Wherever there is judgment there are 2 kinds of error: bias and noise (scatter). It is unacceptable for similar people, convicted of the same offense, to end up with dramatically different sentences—say, 5 years in jail for 1 and probation for another. And yet in many places, that happens. In a large insurance company underwriters set insurance premiums where the sheer volume of noise costs a great deal of money. Wherever there is judgment, there is noise, and more than you think.

The rule of law calls for a body of impersonal rules, applicable across the board, binding on judges as well as everyone else. A study of sentences for defendants in 1974 in hypothetical cases found that absence of consensus was the norm and that variations across punishments were astounding. Judges were more likely to grant parole at the beginning of the day or after a food break than immediately before. If judges are hungry, they are tougher. They are more severe on days that follow a loss by the local city’s football team. When it is hot outside, people are less likely to get asylum.

Sentencing guidelines were enacted in 1984 and they cut the noise. But for unrelated technical reasons the Supreme Court struck them down in 2005, and disparities increased significantly after that. Judges became more likely to base their sentencing decisions on their personal values. Female judges became more likely to exercise increased discretion in favor of leniency. The same is true of judges appointed by Democratic presidents. There was a return to something more like law without order.

Disagreement is unavoidable wherever judgment is involved and the extent of these disagreements is much greater than we expect. System noise (unwanted variability) creates rampant injustice, high economic costs, and errors of many kinds. Noise can be reduced in the judgments of people who make significant financial decisions. And everyone agrees that consistency is desirable.

The median difference in judgments of underwriters and claims adjusters based on written information was 55% and 43%, respectively—all unwanted. But variability in judgments is not always unwanted—like matters of preference or taste. It is also expected and welcome in a competitive situation in which the best judgments are rewarded—like the fair value of a stock.

Noisy systems do not just make multiple judgments of the same case. They make noisy judgments of different cases. Errors do not cancel out. They add up. No one had expected anything like the amount of noise we observed. The common inattention to noise is as interesting as its prevalence. The noise audits suggested that respected professionals maintained an illusion of agreement while in fact disagreeing in their daily professional judgments. Underwriters who heard about the noise audit and accepted its validity never believed that its conclusions applied to them personally. These beliefs, a naive realism, are essential to the sense of a reality we share with others, but we rarely question these beliefs. We can live comfortably with colleagues without noticing that they actually do not see the world as we do.

The illusion of agreement was shattered only by the noise audit. The leaders of the company considered conflict avoidance at least as important as making the right decisions. The amount of noise observed when an organization takes a serious look almost always comes as a shock.

There is a continuum between singular and recurrent decisions. While social scientists have dealt with recurrent decisions, high-stakes singular decisions have been the province of historians and management gurus. Consider how different countries and regions responded to the COVID-19 crisis. Even when the virus hit them roughly at the same time in a similar manner, there were wide differences in responses.

Judgment is measurement in which the instrument is a human mind. Matters of judgment differ from matters of opinion or taste in which unresolved differences are entirely acceptable. They are defined by the expectation of bounded disagreement. Selective attention and recall are a source of variability across people and within a single judge. Many professional judgments are conditional. The aim of judgment is a coherent solution. Focusing on the process of judgment, rather than its outcome, makes it possible to evaluate their quality. As we turn to the question of improving judgments rather than just evaluating them, we will focus on process.

Professionals are usually evaluated on how closely their judgments match verifiable outcomes. They usually claim to strive for a prediction that matches the outcome, the process that produces the best judgment. Evaluative judgments entail an expectation of bounded disagreement. The boundary between predictive and evaluative judgments is fuzzy and people who make judgments are often unaware of it. They develop confidence in their judgments and the justifications for them.

If there are large differences in sentences given to the same defendant, we are in the domain of the “arbitrary cruelties.” System noise is inconsistency which damages the credibility of the system. All we need to measure noise is multiple judgments of the same problem. We do not need to know a true value.

Overall Error = Bias² + Noise². Whenever you observe noise, you should work to reduce it! Noise and bias are independent. Both are costly but not equivalent. Don’t mix values and facts. Good decisions are unaffected by hopes and fears, or by preferences and values. They depend on underlying predictions, which should be value-neutral. New information provides more opportunities for judges to differ from one another.

Some judges have reputations for being harsh “hanging judges,” and others are more lenient “bleeding-heart judges.” We refer to these as level errors. The average level of sentencing functions like a personality trait. We use the term level noise for the variability of the judges’ average judgments. The remaining deviation we call pattern noise.

