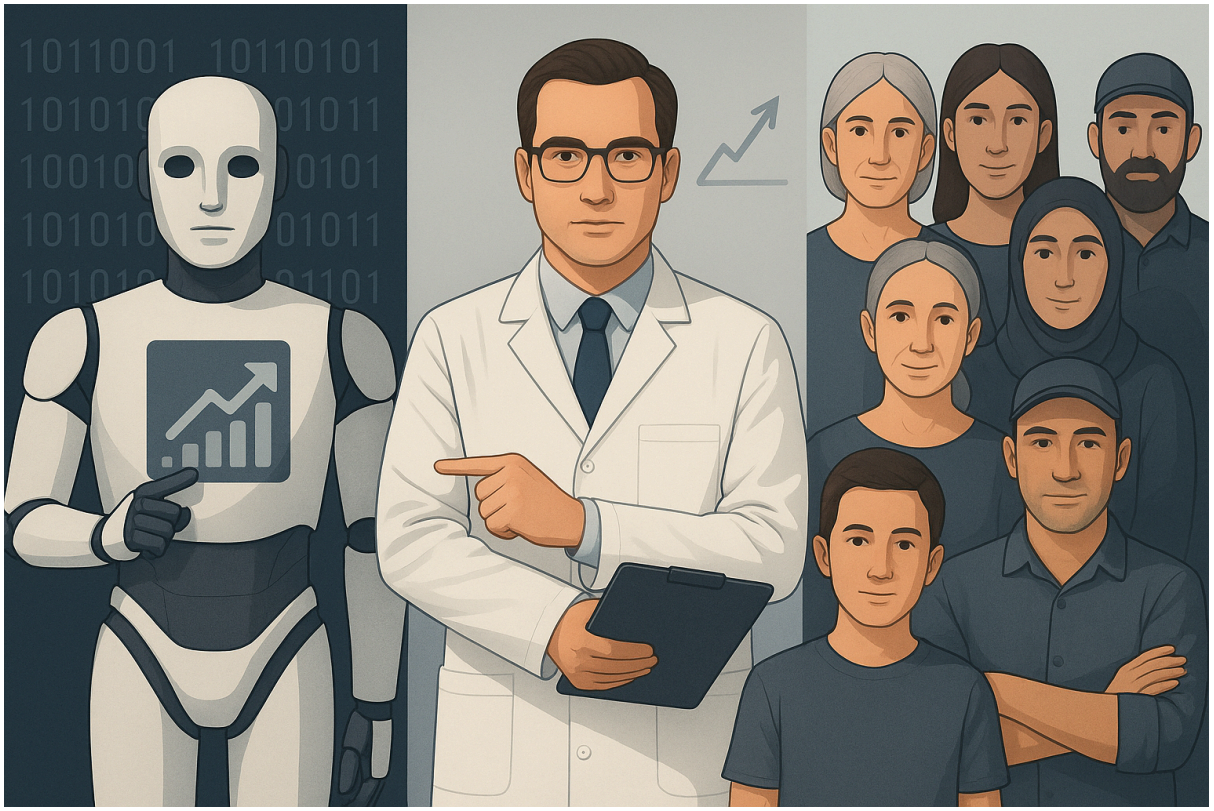
The Machine, The Expert and the Common Folks

A look at noise, consistency and broken legs

by Lars Nørtoft Reiter



Picture a judge about to hand down a sentence just before lunch. Most people would assume the timing doesn't matter for the outcome. They'd be mistaken. Because when judges get hungry, justice gets harsh - a phenomenon known as the *hungry judge effect* [1]. But it's not just a growling stomach and low blood sugar that can influence a judge's, or in fact anyone's, decision. Other seemingly irrelevant factors can also play a role [2,3], such as whether it is the defendant's birthday, whether it's hot outside, or more generally, the judge's mood.

This highlights one of the main concerns in decision-making: where there are people, there is variability ("noise") and bias. So it begs the question: can the machine do better? Before we answer that question, let us first explore in what way people are noisy. Disclaimer: many of the concepts introduced in this article are described in the book *Noise* by Daniel Kahneman (author of *Thinking, Fast and Slow*) and his colleagues Oliver Sibony and Cass R. Sunstein [4].

Noisy people

The authors of Noise identify three sources of human noise.
One is called *level noise*. This describes how soft or extreme an individual's judgement is compared to the average individual. For example, a judge with a high *justice sensitivity* might impose harsher sentences than a more lenient colleague. Level noise is also related to the subjective scale by which we rate something. Imagine that two judges agree on a "moderate sentence", but due to level noise, a moderate sentence in one's perspective is a harsh sentence to the other judge. This is similar to when rating a restaurant. You and your friend might have enjoyed the experience equally. Still one of you "only" gave it four out of five stars, while the other gave it five stars.

