regions. The first has very low model variance across the predictability axis. The second region has higher variance which, uncreases as unpredictability increases.

We see low variance when base rates are very low, five percent and lower. The reason for this result is similar to what drives model bias. It is challenging to predict the minority class with low base rates since false positives are frequent, so the machine is biased towards predicting the majority class. In this situation, the low model variance is not a good thing, but rather, reflects the difficulty of predicting the minority class accurately, especially as the problem gets harder in terms of low base rates and low predictability.

Figure 3 shows that model variance increases with increasing base rate, and more so as predictability decreases. This makes intuitive sense. Imagine a random phenomenon that is well balanced, with about the same number of positive and negative cases on average. For example, the daily direction of the equity market, is fairly well balanced, albeit with a very slightly higher frequency of positive returns. If the problem is random, the learning algorithm will essentially pick up on the noise in the problem, so a learned model's decisions will be highly variable, even though the performance might be stable. In other words, when predictability is low, small changes in the training set lead to significant changes in its decision-making.

The critical question above is whether the level of model variance is acceptable. The strength of the machine learning approach is that it can provide estimates of model variance, such as the model stability measured in terms of flip rates, and estimates of performance variance of metrics such as Sharpe Ratio, AUC and accuracy. With humans, we have no such estimates, and we must rely on their abilities, which, as we have seen, vary significantly.

5 Humans, Machines, or Humans + Machines?

In bringing together the different themes I've been discussing, it is useful to summarize and highlight the key differences in the types of errors that arise in humans relative to those that arise in the use of machine-learning algorithms. Table 1 summarizes this in terms of the origins and impacts of bias and noise in humans and machines.

Table 1 shows that bias and noise have very different origins in humans versus machines. Humans carry prior biases from experience that result in myopic limited attention and a disposition effect. These biases manifest themselves in decisions.

In contrast, machines pick up bias in the data, and amplify it.[3] The extent of bias amplification depends on objective ignorance and base rates, as we demonstrated in the previous section.

Table 1

	HUMANS	MACHINES
BIAS	• limited attention • disposition effect *Comment: These biases often arise from the desire to "enhance the emotional experience" of investing.*	• statistical amplification of: • implicit bias in the data • Implicit or explicit "upstream" bias in data collection, filtering or preprocessing by modeler or others *Comment: Amplification is proportional to the objective ignorance of the problem and unfluenced by the base rate)*
NOISE	• different levels of bias across decision-makers • occasion and pattern noise of individual decision-makers. *Comment: There is generally no consistent mechanism for quantifying this type of instability or "mental model variance"*	• non-stationarity of the underlying phenomenon • variations in the training data lead to instability/variance in decision-making *Comment: Quantitative estimates of such instability (model variance) are feasible*

Specifically, the higher the objective ignorance and lower the base rate, the higher the bias amplification.

With respect to noise, humans tend to be naturally inconsistent for the reasons we have discussed. Quickly recognizing the context of and analogs to a problem, thinking "fast," is both a human strength and a weakness, but it also results in noise, to a degree that we cannot easily estimate.

Decades of research into behavioral psychology has led most of us to expect this type of bias and noise in humans, but we've tended to hold that, in theory, machines are not susceptible to such noise. And, in general they are not susceptible to *that kind of* noise. In practice, however, the collection and selection of training data (by humans) naturally and inevitably introduces noise in statistical learning.

Trust in machines depends on the extent of such noise. The good news is that we have good methods for calculating reasonable quantitative estimates of such model noise (instability), *if we know which aspects of the data and model to look at*. In making a deployment decision, decision-makers must decide whether the level of instability is acceptable, and whether estimates of this instability are sufficiently rigorous and thoughtful.

The chess grandmaster Gary Kasparov, who has played multiple times against progressively better machines, argues for the superiority of the carbon-silicon combination despite the fact that machines handily defeat humans on their own. In a series of tournaments, Kasparov observed that the chess machine *Hydra* (a chess-specific supercomputer like Deep Blue), was no match for a strong human player when that player was aided by access to a simple model. That experience led Kasparov to conclude that, "*Human strategic guidance combined with the tactical acuity of a computer was overwhelming*."

Is Kasparov's result a general one? Is there a Goldilocks strategy that can combine the best of humans and machines and avoid the worst?

This is, of course, a veiled version of a commonly asked question: can humans plus machines outperform machines (or humans) alone. The current conventional thinking seems to be "yes," although the evidence for such optimism is scant. Such a position assumes that the human can recognize situations when judgment should be exercised and will be attentive enough to take the right action. My response to this question is less optimistic. While this may be true for simple problems with solid ground truth where the error of the machine is obvious, as in the cases of the two Boeing Max 737 autopilot failures in October 2018 and March 2019, my own experience in capital markets is that situations in which a human needs to intervene are exceptionally difficult to even recognize, let alone act on correctly.

