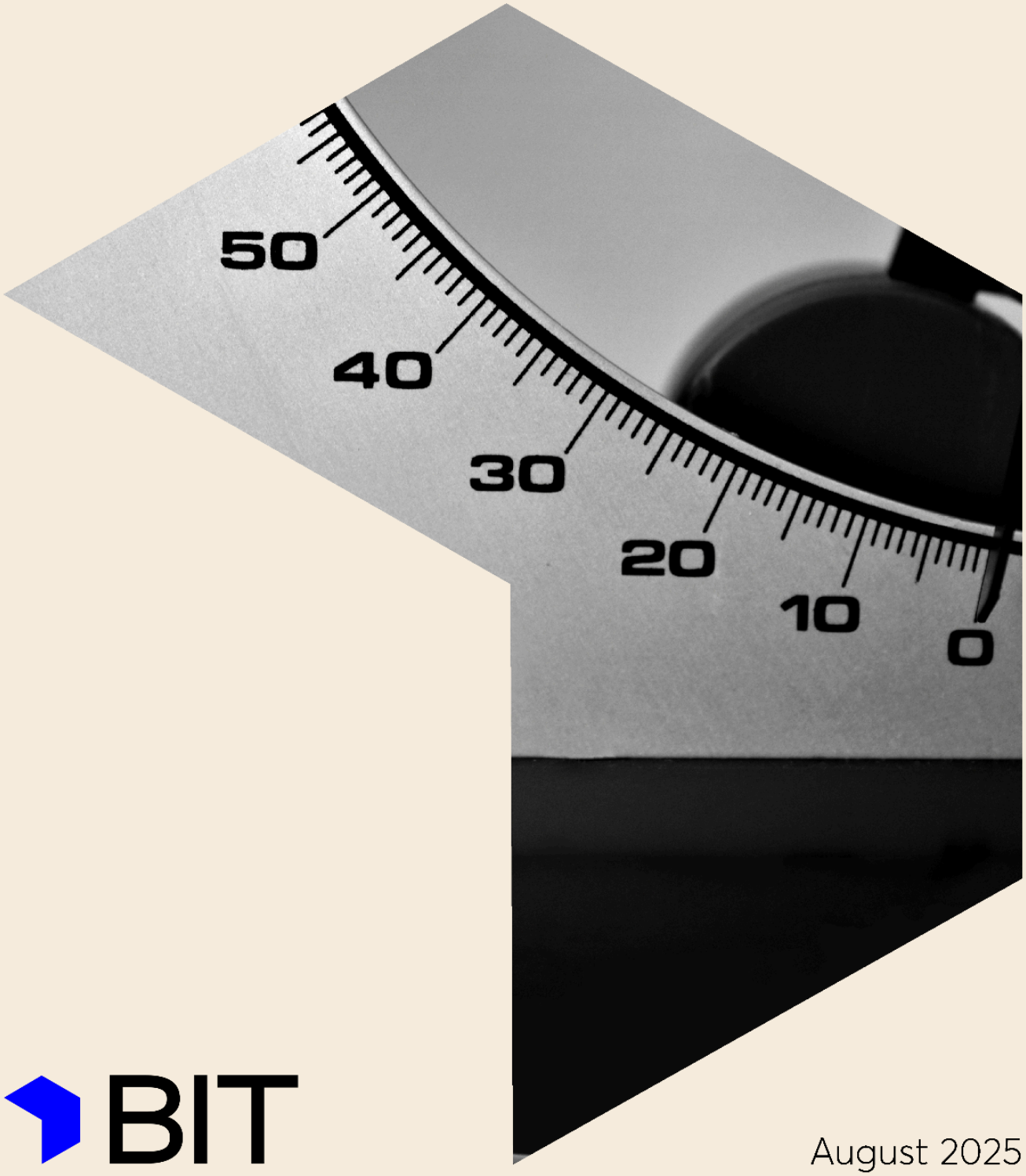
Stay calibrated

A practical guide to debiasing decision-making



BIT

August 2025

BIT is a global research and innovation consultancy which combines a deep understanding of human behaviour with evidence-led problem solving to improve people's lives. We work with all levels of government, nonprofits and the private sector, applying behavioural science expertise with robust evaluation and data to help clients achieve their goals.

Find out more at bi.team

If you'd like this publication in an alternative format such as Braille or large print, please contact us at: info@bi.team

Authors
Dr Mark Egan

Click the **BIT** logo on any page to return to the Contents

Contents

Executive summary

There is now 50 years of research documenting how cognitive biases distort human judgement and lead to worse decisions.

But merely identifying hundreds of types of systematic thinking errors doesn't reliably help people make better real-world decisions.

Instead, behavioural science needs a simple, outcome-focused benchmark we can use to measure the impact of interventions to debias and improve decision-making.

This paper proposes one: **calibration**.

To be well-calibrated means your confidence in your judgement aligns with accuracy. If you say you're 80% sure, you should be right roughly 80% of the time. You are not overconfident or underconfident, but instead lie in between. Although calibration is a simple measure of cognitively-unbiased judgement, it is relatively unknown as a concept both to ordinary people and many behavioural scientists.