Another source is called *(stable) pattern noise*. This describes how an individual's decision is influenced by factors that should be irrelevant in a given situation. Say, if a judge is more lenient (compared to the judge's baseline level) when the defendant is a single mother - perhaps because the judge has a daughter who happens to be a single mother. Or going back to the restaurant rating example, if, for whatever reason, your rating system is different based on whether it is an Italian or French restaurant.

The final source of noise is occasion *noise*. It is also called *transient* pattern noise, because like pattern noise, it involves irrelevant factors influencing decisions. But unlike pattern noise, occasion noise is only momentary. The

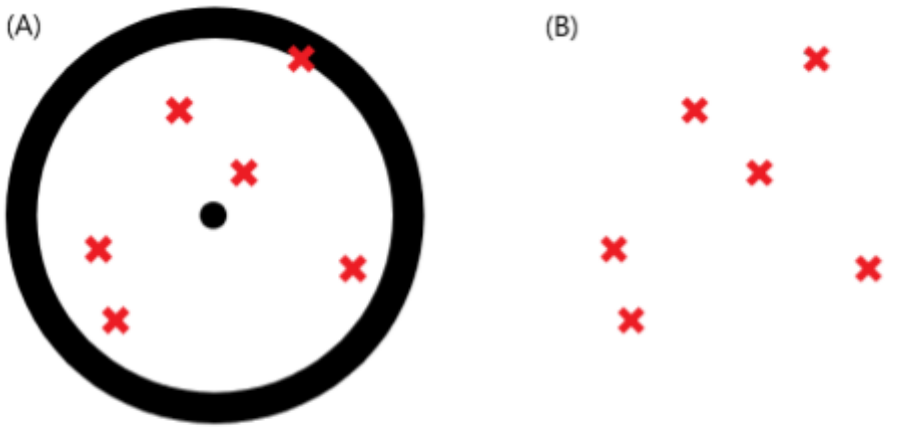
hungry judge from the introduction shows occasion noise in action, where the timing (before/after lunch) changes the severity of the sentence. More generally, mood causes occasion noise and changes how we respond to different situations. The same experience can feel very different depending on your mental state.

Now that we better understand noise, let's now look at two types of decisions where noise infiltrates.

Prediction and evaluation

Often we want the quality of a decision to be measurable. When we go to a doctor, it is nice to know that many patients before you got the proper treatment: the assessment of the doctor was correct. On the other hand, when you're watching the Lord of the Rings movies with friends who have wildly different opinions about how to rate it, you have to respect that there's no universal truth (and if there were, it would obviously be that Lord of the Rings is the greatest film series ever).

With that in mind, we need to distinguish between *predictions* and *evaluations*. Predictions imply a single (verifiable) truth, evaluations do not. This in turn implies that predictions can be biased, since there is a universal truth, whereas evaluations cannot be biased per se. Both can still be noisy however. See the Figure below.



(A) Bias and Noise (B) Noise without bias

My movie example likely made it seem as if cases of evaluations are unimportant. It is a matter of taste, right? But even if there is no bias (in the statistical sense), there is still noise. The example given in the introduction is a case of evaluation. There is no universal correct sentence. Still, if different judges impose different sentences the result is a noisy and unjust judicial system. Thus, cases of evaluations can be equally important.

Next I will show that what distinguishes humans from machines is (among many other things) our lack of consistency.