One reason for this is that humans and machines can have very different prediction horizons for investment decisions. Under the Efficient Market Hypotheses, information is reflected instantaneously in prices. In other words, the time taken to adjust prices is zero. After being trained on historical market data, machine learning-based models, typically assume that there is a nonzero delay (that's the whole point, after all), and attempt to learn the rules that capture short-term mispricing. Unlike humans, machines have little situational awareness or larger context, and for such models, long-term positions are simply an aggregation of a series of "greedy," if myopic, decisions. There is generally no Damodaran-style narrative to anchor such predictions, such as how the policy of the Federal Reserve is likely to play out, whether it is likely to fuel inflation, and the impact of such inflation on asset prices.

Could humans profitably add a layer on the machine's decisions through the inherently human appreciation of a larger context?

Andrew McAfee has taken up this question in his research. McAfee's conjecture for Kasparaov's results is that the benefit of the combination was driven by the application of a "process" followed by the hybrid computer-human combination (McAfee, 2010). He argues that the better results consistently enjoyed by some teams of human-machine decision makers ensue from their employment of "better" processes in their hybrid decision-making. These teams apply their processes consistently across tasks. Developing such a process could reduce the pattern and occasion noise that Kahneman described.

McAfee summarizes the result as follows:

> *Weak human + machine + better process was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process.*

This also sounds strikingly similar to Tetlock's teams of superforecasters, who also followed a specific process to make their decisions, but in McAfee's case, this process (or team) has been augmented through the introduction of computer algorithms to gather and analyze data, which allows the team and team members to iteratively refine their estimates and focus on evaluating patterns.

Kahneman, however, cautions a more nuanced perspective on hybrid strategies than Kasparov and McAfee. His perspective emphasizes distinguishing between problems for which there is currently too much epistemic uncertainty and insufficient data to build reliable models, on the one hand, and those for which computers already perform comparably to, or better than, humans, on the other. For these latter problems, Kahneman notes:

> *You can see why if a diagnostic program comes close to being as good as a diagnostic physician, it will not take very long for the program to do much better than the physician. And this is simply because programs can assimilate, accumulate, learning different programs in the offices of different physicians, they all learn together. A single physician learning could never be able to match this. If you let the human have the last word, there will be more mistakes than if you let the machine have the last word. So, we really have to accept the fact that* ***this machine human combination is unstable****. You should let the human have the last word only in cases that lie completely out of the distribution.* (Brave New World, Episode 21).

Kahneman's observation has important implications for determining meta-decision policies governing when a human should take control of an automated decision-making system. It stands to reason that when computers handily outperform humans, the less human involvement the better. However, even here, the role for the human operator is still to figure out whether the current situation is one that had been represented in the training data, or whether it is completely novel, in which case an intervening human decision might perform better.

What this implies is that the real challenge for an operator of an automated system is not to outperform the machine most of the time, but rather to be confident enough in identifying those uncommon settings in which a situation is out of the distribution of the training data, and to act. This can be difficult for a human in the heat of the moment. I wrote about this in my discussion of running a trading system during the onset of the COVID *pandemic* (Dhar, 2020):

> *Crisis environments create considerable uncertainty regarding the cost of errors. In such settings, an airline pilot (or an algorithm) must be able to make the right decision in the heat of the moment and in mid-flight. The same applies to the healthcare professional (or a decision-making system) in a life and death situation when a patient*

is in critical condition and worsening. Or to an investment professional (or algorithm), confronting a new "unknown unknown." In crises, the spike in the cost of error shifts the problem upwards into the red zone (in Figure 1). *Furthermore, in some cases, because the environment is often quickly changing, and typically unfamiliar, the likelihood of errors increases, pushing these decisions towards the red zone.* (Dhar, 2020)

The COVID crisis was challenging for portfolio managers. As unbelievable as it might seem in retrospect, it was difficult to ascertain in the early days of COVID whether it represented the beginning of a unique crisis or a blip in business as usual. Human attention in February 2020 was focused on the US presidential elections, especially the primaries. Even as the crisis started to become apparent, it was difficult to anticipate the nature of the step function interventions that leaders and central banks would make to deal with the impending crisis, and what the impact of such interventions might be.

This question of how to determine when human plus machine works better than the machine alone remains a major open research question at this time. Humans are noisy and often misled by context, whereas algorithms are more consistent, but incapable of recognizing the larger context. But it also seems that some humans with eclectic cognitive styles are capable of seeing well into the future without getting misled by context. How can decision-makers and organizations combine such human strengths with algorithms, and how can we create the right interface between the two? From the discussion above, we might conjecture that the answer should depend on the nature of the problem, including our objective ignorance, phenomenological base rates, and the practical costs of errors.

Perhaps the biggest challenge for humans and machines in making predictions is determining how to ask the right questions. In a previous article on *Data Science and Prediction* Dhar (2013), I argued that machines are becoming capable of asking good questions. However, such questions are currently narrowly defined and circumscribed by limitations on available data. Tetlock's superforecasters are skilled at answering questions, but how do they know how to ask the "right" questions in the first place?