Being well-calibrated doesn't require you to have perfect logic or encyclopaedic knowledge. It also doesn't guarantee that you will always make a good decision. But, it matters because:

- **Overconfidence is rampant** in many domains from surgery to strategy – people are more sure in their judgement than they should be, and this leads to bad outcomes.
- **Well-calibrated thinkers are rare** but superforecasters, bridge players, and weather forecasters show that it's possible, and give us hints about what types of environments foster and reward it.
- **Calibration is trainable.** Like physical fitness, it can be built through habit, feedback, and repetition, without requiring innate brilliance.

The goal of calibration isn't to *eliminate* cognitive bias, but to enable good decisions despite it. This report offers:

- A clear framework for understanding and measuring calibration
- Practical tips for individuals, teams, and leaders to achieve it
- Design principles for building calibration into organisational systems

1. How to think about thinking

Cognitive bias research is over 50 years old.

In that time there have been thousands of studies showing how these 'systematic thinking errors' distort judgement and decision-making in domains as diverse as psychology, economics, medicine, business, law, public policy, climate change, and now, artificial intelligence.

Once seen purely as flaws in human reasoning, cognitive biases are now recognised as ecologically adaptive – mental shortcuts that often serve us well but which can misfire in modern environments. Take negativity bias. In an ancestral environment, over-reacting to danger signals like rustling in the bushes could save your life. In a modern workspace, the same tendency can turn receiving critical feedback into feelings of unproductive anxiety.

The existence of these cognitive shortcuts, of which around 200 have now been documented, form one of the most empirically robust findings in the social sciences. The award of the Economics Nobel to Daniel Kahneman in 2002 and Richard Thaler in 2017, both pioneers in the field, further cemented the legitimacy and influence of this line of research. But although the field is now mature, it has not slowed down – recent studies continue to identify biases harming the success of surgical procedures (Armstrong et al., 2023), intelligence analysis (Belton & Dhami, 2021), and policymaking generally (Hallsworth et al., 2018).

So, cognitive biases are common and capable of undermining decision quality in virtually every field of human activity. What to do about it? Most responses fall into two categories. One is to *change the thinker* – through training, critical thinking prompts, or reflective techniques like "consider the opposite" to reduce confirmation bias. The other is to *change the environment* – by redesigning systems, tweaking incentives, or outsourcing decisions to algorithms in the hope that they prove less biased than the humans.

Both strategies have value. But they reveal two problems.

First, we rarely agree what 'good' looks like. If biased judgement is the problem, what does unbiased judgement look like? Definitions and metrics vary. Some focus on *logical consistency* – does a person's reasoning follow valid rules? Others

emphasise *coherence with expected utility theory* – does the person maximise outcomes based on preferences and probabilities? *Procedural correctness* – did the person follow a checklist, or consult an external view? These are all useful in context, but none offer a universal benchmark for good judgement.

Second, we focus too much on faulty inputs instead of achieving good results. Imagine you go to the optician to get help for blurry vision. The optician lists all the ways your eyes might be faulty. They run tests, show you diagrams, explain the structure of the retina. They then end the appointment by airily telling you to try squinting to bring reality into sharper focus, as they turn back to continue studying the cause of the problem in greater detail. You would not find this very helpful. A good optician doesn't focus just on diagnosis. They give you a pair of glasses and you walk out seeing better even if *the underlying fault is still there*. That's how we should think about improving decision-making. The goal shouldn't be to name and tame all 200 cognitive biases, or suggest people just 'try harder' to avoid them, or to believe we can prevent them from occurring altogether. It's to help people think clearly enough to make good decisions despite these cognitive flaws. We have identified enough cognitive biases, we need to focus more on how to practically address them.

To achieve that we need two things:

1. **A clear, consistent way to tell whether someone's judgement is *working*** – something as simple as reading the letters on the optician's chart.
2. **A commitment to work backwards from improving it.** Less focus on hunting for the 201st cognitive bias, more on ensuring our solutions actually improve people's decision-making.

This report offers my answer to what cognitively unbiased judgement looks like and provides practical strategies to achieve it. The ideas are grounded in academic research. The solutions are shaped by a decade of applied consultancy work with hundreds of organisations and thousands of professionals.

What follows is not a new taxonomy of biases, but a simple organising principle and a toolkit for helping people think more clearly and act more wisely.