One judge may be lenient toward white-collar criminals and another inclined to punish more severely when the offender is a recidivist. A 3rd may be tough when the victim is an older person. System Noise² = Level Noise² + Pattern Noise². Occasion noise can be treated as random error. Our opinions change without apparent reason—mood, the sequence of cases that are fresh in our mind, or countless other features of the occasion.
When people form a carefully considered professional opinion, they associate it with the reasons that justify their point of view. When presented with the same problem a second time and they recognize it, they will reproduce the earlier answer both to minimize effort and maintain consistency. People in a good mood are more cooperative and elicit reciprocation, are more likely to let their biases affect their thinking and more gullible in general. You are not the same person at all times, changing with stress and fatigue. Physicians are significantly more likely to prescribe opioids at the end of a long day and under time pressure, they are apparently more inclined to choose a quick-fix solution. Bad weather improves memory; judicial sentences tend to be more severe when it is hot outside; and stock market performance is affected by sunshine.

Another source of random variability in judgment is the order in which cases are examined. When a person is considering a case, the decisions that immediately preceded it serve as an implicit frame of reference. Restoring a form of balance after a streak, asylum judges in the United States are 19% less likely to grant asylum to an applicant when the previous two cases were approved. Just as our muscles never execute exactly the same gesture, our neurons never operate in exactly the same way.

Group decisions can go in all sorts of directions, depending in part on factors that should be irrelevant. Who speaks first, last, or with confidence, who is wearing black, who is seated next to whom, or who smiles or frowns or gestures at the right moment—all these factors, and many more—affect outcomes. Groups dynamics can add noise, too. Popularity self-reinforces. An initiative might go nowhere, simply because those who supported it did not voice their opinion.

Independence is a prerequisite for the wisdom of crowds, but social influences reduce group diversity. We see events in informational cascades. We are able to see history only as it was actually run. Because some do not want to look disagreeable, it can lead to confidence about, and unanimous support for, a judgment that is quite wrong. Group polarization can occur when internal discussions create greater confidence, unity, and extremism. Deliberating juries are far noisier than statistical juries—a clear reflection of social influence noise. Deliberation increases noise. After people talk with one another, they typically end up at a more extreme position. In business, government, and everywhere else, cascades and polarization can lead to wide disparities.

In ranked accuracy of predictions—those made by professionals, by machines, and by simple rules—professionals come last. A measure that captures this is the percent concordant (PC) using the correlation coefficient (r), which varies between 0 and 1 (max): r .3 => PC 60% ; r .4 => PC 60 %; r .6 => PC 71%. The PC for foot length and height in adult men is 71%, with r ~ .60.

Many are interested in forecasting people’s future performance on the job—their own and others. Based on current rankings in factors of leadership, communication, interpersonal skills, technical skills, and motivation, future performance evaluations of 2 individuals after 2 years were predicted. This informal approach is known as clinical judgment. It a quick computation, based on intuition, to come up with a judgment.

Simple, linear models base on simple mechanical rules are generally superior to human judgment. Professionals are distressingly weak in what they often see as their unique strength: the ability to integrate information. In mechanical prediction models the same rule is applied to all cases. Clinical intuition often violates this rule—only .15 correlation (PC = 55%) vs .32 (PC = 60%) with a mechanical rule. The illusion of validity is found wherever predictive judgments are made, because of a common failure to distinguish between stages of the prediction task: evaluating cases on the evidence available and predicting actual outcomes. The evidence for the advantage of the mechanical approach to combining inputs was “massive and persistent,” including diagnosis of jaundice, fitness for military service, and marital satisfaction. The mechanical is faster and cheaper than clinical judgment. The noisiness of people is their critical weakness. Removing noise from your judgments will always improve your predictive accuracy. Replacing you with a model of you eliminates your subtlety and your pattern noise.

All mechanical approaches are noise-free and far superior to clinical judgments. This includes defendant’s liability to jump bail, since the model uses age (older people are lower flight risks) and the number of past court dates missed. Its superiority applies to determining the severity of a tumor from mammographic data, diagnosing heart disease, and predicting credit risk. This includes the judicial problem: recidivism prediction, since the model uses predictors (age and the number of previous convictions). The appeal of rules is that they are transparent and easy to apply. They are capable of discovering many more details and special circumstances than humans. If they were implemented, crime rates could be reduced by up to 24% because the people behind bars would be the ones most likely to recidivate. The number of people detained could be reduced by up to 42%.