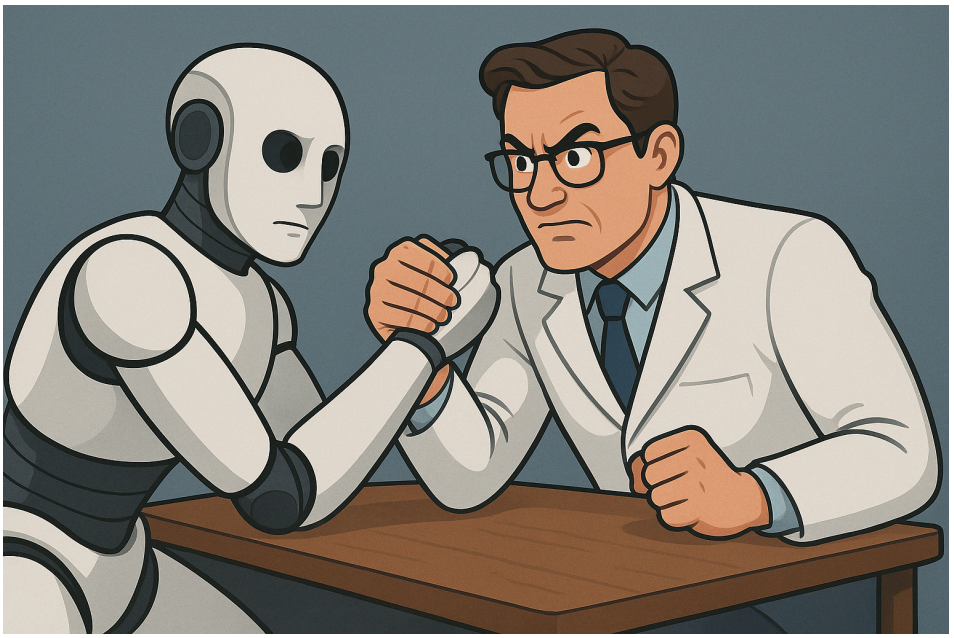
Consistency beats complex rules

In a study from 2020, researchers wanted to see how experts matched up against simple rules in predictive tasks [5]. The researchers acquired archival assessment validation datasets (three batches/groups of candidates) supplied by a large consulting firm, which contained performance information on a total of 847 candidates, such as the result of personality tests, cognitive tests and interviews. Experts were then asked to assess all 847 candidates across 7 categories (such as Leadership, Communication, Motivation, etc.) by assigning scores from 1 to 10 points. Based on their assigned scores across these 7 categories, the experts then had to predict what score the candidates would

achieve in a performance evaluation (also from 1 to 10 points) which were conducted two years later.

The researchers then built more than 10,000 linear models, where each model generated its own random weights for each of the 7 categories. Each model then used the randomly generated weights along with the points given by experts for each of the seven categories to make consistent (i.e. fixed weight) performance evaluation predictions across all 847 candidates. Finally, these predictions were compared against the experts' predictions.

The result was thoughtprovoking: in two out of the three candidate groups, *every single* model was better at predicting the performance evaluation scores than the experts. In the remaining group, "only" 77% of the models came closer to the final evaluation than the human experts did.

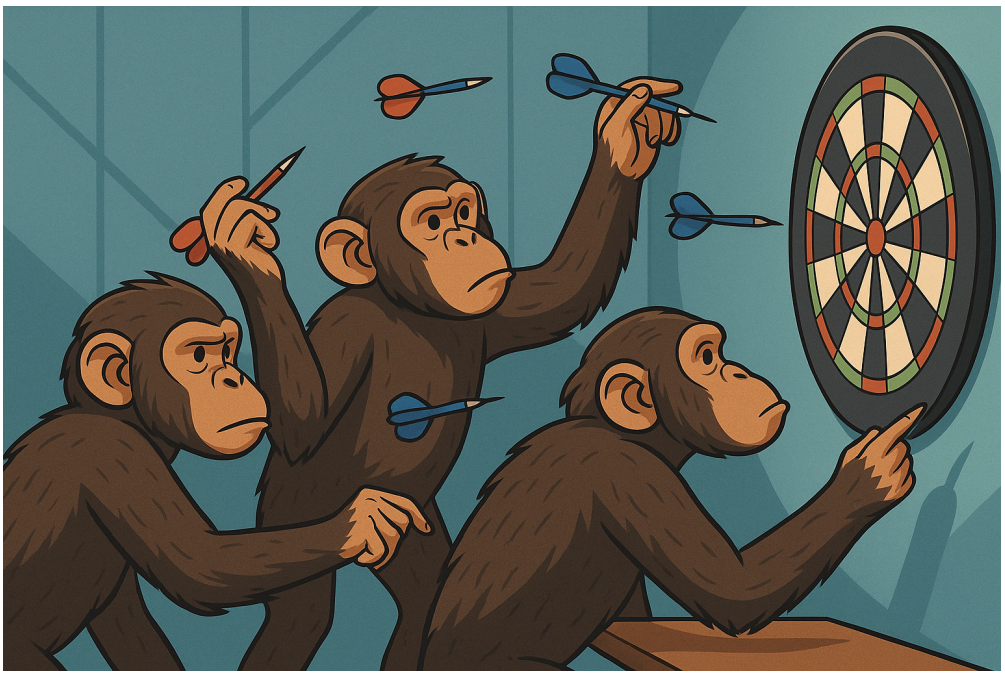


So how could simple mathematical models beat experts? According to the authors of Noise (from which the example is taken), we humans weigh different categories similar to the simple models. But unlike the simple models, our own mental models are so complex that we lose the ability to reproduce our own rules, and *noise* takes over. The simple models, by contrast, are both consistent

and partly noise free. They are only affected by whatever occasion noise (mood for example) or pattern noise that went into the category evaluation score, but not in the final performance evaluation.