It can be summarised in two words: **stay calibrated.**

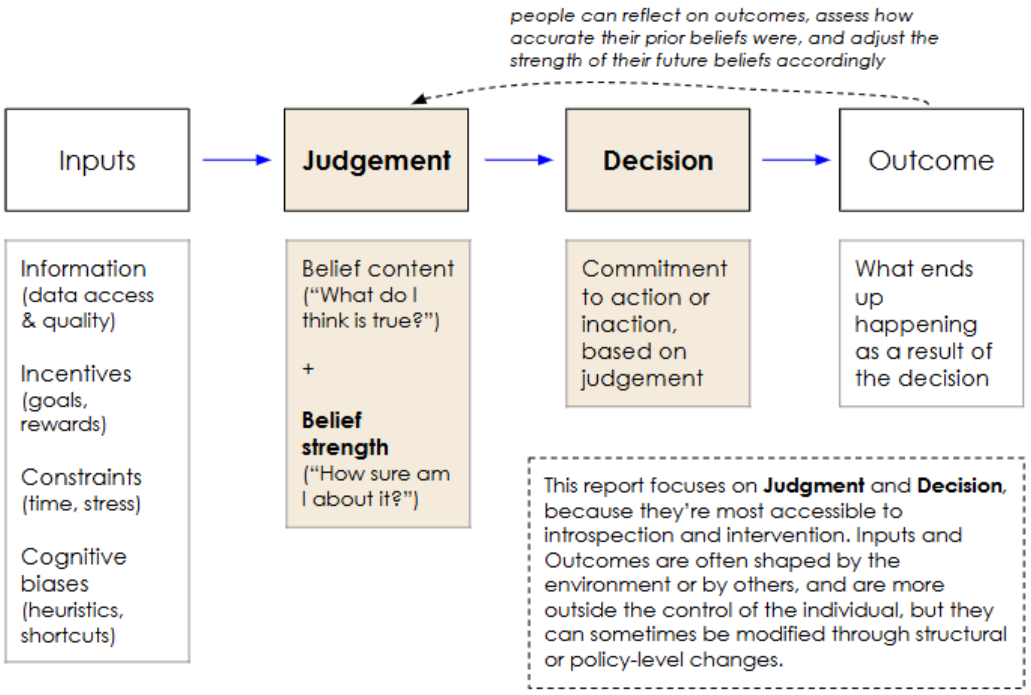
2. A practical model of decision-making

Behavioural scientists should try to help people make better judgements and decisions.

'Judgement' means assessments and evaluations: "What do I think is true?". Decisions mean a commitment to act: "What should I do next?" Judgements usually come first and shape the following decision. Imagine a manager deciding whether to promote someone. They first make a judgement: "Do I believe they're ready?" Confirmation bias might lead the manager to selectively focus on critical feedback that supports a pre-existing doubt about the employee, while ignoring evidence of recent improvement. Anchoring might play a role if the manager's view is still overly shaped by an early stumble the employee made, even if their performance since then has improved. The decision follows: "Should I promote them?" That act – making the call – depends on the underlying judgement. Here status quo bias might make the manager default to inaction – not because it's the best decision, but because it feels safer than change. Both judgement and decision can be affected by cognitive bias but for simplicity, we will mostly use the language of 'decisions' and 'decision-making' from here on – because what people ultimately do is what matters most.

We want to help people make *better* decisions – not necessarily perfect ones. This is a humble ambition, because it is built on a recognition that cognitive biases are only one part of the decision-making picture, as per Figure 1.

Figure 1. A practical model of decision-making.



A range of inputs, such as information quality, incentives, time pressure, as well as cognitive biases, shape the judgments people form and how sure they are about them. Those judgements then inform the decisions they make: the actions they take or avoid. Those decisions lead to outcomes – improving these is what we ultimately care about. People may then observe the outcome of their decisions and use them to update their future process of forming judgements and making decisions. Imagine someone trying a new route to work. They judge that it will be quicker than their usual one and feel fairly confident in that belief, so they decide to take it. It ends up taking longer than expected. The next day, they revise their belief – lowering their confidence in that route and checking traffic first.

Any of the inputs can lead to poor decisions on their own, and they can also interact. Confirmation bias might reduce your willingness to seek new information, leaving you excessively confident in your decision-making. Strong incentives, like financial rewards for achieving a sales target, might increase stress and reduce the quality of your judgement even without cognitive biases coming into play.

Even if we could eliminate cognitive biases entirely, judgment quality would still depend on the other inputs. This means someone could be completely free of cognitive bias and still make decisions which lead to bad outcomes. Consider:

- A policymaker receives flawed data about a new virus. They weigh the evidence calmly, avoid availability bias or political framing, and make a well-reasoned decision. But the data are wrong, so the decision leads to disaster.
- A doctor methodically evaluates a patient's symptoms with the aid of a checklist designed to mitigate overconfidence and avoid representativeness bias. But they're working in a very busy ward, and do not spend as much time as they should on the assessment, and end up missing a rare but critical diagnosis.
- A product manager makes a fair and unbiased evaluation of two competing designs. But their incentive structure rewards short-term engagement over long-term user trust, so they choose the design that ultimately undermines the product's success.

So we can't control all the inputs that shape decisions, guarantee good outcomes or even eliminate cognitive biases. I recommend accepting this reality with a sigh of relief. Behavioural scientists are not on the hook for guaranteeing perfect decision-making. Phew.

This means that in our model, we are going to focus on the *Judgement and Decision* stages, because these are the most amenable to practical behaviour change.

Within those stages, the single most important thing about the model is its emphasis on the **strength of belief** (or confidence) that underpins a judgement, instead of just the content of the belief. So not just "I want to promote this person", but "I'm 90% sure that promoting this person is a good idea".

Confidence has been extensively studied in metacognition and forecasting research (Moore et al., 2017). But many models of judgment still focus primarily on what people believe, not how sure they are in those beliefs. This distinction matters because misplaced confidence is often where the harm arises. A misjudgment held with doubt is usually harmless. I may vaguely think my meeting is at 3pm, but I know I'm not sure so I double-check my calendar and learn it's at 2pm – no harm done.