Mechanical prediction techniques, represent significant improvements on human judgment. Simplicity and noiselessness are sizable advantages. But their use remains limited because many experts ignore the clinical-versus-mechanical debate, preferring to trust their judgment. They regard the idea of algorithmic decision making as dehumanizing and an abdication of their responsibility. People are willing to give an algorithm a chance but stop trusting it as soon as they see that it also makes mistakes. As humans, we are keenly aware that we make mistakes, but that is a privilege we are not prepared to share. Because of this intuitive expectation, however, people are likely to distrust algorithms and keep using their judgment, even when this choice produces demonstrably inferior results. Fortunately, much of what makes rules and algorithms better can be replicated in human judgment.

What makes the internal signal important—and misleading—is that it is construed not as a feeling but as a belief. But confidence is no guarantee of accuracy. What is the probability someone you thought had more potential turned out to be the higher performer? On average, we achieve a predictive correlation of .28 (PC = 59%). Intractable uncertainty includes everything that cannot be known about the outcome that you are trying to predict. Much about the candidates is not known when you make your judgment. Unknowns are not problems of bias or noise in your judgment; they are characteristics of the task.
Overconfidence is one of the best-documented cognitive biases. Wherever there is prediction, there is ignorance, and more of it than you think. Experts tell great stories. They can analyze a situation, paint a compelling picture of how it will evolve, and refute, with great confidence. But unforeseeable events will occur, and the consequences of these events are also unforeseeable. As a result, objective ignorance accumulates steadily the further you look into the future. Experts deserve criticism for attempting an impossible task and for believing they can succeed in it.

The future is unpredictable. However, the obviousness of this fact is matched only by the regularity with which it is ignored. People often mistake their subjective sense of confidence for predictive validity. They also believe in the predictability of events that are in fact unpredictable, implicitly denying the reality of uncertainty. This attitude amounts to a denial of ignorance. When they listen to their gut, decision makers hear the internal signal and feel the emotional reward it brings. Giving up the emotional reward of intuitive certainty is not easy. In something as important as decisions about a large mass of predictive and more of it than you think. Experts tell great stories. They can analyze a situation, paint a compelling picture of how it will evolve, and refute, with great confidence. But unforeseeable events will occur, and the consequences of these events are also unforeseeable. As a result, objective ignorance accumulates steadily the further you look into the future. Experts deserve criticism for attempting an impossible task and for believing they can succeed in it.

The future is unpredictable. However, the obviousness of this fact is matched only by the regularity with which it is ignored. People often mistake their subjective sense of confidence for predictive validity. They also believe in the predictability of events that are in fact unpredictable, implicitly denying the reality of uncertainty. This attitude amounts to a denial of ignorance. When they listen to their gut, decision makers hear the internal signal and feel the emotional reward it brings. Giving up the emotional reward of intuitive certainty is not easy. In something as important as decisions about whom to hire, however, an increase in validity has a great deal of value.

Giving up the emotional reward of the internal signal is a high price to pay when the alternative is a mechanical process that does not claim certainty. As long as algorithms are not nearly perfect, human judgment will not be replaced. That is why it must be improved.

The prevalent and misguided sense is that unpredictable events can be understood. The large mass of predictive information is not enough for the prediction of single events and the prediction of groups of events is even more limited. To understand is to describe a causal chain. Correlation does not imply causation, but causation implies correlation. Causal thinking creates stories in which specific events, people, and objects affect one another. Whatever the event, once it has happened, causal thinking makes it feel entirely explainable, indeed predictable. The process of understanding reality is backward-looking. This is what we mean by understanding a story, and this is what makes reality appear predictable—in hindsight.

Genuine surprise occurs when routine hindsight fails. This continuous causal interpretation of reality is how we “understand” the world. Even when subjective uncertainty exists for a while, memories of it are largely erased when the uncertainty is resolved. There are two ways of thinking about events: statistical and causal. The flow of events in the valley of the normal requires little mental work. In contrast, statistical thinking is effortful. Relying on causal thinking about a single case cases errors. The statistical view, the outside view, avoids them.