The study is interesting, because it shows the extent of human noise in predictive tasks, where mindless consistency appears superior to mindful expertise. But as the authors also warn, we should be careful not to overgeneralize from these three datasets focused on managerial assessment, as different settings and other types of expertise may yield different results. In this study, it was also shown that the experts outperformed pure randomness (where the model used different random weights for each candidate), indicating the presence of valid expert insight. Consistency was the critical missing ingredient.

This finding isn't unique. There are several studies that similarly document how “machines” (or simple rules) tend to outperform humans and experts. Another example is in the book Expert Political Judgment by Philip Tetlock who became famous for the statement that “the average expert was roughly as accurate as a dart-throwing chimpanzee”. Behind this statement lies a study involving 80,000 predictions made by 284 expert forecasters across different fields, all assessed after a 20-year period. You can imagine how that turned out.



Since mathematical models are the backbone of machines, the examples provide evidence that machines can outperform humans. It is not hard however to think of examples, where the complexity and nuanced view of the expert will be superior to a simple machine. Consider a famous example by the psychologist Paul Meehl. If a machine confidently predicts that a person will go to the movies with a 90% probability, but the clinician knows that the same person has just broken his leg, the clinician (who now takes the role of “the expert”) has access to information that should overwrite the machine prediction. The cause is obvious, however: the machine is lacking data while the human is more informed.

Both the movie-goer and performance evaluation examples consider predictions. But when it comes to evaluations, machine limitation becomes even more obvious in domains that demand contextual judgements. Such as providing emotional support or giving career advice to an individual. Both situations demand a deep understanding of the subtle details that make up this individual, something humans understand better, especially those who know the person

well. Ethical decisions are another example, which frequently involve emotions and moral intuitions that many machines currently struggle with understanding. Despite these few human advantages, there is much literature supporting that machines are generally better at prediction, but only little evidence documenting that machines are *much* better. Since many of us are skeptical toward decisions made solely by soulless machines, it would require great technological advancement and documented performance superiority to overcome our reluctance.

AI: Finding the broken legs

It is well known that complex (unregularized) models are prone to overfitting, especially on small datasets. Luckily, in many domains today, datasets are large enough to support more complex deep learning models. If we go back to Paul Meehl's example with the movie-goer and the broken leg, this was a data problem. The clinician was more informed than the machine. Now imagine that the machine was more knowledgeable, in the sense that it is trained on more data. For example, it might have discovered a connection between hospitalisation and the lower probability of going to the cinema. There is a good chance that this model now correctly predicts a low probability of seeing this person at the movie, rather than the 90% the simple model produced.

In Meehl's example, a broken leg was a metaphor for something unforeseen by the machine, but understood by the human. For the complex model (lets call it AI) the roles have changed. This AI has not only removed the broken leg, it might also be able to see patterns that we, as humans, cannot. In that sense, the AI is now more knowledgeable and able to foresee broken legs that we could not have imagined. We are in a weaker position to overwrite or question the predictions.

We can only understand so much

If we go back to Philip Tetlock's study, and the dart-throwing chimpanzees, the problem leading to the inaccurate forecasts of the experts is likely caused by a well established cognitive bias: overconfidence. Specifically, confidence that one has enough details to make a plausible forecast of (highly uncertain) events in the future. In fact, one typically underestimates how little we know, and what we don't know (for whatever reason) is called *objective ignorance*. AI is impressive, but also suffers from the same limitation. No matter how much data we feed it, there are things that it cannot anticipate in this wildly complex world of billions and billions of interacting events. So while AI might do better than humans in keeping objective ignorance to a minimum, it will, as with human experts, have a natural limit where predictions become no better than those of a dart-throwing chimpanzee. Consider weather prediction. Despite modern and complex methods, such as ensemble forecasting, it remains hard to make predictions more than 2 weeks forward. This is because weather systems are chaotic, where small perturbations in the initial atmospheric conditions of the models can lead to entirely different chain of events. There is a lot of objective ignorance when doing weather forecasts.