But a misjudgment held with high confidence can block correction, mislead others, and lead to seriously bad outcomes. Think of someone who bets a small amount of money on a football game because they have a vague feeling their team might win, vs someone who bets a very large amount because they feel certain about it. The team loses badly, and the second person who was 'confidently wrong' loses much more than the first. Strength of belief matters.

Belief strength or confidence is what we will focus on fine-tuning. Earlier we talked about opticians. By giving you glasses, opticians can improve your vision without actually fixing the underlying fault in your eyes. Similarly, learning to fine-tune confidence in your decisions will give you 'glasses for your brain' – a method for thinking clearly even despite the continued existence of cognitive biases.

3. Try to be well-calibrated

3.1 Well-calibrated judgement is cognitively unbiased

Cognitive biases distort our perceptions of the world. Confirmation bias leads us to seek out information that supports what we already believe, while ignoring or dismissing what contradicts it. Availability bias makes dramatic or recent events feel more common than they really are. Anchoring tethers our estimates to irrelevant starting points. And so on.

It follows then that unbiased decision-making is *undistorted*. Like looking through a clear window and seeing the world as it is, not as we wish or fear it to be.

Here's how we measure that:

1. Test a person's knowledge. About anything, but general knowledge and trivia are a good place to start.
2. Ask them how sure they are about their answers.
3. Compare the two.

This gives you two data points:

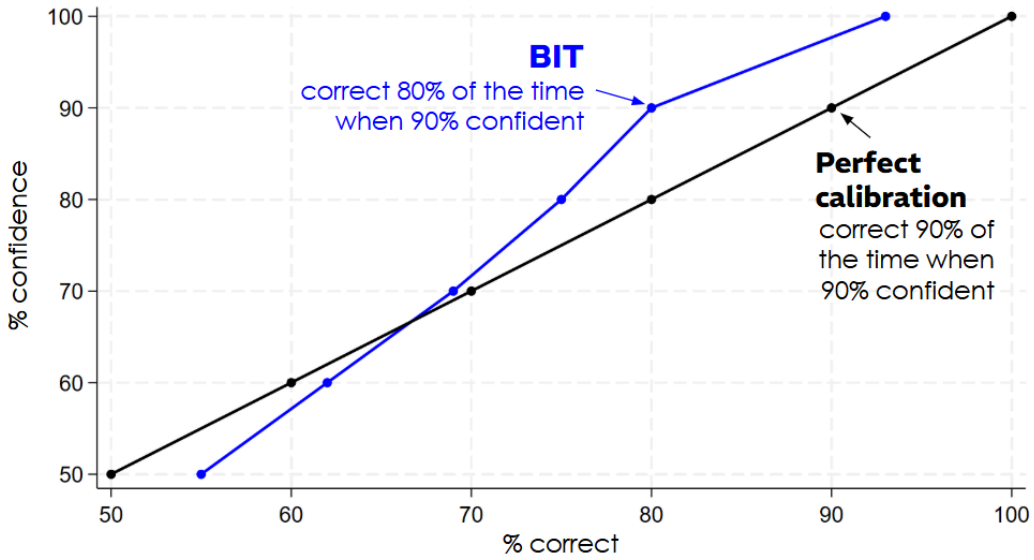
1. Confidence, how sure a person is in their judgements, and
2. Accuracy, how correct they turn out to be about them.

When you compare the two, you will find that people fall into one of three categories:

1. Overconfident, a person who is more confident than correct.
2. Underconfident, a person who is more correct than confident.
3. Well-calibrated, the group in between, whose confidence and accuracy closely track each other.

Figure 2 shows what being well-calibrated looks like – when you're 90% sure about something, you're right 90% of the time. That may seem like a low bar. But in practice it is not always achieved even by the behavioural science experts at BIT.

Figure 2. What calibration looks like, in theory & practice.



BIT data collected internally over 2017-22.

Why does this simple measurement matter so much? Calibration matters because cognitive biases don't just make us wrong – they make us *confidently* wrong. Confirmation bias doesn't just lead us to ignore contradictory evidence; it makes us feel more certain about conclusions drawn from incomplete information. Anchoring doesn't just skew our estimates; it makes us inappropriately confident in those skewed numbers.

This is what makes poor calibration – especially overconfidence – such a useful diagnostic. It often signals that biases are systematically distorting someone's judgment. Well-calibrated people have learned to recognise the limits of their knowledge and adjust their confidence accordingly. Their decisions may still be imperfect, but they're not amplified by misplaced certainty. Calibration is not an unknown concept in academia. But its full potential has been overlooked by two major research traditions – see Box 1.

Box 1. Calibration sits between two research worlds

Calibration – confidence aligned with accuracy – bridges two major areas of study in human decision-making.

1. **Cognitive bias research** focuses on inputs: faulty reasoning driven by heuristics, perceptual distortions, and belief errors. Calibration rarely features, and when it does it's often buried as just one more bias in a list of 200.
2. **Forecasting research** focuses on outcomes. Here, calibration is foundational – but it ends up getting overlooked, because great forecasting requires both calibration and resolution (ie, the ability to make very accurate predictions across a range of domains)

Both traditions acknowledge calibration. Neither puts it at the centre.