Conclusion bias, or pre judgment occurs if we start the process of judgment with an inclination to a particular conclusion. We tend to collect and interpret evidence selectively to favor a judgment that we already believe or wish to be true. People may also determine what they think by consulting their feelings. We will be inclined to accept any argument that appears to support it, even when the reasoning is wrong. Anchoring is a very robust effect and often deliberately used in negotiations—the effect that an arbitrary number has on those who make a quantitative judgment. The recipient of the anchor is involuntarily drawn to think of ways your offer could be reasonable. People always attempt to make sense of what they hear. Excessive coherence: we form coherent impressions quickly and are slow to change them. The halo effect of initial impressions is a serious problem in hiring decisions. We tend to jump to conclusions, then stick to them.

Substituting one question for another inevitably causes errors. Predictions that match the evidence are hard to resist. Intuitions must be endorsed by the more reflective System 2 before they become beliefs. The outside view is a corrective for intuitive predictions of all kinds.

A respect-expert’s credibility depends on the respect of their students, peers, or clients. Part of this arises from the existence of shared norms, or professional doctrine. General mental ability (GMA, the term now used in preference to intelligence quotient, or IQ) predicts both occupational level attained and performance within one’s chosen occupation and does so better than any other ability, trait, or disposition and better than job experience. For jobs of high complexity, the correlations between standardized test scores and job performance are in the .50 range (PC = 67%).

Cognitive reflection test (CRT) measures propensity to rely on slow, System 2 thinking rather than on fast, System 1 thinking. Actively open-minded thinking is a teachable skill.

Automatic enrollment in pension plans is designed to overcome inertia, procrastination, and optimistic bias, ensuring that employees save for retirement unless they deliberately opt out. If healthy foods are put in prominent places, more people buy them. Teaching people to avoid biases is hard. The lack of awareness is itself the bias blind spot. Detecting biases is useless if we aren’t committed to fighting them. Bias is error we can often see and explain, so that we can combat it. Noise, on the other hand, is unpredictable error that we can’t easily see or explain. That is why we so often neglect it—even when it causes grave damage. Noise reduction decision hygiene adopts techniques that reduce noise without knowing which underlying errors you are avoiding. Hygiene measures can be tedious and their benefits invisible. They are invaluable but thankless, preventing many errors without knowing which ones.

Decision hygiene is a strategy that can apply to all environments: a tight control over the flow of information used to make judgments. Latent fingerprints are often partial, unclear, smudged, or otherwise distorted. Deciding whether they match a suspect’s exemplar prints requires expert judgment. Fingerprints were accepted—at least until the advent of DNA analysis—as the most credible form of evidence. It remained unchallenged for so long in part because of the difficulty in proving it wrong. Occasion noise: the variability between repeated judgments of the same experts. Biasability: expert fingerprint examiners made decisions on the basis of context, rather than the print. When the latent print is accompanied by a target exemplar print, the examiners observe fewer details (called minutiae) than when they see the latent print alone—confirmation bias. If fingerprint examiners can be biased, so can experts in other fields. Misapplication of forensic science was a contributing cause in 45% of cases of the 350 exonerations obtained by the Innocence Project.
Decision hygiene strategy applies to many domains: sequence information to limit the formation of premature intuitions. More information is not always better, especially if it biases judgments by leading the judge to form a premature intuition. Linear sequential unmasking: document analysis of a latent fingerprint before they look at exemplar fingerprints to decide whether they are a match. Our ignorance of the possibility of error in fingerprint identification shows how our confidence in expert human judgment can sometimes be exaggerated. The ability to mitigate these shortcomings through simple process changes should encourage those who care about improving the quality of decisions.

Official agencies show unrealistic optimism in their budget forecasts, economic growth and deficits. Forecasters tend to be overconfident. CFOs are far too confident in the precision of their forecasts. Good calibration is one requirement for good forecasting.

The BIN (bias, information, and noise) model intervenes to minimize each source of error. Apgar’s score is a guideline for diagnosis on the distress of a newborn baby. The “backronym” for Apgar’s name: appearance (skin color), pulse (heart rate), grimace (reflexes), activity (muscle tone), and respiration (breathing rate and effort). Each of these is given a score of 0, 1, or 2. The highest score is 10, which is rare. A score of 7 or above indicates good health. Because the judgment is decomposed into individual elements, each of which is straightforward to assess, practitioners with even modest training are unlikely to disagree a great deal—and hence Apgar scoring produces little noise. The Apgar score exemplifies how guidelines work and why they reduce noise, as they decompose a complex decision into a number of easier subjudgments on predefined dimensions. It focuses the health professional on the 5 that matter. Clear description of how to evaluate each cue of the Apgar score specifies how to weight the predictors.