Perhaps those who make better predictions also tend to ask better questions. This still seems to be more of an art than a science, and it is an area in which humans still have an edge over machines in most settings. These types of "Rumsfeldian unknown unknowns" remain a major factor in policies that include a human in the loop at some level. In other words, the buck ultimately stops with the human, and ultimately, humans must live with the decisions that they, or their machines make.

Notes

[1] By "programmability" I mean the degree to which a problem can be well represented algorithmically as a set of operations that can be done by a computer. For more on programmed vs. non-programmed decision-making, see Simon (1965).

[2] I do not consider "stochastic prediction," such as when one deliberately adds noise to the model to produce more varied outputs, or considers an additional parameter during prediction, such as the "temperature parameter" in a large language model.

[3] The source of the bias in the data may be from the phenomenal itself, as in capital markets where there is a "upward bias," or from human decisions reflected in the data, such as in credit markets or the legal system.

Acknowledgment

Many thanks to Roger Stein for reading several drafts that sharpened the article considerely.

Podcast Links

Daniel Kahneman, Brave New World Episode 21.
https://bravenewpodcast.com/episodes/2021/09/16/episode-21-daniel-kahneman-on-how-noise-hampers-judgement/

Philip Tetlock, Brave New World Episode 31.
https://bravenewpodcast.com/episodes/2022/02/03/episode-31-philip-tetlock-on-the-art-of-forecasting/

Terry Odean, Brave New World Episode 23.
https://bravenewpodcast.com/episodes/2021/10/14/episode-23-terry-odean-on-how-to-think-about-investing/

Aswath Damodaran, Brave New World Episode 33.
https://bravenewpodcast.com/episodes/2022/03/03/episode-33-aswath-damodaran-on-investing/

References

Barber, B., Huang, X., Odean, T., and Schwarz, C. (2021). "Attention Induced Trading and Returns: Evidence from Robinhood Users," *Journal of Finance*, forthcoming, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3715077.

Barber, B. and Odean, Y. (2000). "Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors," *Journal of Finance* **55**(2), https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00226.

Barber, B. and Odean, T. (2008). "All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors," *The Review of Financial Studies* **21**(2), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=460660.

Barocas, S. and Selbst, A. (2016). "Big Data's Disparate Impact," *California Law Review* **104**(3), 671–732.

Breiman, L. (2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," *Statist. Sci.* **16**(3), 199–231, https://doi.org/10.1214/ss/1009213726.

Damodaran, A. (2017). "Narratives and Numbers: The Value of Stories in Business," *Columbia Business School Publishing*.

Damodaran, A. (2021). "Musings on Markets, The Zomato IPO: A Bet on Big Markets and Platforms," https://aswathdamodaran.blogspot.com/2021/07/the-zomato-ipo-bet-on-big-markets-and.html.

Dhar, V. (2013). "Data science and prediction," *Communications of the ACM* **56**(12).

Dhar, V. (2016). "When To Trust Machines With Decisions and When Not To," *Harvard Business Review*.

Dhar, V. (2020). "Algorithms in Crisis: When Context Matters," *Medium*, https://medium.com/firmai/algorithms-in-crises-when-context-matters-6c87e26fc3aa.

Dhar, V. and Yu, H. (2020). "On the Stability of Machine Learning Models: Measuring Model and Outcome Variance," *Journal of Investment Management* **18**(2).

Hall, M., Van der Maaten, L., Gustafson, L., Jones, M., and Adcock, A. (2022). A Systematic Study of Bias Amplification, https://arxiv.org/abs/2201.11706.

Hüllermeier, E. and Waegeman, W. (2021). "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods," *Machine Learning* **110**, 457–506, https://arxiv.org/abs/1910.09457.

Israel, R., Kelly, B., and Moskowitz, T. (2020). "Can Machines "Learn" Finance?," *Journal of Investment Management* **18**(2).

Kahnaman, D. (2011). "Thinking Fast and Slow Farrar," (Straus and Giroux, New York).

Kahneman D. and Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk Econometrica," **47**(2).

Kahneman, D., Sibony, O., and Sunstein, C. (2021). "Noise: A Flaw in Human Judgment," (New York: Little, Brown Spark).

Karger, E., Monrad, J., Mellers, B., and Tetlock, P. (2021). "Reciprocal Scoring: A Method for Forecasting Unanswerable Questions," Available at SSRN: https://ssrn.com/abstract=3954498 or http://dx.doi.org/10.2139/ssrn.3954498.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm, Propublica, May 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

McAfee, A. (2010). "Did Garry Kasparov Stumble Into a New Business Process Model?," *Harvard Business Review*.

Odean, T. (1998). "Are Investors Reluctant to Realize Their Losses?," *The Journal of Finance* **53**(5). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=94142.

Ozkan, E. and Naci, M. (2018). "Emotional Judges and Unlucky Juveniles," *American Economic Journal: Applied Economics* **10**(3).

Shmueli, G. (2010). "To Explain or to Predict?," *Statist. Sci.* **25**(3), 289–310, https://doi.org/10.1214/10-STS330.

Tetlock, P. and Gardner, D. (2015). "Superforecasting: The Art and Science of Prediction," (Broadway Books).