This report does. Calibration is the outcome measure we should prioritise when trying to improve everyday judgment and decision-making.

Consider weightlifting. Progress requires two things: good form and adding weight. Good form means moving the bar correctly. Adding weight means increasing load. Form comes first, and is something everyone can learn. Not everyone can lift heavy.

Calibration is good form for thinking. You can have clean form whether you're lifting 5kg or 100kg – and whether you're getting things right 20%, 50%, or 80% of the time. Calibration shows whether your cognitive movements are aligned, regardless of difficulty or domain.

Forecasting demands good form and heavy loads – high calibration and high resolution. That's useful, but unrealistic for most people. Most people who go to the gym aren't going to become powerlifters.

But **good form is achievable**. It's the part of expert judgement that everyone can master.

Calibration has clear boundaries. It won't fix problems with bad data or misaligned incentives. But it does something important: it prevents cognitive biases from compounding other problems. Well-calibrated individuals handle flawed information differently – they recognise uncertainty, hedge appropriately, and stay open to

updates. Miscalibrated people treat uncertain inputs as certain, making bad situations worse.

Understanding this clarifies overconfidence's special role. It's often listed as just another bias among 200, but that misses something crucial. Overconfidence isn't a distinct bias operating in parallel with others – it's what happens when bias-distorted beliefs meet misplaced certainty. Confirmation bias shapes what you believe; overconfidence determines how sure you feel about it. Treating it as just another entry in the catalogue obscures its role as a summary failure of judgement. Overconfidence isn't a sibling to other biases – it's their offspring.

3.2 Overconfidence is much more common than good calibration or underconfidence

When you measure calibration, overconfidence tends to be the norm.

A well-known example is the 1980 Swedish study which found that 80-90% of people considered themselves to be above-average drivers; admirable self-belief but statistically impossible. More recently, Alba & Hutchinson (2000) found an average 'overconfidence gap' of 15 percentage points (eg, 70% confident vs 55% correct) across several hundred studies about cognitive calibration and across topics such as general knowledge, memory for events, predictions about the future, and assessments of one's own abilities.

Subsequent reviews have confirmed the same pattern across more applied domains. Koehler, Brenner, and Griffin (2002) examined over 100 studies and found consistent overconfidence among doctors, stock analysts, and sports commentators – though economists and weather forecasters tended to be better calibrated. Sanchez and Dunning (2023) reviewed evidence from psychology, economics, medicine, and management. They concluded that experts were often too certain their answers were right, and often overrated their performance. These findings were stronger when 'expertise' was defined by credentials or job titles rather than tested knowledge. Interestingly, experts sometimes displayed *underconfidence* when comparing themselves to others, showing that calibration failures can cut both ways. Finally, recent studies using representative samples of the general public have found 80–90% of people, from both WEIRD and non-WEIRD populations, were overconfident when tested on general knowledge (Egan, 2024; Egan et al., 2025).

3.3 Characteristics of the well-calibrated

A small set of studies has identified groups with excellent calibration. Keren (1987) found that bridge players had nearly perfect calibration when making judgements within that domain of expertise. Mellers et al. (2016) found that 'superforecasters' had essentially perfect calibration over a two year period, such that they were capable of accurately distinguishing between fine-grained probabilities (eg, when they forecast events with 68% likelihood, they really are more likely to occur than events predicted with 63% likelihood). Psychometrically, these individuals tended to have high crystallised intelligence, a high need for cognition, high open-mindedness, and 'scope-sensitivity' (ie, they are good at fine-tuning assessments based on new information). A greater tendency towards analytic (vs intuitive) thinking and active open-mindedness was also found to be predictive of better calibration in a multi-country study (Egan et al., 2025).

Morgan (2014) illustrates how calibration can depend on the feedback environment experts operate in, rather than just their individual characteristics. He contrasts two groups: doctors diagnosing pneumonia, and weather forecasters predicting rain. In a study of over 1,500 diagnoses, nine physicians were asked to estimate whether patients had pneumonia and how confident they were. Their confidence bore little resemblance to reality: when doctors were 80% sure, they were right only about 20% of the time – a striking case of poor calibration. By contrast, US weather forecasters were almost perfectly calibrated. When they predicted an 80% chance of rain, it rained 80% of the time. This difference is driven by feedback rather than talent. The weather forecasters made thousands of predictions and got timely, accurate feedback – often within hours. The doctors worked in slower-feedback environments where it typically took much longer to learn whether a judgement was correct.

Calibration can be improved with training. In a multi-year geopolitical forecasting tournament, Moore et al. (2017) found that just one hour of structured training halved overconfidence, with the effect persisting into the following year. The training covered four simple habits: taking the outside view (drawing on base rates), averaging across judgments (to reduce noise), using basic models where possible, and being alert to cognitive biases. Kelly and Mandel (2024) found that calibration training helped overconfident intelligence analysts improve on estimation tasks – but made underconfident analysts even more underconfident for some tasks. This suggests that while calibration training can help, it risks overcorrecting unless it's

tailored to the task and the starting bias. Gutierrez de Blume (2022) reviewed 56 studies and found that teaching learning strategies reliably improves calibration, with a moderate effect size ($g \approx 0.57$). The most effective interventions helped people reflect on their knowledge: using techniques like prediction-postdiction, self-explanation, and external benchmarks. When learners were trained to monitor their own understanding more deliberately, their confidence started to track their accuracy more closely.