Similarly, the Centor score guides diagnosis of strep throat. Breast Imaging Reporting and Data System (BI-RADS). Noise reduction has been a priority for the psychiatric community since at least the 1940s. In medicine in general, guidelines have been highly successful in reducing both bias and noise.

Everyone hates performance reviews and knows they are subject to both bias and noise. Studies based on 360-degree performance reviews show that 70 to 80% of variance in the ratings is system noise. They also have undesirable side effects on morale and teamwork. A system that depends on relative evaluations is appropriate only if an organization cares about relative performance. A small but growing number of companies are considering the radical option of eliminating evaluation systems altogether. These companies focus on developmental, future-oriented feedback rather than on evaluative, backward-looking assessments to ensure a common frame of reference.

We can find logic in perfectly meaningless answers. Structured complex judgment is defined by 3 principles: decomposition, independence, and delayed holistic judgment. Independence requires that information on each assessment be collected independently. Delayed holistic judgments do not exclude intuition, but delay it. Google allows judgment and intuition in its decision-making process only after all the evidence has been collected and analyzed. Diverse results of the various steps will not make the decision easier, but they will make it better.

Comparative judgments become much easier in the context of a recurring decision. Create a case scale defined by anchor cases. Content is specific; process is generic. Using intuition and judgment is fun; following process is not. It promotes challenge and debate, not the stifling consensus that characterizes bureaucracies. When people decompose a problem into its component parts, their judgments need not be mechanical.

The costs of eliminating noise might exceed the benefits. A noise-free system might freeze existing values. Whether an objection is convincing depends on the particular noise-reduction strategy to which it applies. We must compare the benefits of noise reduction with the costs—1 reason noise audits are so important. Try to devise a better noise-reduction strategy. They might know that their approach is noisy, but if it ensures that people feel treated with respect and that someone has listened to them, they might embrace it anyway. We can do a great deal to reduce noise or even eliminate it while still designing processes to allow values to evolve.

A rule might say that no one may exceed a numerical speed limit, that workers may not be exposed to carcinogens, or that all prescription drugs must come with specific warnings. By contrast, a standard might say that people must drive “prudently,” that employers must provide safe workplaces “to the extent feasible,” or that in deciding whether to offer warnings for prescription drugs, companies must act “reasonably.” Rules eliminate discretion by those who apply them. Standards mean numerous judgments but might be far easier to generate than rules. They may be the best that leaders can do.

Those who decide between rules and standards must focus on the problem of noise, bias, or both. The choice depends on the costs of decisions and the costs of errors. With rules, the costs of decisions are typically much lower. But before a rule is put in place, someone has to decide what it is and producing a rule can be hard. The pervasive question is whether agents are knowledgeable and reliable, and whether they practice decision hygiene. Organizations all over the world see bias as a villain. They are right. They do not see noise that way. They should.

[[The way society wrestles with the judgment process is the fascinating subject of this book.] Noise audits suggest that respected professionals maintain an illusion of agreement while in fact disagreeing in their daily professional judgments. Underwriters who heard about the noise audit and accepted its validity never believed that its conclusions applied to them personally. Judgment is measurement in which the instrument is a human mind. Professionals are distressingly weak in what they often see as their unique strength: the ability to integrate information. The illusion of validity is found wherever predictive judgments are made. All mechanical approaches are noise-free and far superior to clinical judgments. Rules are transparent and easy to apply. Professional regard the idea of algorithmic decision making as dehumanizing and an abdication of their responsibility. Confidence is no guarantee of accuracy. Unknowns are characteristics of the task of prediction.]}
Overconfidence is one of the best-documented cognitive biases. The future is unpredictable. However, the obviousness of this fact is matched only by the regularity with which it is ignored. People often mistake their sense of confidence for an indication of predictive validity. When they listen to their gut, decision makers hear the internal signal and feel the emotional reward it brings. Human judgment will not be replaced, so it must be improved. Causation implies correlation. Whatever the event, once it has happened, causal thinking makes it feel entirely explainable, indeed predictable. This continuous hindsight causal interpretation of reality is how we “understand” the world.