Expert Proficiency and the Crowd

Human experts are inherently biased and noisy due to our complex, individual nature. This raises a natural question: Are some people less susceptible to noise, bias, and objective ignorance than others? The answer is yes. Generally speaking, there are two major categories that contribute to performance within decision-making. One is general intelligence (or general mental ability; GMA), the other we can call your Style Of Thinking (SOT). Concerning GMA, one would assume that many experts are already high-scorers, and one would be correct. Still, even within this group of high-scorers there is evidence on how the top quantile outperforms the lower quantiles [6]. The other factor, SOT, addresses how people engage in cognitive reflection. Kahneman is known for his system 1

and system 2 model of thinking. In this framework, people with an advanced style of thinking are more likely to engage in slow thinking (system 2). Thus these people are likely to overcome the fast conclusions of system 1, an inherent source to cognitive biases and noise.

These performance traits are also found in so-called Superforecasters, a term invented by Philip Tetlock, author of Expert Political Judgement and inventor of the dart-throwing chimpanzees. Following his study on expert forecasting, Tetlock founded The Good Judgement Project, an initiative that wanted to exploit the concept known as Wisdom of the Crowd (WotC) to predict future world events. Around 2% of the volunteers that entered the program did exceptionally well and were recruited into Tetlock's team of Superforecasters. Not surprisingly, these forecasters excelled in both GMA and SOT and, perhaps more surprisingly, these forecasters reportedly offered 30% better predictions than intelligence officers with access to actual classified information [7].

The motivation for using WotC for prediction is simple: people are noisy, and we should not rely on a single prediction, be it expert or non-expert. Aggregating several predictions however, we can hope to eliminate sources of noise. For this to work, we need of course many forecasters but equally important, if not more so, is diversity. If we were predicting the next pandemic using a crowd high in neuroticism, this homogeneous group might systematically overestimate the risk, predicting it would occur much sooner than in reality.

One must also consider how to aggregate information. Since one person might be more knowledgeable about a subject than the next person (experts being the extreme), a simple average of votes might not be the best choice. Instead, one could weight the votes by each person's past accuracy to promote more robust predictions. There are other ways to strengthen the prediction, and in the Good Judgement Project they have developed an elaborate training program with the goal of reducing noise and combat cognitive bias, thus enhancing accuracy of their Superforecasters (and in fact anyone else). It goes without saying that when it comes to domain specific predictions, a crowd needs expert knowledge. Letting the common folks try to predict when the sun burns out might yield alarmingly variable predictions, compared to those of astrophysicists.

Prediction without understanding

We have seen that machines can offer certain advantages over individual humans, partly because they process information more consistently, although they remain vulnerable to the biases and noise present in their training data. Even if some humans tend to overcome their own noise and bias owing to refined cognitive abilities (measured by GMA and SOT) they can still produce inaccurate decisions.

One way to mitigate this is aggregating different opinions from several people, preferably those less influenced by noise, bias and objective ignorance (such as

the Superforecasters). This approach recognizes that each person functions as a repository of vast information, though individuals often struggle to use that information consistently. When we aggregate predictions from multiple such "data-rich" individuals to compensate for their individual inaccuracies, this process bears some resemblance to how we feed large amounts of data into a machine and ask for its prediction. The key difference is that humans already contain extensive knowledge without requiring external data feeding.

One important distinction between people and current machine learning systems is that people can engage in explicit causal reasoning and understand underlying mechanisms. So while many deep learning models might produce more accurate predictions and discover subtler patterns, they typically cannot match humans' ability to reason explicitly about causal structure - though this gap may be narrowing as AI systems become more sophisticated.

[1] Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. Proc Natl Acad Sci U S A. 2011 Apr 26;108(17):6889-92. doi: 10.1073/pnas.1018033108. Epub 2011 Apr 11. PMID: 21482790; PMCID: PMC3084045.

[2] Chen, Daniel L., and Arnaud Philippe. "Clash of norms: judicial leniency on defendant birthdays." *Journal of Economic Behavior & Organization* 211 (2023): 324-344.

[3] Heyes, Anthony, and Soodeh Saberian. "Temperature and decisions: evidence from 207,000 court cases." *American Economic Journal: Applied Economics* 11, no. 2 (2019): 238-265.

[4] Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*.

[5] Yu, Martin C., and Nathan R. Kuncel. "Pushing the limits for judgmental consistency: comparing random weighting schemes with expert judgments." *Personnel Assessment and Decisions* 6, no. 2 (2020): 2.

[6] Lubinski, David. "Exceptional cognitive ability: the phenotype." *Behavior Genetics* 39, no. 4 (2009): 350-358. doi: 10.1007/s10519-009-9273-0.

[7] Vedantam, Shankar. "So you think you're smarter than a CIA agent." NPR, April 2, 2014. https://www.npr.org/sections/parallels/2014/04/02/297839429/-so-you-think-youre-smarter-than-a-cia-agent.