In summary, excellent calibration is most evident among people who develop domain-specific expertise in feedback-rich environments which provide many opportunities for repetitive, iterative learning. Being well-calibrated is rare, but not impossible. It is like maintaining physical fitness – it requires relatively uncomplicated habits that nonetheless need consistent application over time. For some people those habits come naturally, for others they require more work. But they are learnable.

4. Practical tips for becoming well-calibrated

Calibration is good form for thinking. Like any form, it can improve with feedback and deliberate practice. Some individuals – like superforecasters or expert bridge players – maintain remarkably well-calibrated judgement. Their edge isn't raw brainpower alone. It came from habits, environments, and feedback loops that reinforce clear thinking. Those habits include pausing to estimate confidence, considering alternative scenarios ("If X happens, then…"), and deliberately seeking evidence that challenges your current view.

This section sets out practical ways to encourage and sustain calibration. These tips are drawn from experience working with teams and organisations to improve decisions, and can be deployed at multiple levels – from individuals to leaders to system designers. The core principle is the same throughout: treat calibration not as a trait, but as a skill which can be measured, monitored, and improved.

4.1 As an individual and team member

4.1.1 Get in the habit of doing internal calibration checks

The first tip is foundational – think of calibration as a basic life maintenance activity akin to brushing your teeth or regular exercise. Just as musicians practise scales or gymnasts drill their technique, we should regularly test whether our confidence matches reality. If we're in the business of helping others think clearly, we need to be sure our own judgement is in shape.

Staying calibrated starts with a simple habit: pause and ask yourself "*How sure am I?*" before making a judgement. You don't need special tools or data – just a moment of honest reflection. That small step builds the muscle of metacognition. Box 2 provides a practical example.

Box 2. The Egg Test

Do you have eggs in your fridge?

Don't go and check – just answer yes or no.

Now pause and ask: *How sure am I?*

Totally sure? Pretty sure? A little bit? No idea?

Whatever your answer, hold onto that feeling of confidence – or doubt – and *then* go check.

If you were confidently wrong in either direction – "I'm sure I have eggs" and it turns out you don't, or "I definitely don't have eggs" and it turns out you do – that's a useful signal. It's telling you the strength of your belief doesn't necessarily align well with reality. At least not in this case.

Keep doing little calibration checks like this. Once the process starts to feel familiar, start putting numbers to it. Maybe "pretty sure" for you means 80% and "a little bit sure" means 20%. Get used to thinking in quantified degrees of certainty, and keep testing how well those correspond to reality.

Try it on everyday questions:

- Did I send that email?
- Will I make my train?
- Will I make this lift in the gym?

The goal is simple: the more confident you are about something, the more likely you should turn out to be right. When you're 100% sure, you should never be wrong.

The thing to watch for is major **miscalibration** – moments when you're very confident and wrong, or very doubtful but turn out to be right. Those tell you when your internal compass is off and give you a chance to adjust it.

Do this until it's as automatic and low-effort as brushing your teeth. Your epistemic hygiene matters just as much!

4.1.2 Use ThinkGroups to avoid groupthink

Group discussions are a core part of decision-making. They're how we generate ideas, share updates, and try to pressure-test proposals. But they're also prone to bias. Instead of improving judgement, traditional group settings sometimes make it worse. They can:

- Focus attention on what most people already know
- Push people toward more extreme versions of the majority view
- Give disproportionate weight to whoever speaks first
- Reinforce existing hierarchies
- Crowd out quieter voices

Worse still, they create false consensus that artificially inflates confidence. ThinkGroups are a simple way to disrupt these patterns. They flip groupthink on its head. Instead of rewarding fluency, confidence, or status, they make ideas compete on merit – shared anonymously, in parallel, with no way to defer to the loudest or most senior voice in the room.

How to run one

- Use a shared virtual document that everyone can work in at the same time (Google Docs or similar)
- Everyone joins anonymously
- Pose a few structured prompts or questions
- Run it live: ask people to contribute ideas over a set time (eg, 15–30 minutes)
- Start with idea generation only – no discussion yet
- Then open it up for commenting, refining, or building on others' suggestions
- Finish by asking participants to +1 or rank the ideas they find strongest

Why does this matter for calibration? Anonymous parallel input reveals the true distribution of opinions and concerns. Instead of inheriting the group's artificially confident consensus, you see the full range of views – including doubts and

alternatives that would never surface in traditional discussion. This gives you a much more realistic picture of the uncertainty surrounding a decision, allowing you to calibrate your confidence appropriately rather than just assuming everyone agrees.

It doesn't replace open discussion – but by improving the raw material and revealing the true landscape of belief, it produces a discussion which is more likely to lead to accurate assessments.

4.1.3 Run premortems to 'fail in advance'

If overconfidence is the default, then many projects will fail for predictable reasons: unrealistic expectations, untested assumptions, and failure to plan for things going wrong. A premortem is a simple way to counter that.

Instead of asking "What could go wrong?", you assume it has gone wrong – and then ask why. That small shift creates a big difference in how people think. It gives teams permission to challenge the plan, spot blind spots, and flag risks that might otherwise go unspoken.

How to run one:

- Set the scene: "It's 3/6/12 months from now. The project has failed. What went wrong?"
- Give everyone a few minutes to write down plausible reasons individually
- Collect and cluster the responses
- Discuss: Which risks are most likely? Which would matter most?
- Agree on a set of mitigations, contingency plans, or next steps

You can do this on paper, in a doc, or in conversation. What matters is the shift in mindset – from defending the plan to trying to break it. The link to calibration here is direct. A premortem systematically generates negative evidence that an overconfident team would otherwise ignore. When you start a project feeling 90% sure it will succeed, and the premortem surfaces ten plausible and serious risks, you are forced to update that initial belief. You can no longer honestly say you are 90% confident. Your understanding of the newly visible risks might drop your confidence

to 70% – you might decide it's still worth proceeding with the project, but now you know you need to manage it more carefully.

A good premortem surfaces the exact problems and cognitive biases that a real postmortem might reveal later – except this time, while there's still time to fix them.

4.2 As a leader

People in leadership roles can have an outsized impact on organisational culture, through their effect in managing individuals and teams, and their influence over the long-term progression and development of junior colleagues.

One most powerful lever is role-modelling. Try saying "I'm 80% sure" out loud. Admit when you don't know. Praise team members for highlighting flaws, flagging doubts, or revising their views. Show that epistemic humility is a strength, not a weakness.

Here are three additional ways leaders can help keep their teams cognitively calibrated.

4.2.1 Be more tolerant of underconfidence

The organisation where I work, BIT, tends to do well on calibration tests. The reason why is not what you might think. It is not the case that every individual at BIT is superbly well-calibrated – we do have our share of overconfident people. But, uniquely among the organisations I've worked with, BIT's culture appears to be unusually tolerant of underconfidence, and has a relatively high proportion of underconfident staff. This group in turn effectively counterbalances our overconfident colleagues, and on average we tend to be well-calibrated.

This balance may be a byproduct of BIT's mixed identity. As a research consultancy, we blend two professional cultures. The research side attracts people trained to second-guess themselves and hedge carefully – traits encouraged in academia. The consultancy side demands clarity, speed, and confident recommendations. While these instincts can be in tension, the result is a kind of functional equilibrium: evidence-weighers and decision-drivers in productive coexistence. Another reason may be demographic. Although the evidence on how overconfidence varies by gender is relatively mixed, when we examined the calibration data within BIT we found that our younger female staff tended to be the most underconfident,

relatively speaking. This underconfidence often moderates with experience and seniority, but in the meantime, it adds useful ballast by tempering the risk of group-level overconfidence.

In other words, one reason BIT has sustained good calibration is that our culture, structure, and staff mix allow underconfidence to persist, rather than squeezing it out in favour of surface-level certainty. If you want your own team to be better calibrated on average, you need to create space for both over- and underconfidence. Think of your colleagues as providing you with a portfolio of confidence levels – some overconfident, others underconfident, a few well-calibrated. It may be possible for some or all of them to become well-calibrated with time and practise. But in the meantime, you can aim to manage that portfolio wisely.

Start by learning to recognise the signals. Signs of good individual calibration include:

- **Caveated language that reflects uncertainty.** Phrases like "this depends on…" or "unless new data changes this…" are good signs of someone tracking the limits of their knowledge.
- **Conditional thinking and scenario planning.** Calibrated thinkers may frame predictions with assumptions ("If interest rates stay flat, then…"). They don't treat their forecasts as absolute.
- **Probing questions that surface hidden assumptions.** Rather than jumping to answers, calibrated thinkers often slow the group down: "What would change our mind here?", "What are we assuming?"

Signs of overconfidence can include:

- **Binary thinking.** Overconfident individuals tend to frame issues in black-and-white terms.
- **Unwillingness to revisit past views.** A refusal to admit earlier mistakes or shift views in light of new evidence is a red flag.
- **Volume over substance.** Some people dominate airtime with confident language and decisive tone but offer little underlying reasoning or evidence. Don't confuse presence with insight.

Signs of underconfidence can include

- **Reluctance to voice doubts** or alternative views, especially in group settings
- **Frequent hedging** or qualifying language that understates one's knowledge
- **Deferring to louder or more assertive colleagues**, even when privately disagreeing

As a leader, your instinct might be to demand confidence from your team – but keep in mind that the world already oversupplies overconfidence. A more challenging task is to create an environment that can tolerate underconfidence. When knowledgeable colleagues are reluctant to speak up, their silence can be mistaken for agreement, masking critical insights or risks. By seeking out and valuing calibration, you can identify those who are frequently correct but lack self-belief. This allows you to support their growth and help them develop genuine, well-earned confidence, rather than demanding they perform a brittle, artificial certainty. This strategy will help keep your team intellectually honest and ensure that the best ideas win out, not just the loudest voices.

4.2.2 Don't say "are you sure?"

This tip is simple but high-leverage.

In leadership roles, part of the job is to check in with less experienced colleagues on how their work is going. That means asking about plans, surfacing assumptions, and spotting potential problems before they escalate. One frequent challenge is that the power dynamic may discourage frank responses, as more junior team members may feel reluctant about admitting doubt or confusion in front of someone more senior.

This dynamic can be exacerbated by that common question "Are you sure?". It sounds innocuous, but it's a binary prompt with a socially loaded default. In professional settings, people will often think they're supposed to say yes. It nudges them toward feigned certainty – even when uncertainty would be more appropriate.

A tiny tweak makes a big difference. Instead of "Are you sure?", ask "How sure are you?"

This opens the door to graded responses: "I think I'm about 80%," "I'm not totally sure," "I'm confident about X, but less sure about Y." Suddenly you're having a richer, more realistic conversation – one where uncertainty isn't treated as weakness, and where confidence levels can be inspected rather than performed.

It also gives you better follow-up questions:

- "Why 80%?"
- "What's in the 20%?"
- "What would change your mind about this, or make you much more confident?"

This isn't just semantics. It's culture-setting. You're signalling that calibrated confidence is the desirable characteristic you're actually looking for from an experienced professional.

4.2.3. Reward good judgment, not loud certainty

To paraphrase Tetlock & Gardner (2015): if you reward confidence without accuracy, you'll get more confident errors. Organisations move toward whatever gets rewarded. And in many teams, what gets rewarded is fluency, conviction, and quick answers rather than strict accuracy. The result: overconfident errors, poor learning, and misallocated influence.

Flip the incentive structure. Make good judgment the thing that earns respect. That means rewarding people who:

- Calibrate their confidence appropriately
- Think in conditional terms
- Update based on new evidence
- Own and learn from mistakes

You can build this into culture through small rituals:

- Prediction pauses: "Before we go ahead, how confident are we this will work?"

- Calibration debriefs: “Which of our calls were off, and by how much? Who was appropriately cautious?”
- Public praise for good updates: Celebrate when someone revises a view based on new information. For example, in a performance review, you could explicitly praise an employee by saying, "The way you updated your view on the X project after seeing the new data was a great example of good judgment, and it helped us avoid a key mistake."

4.3 As a system designer

Many behavioural scientists – and likewise policymakers, product managers, team leads, and educators – work in organisations where they have influence over system design (eg, designing decision processes, setting performance metrics, structuring meetings, shaping hiring and promotion criteria, or choosing which data gets surfaced). They help shape the workflows, defaults, incentives, and rituals that govern how decisions are made and how thinking is rewarded. These are high-leverage roles.

Even if individuals model calibration and team leaders encourage it, it may not stick unless the broader system stops rewarding surface-level confidence and starts valuing epistemic accuracy. Good judgement won’t survive in a hostile environment. The culture, incentives, and decision architecture must align.

Here are system-level interventions that can help embed calibration into an organisation’s DNA:

4.3.1 Separate the judgment from the person

Good calibration requires honest judgments, but honesty is difficult in a social context. People are often judged not just on the quality of their ideas, but on their status, their relationship with the boss, or their willingness to support the prevailing consensus. To get an accurate reading of what people really think, you need to design processes that systematically separate the judgment from the person making it.

- **Use tactical anonymity to bypass hierarchy.** Anonymity isn't a long-term strategy for building team culture, but it's an incredibly effective tactical tool

for specific situations. For a strategy offsite or a project retrospective, using anonymous idea submission strips away job titles and focuses everyone purely on the content. It ensures that a brilliant idea from a junior analyst gets the same initial hearing as a mediocre one from a senior executive.

- **Pre-register forecasts to create an objective record.** Before a decision is made, have team members log their individual predictions (eg, "I'm 70% confident this feature will increase user retention by 5%"). This creates a record of what people actually thought before the outcome was known and before group consensus formed. It prevents hindsight bias and allows for an honest, data-driven review of whose judgment was well-calibrated.
- **Separate idea generation from critique.** When you mix generating ideas with evaluating them, people become defensive and attached to their own suggestions. This makes them less likely to update their views. A better process is to first generate a wide range of possibilities without judgment ("divergent thinking"), and only then move to a separate phase of assessing those ideas ("convergent thinking"). This lowers the emotional stakes and focuses the team on finding the best answer, not on defending their own contribution.

4.3.2 Build intellectual humility into your culture

Even the best-designed systems will fail if your culture rewards the wrong things. To make good judgment the default, you must actively and formally embed the values of intellectual humility into your organisation's DNA – how you hire, how you promote, and how you talk. You need to make it clear that curiosity and adaptability are valued more than declarative certainty.

Here's how:

- **Hire and promote for curiosity, not conviction.** During interviews, look for candidates who naturally use conditional language, who can articulate the weaknesses in their own arguments, and who are comfortable saying, "I don't know." When considering promotions, explicitly reward those who have a track record of updating their views in light of new evidence, not just those who defended their initial positions most loudly.

- **Standardise the language of uncertainty.** Build prompts into your core rituals and templates to make calibration a normal part of daily work. Add a section to your standard meeting agendas or project kickoff documents with questions like:
    - *"On a scale of 0-100%, how confident are we in this forecast?"*
    - *"What is the outside view or base rate for this kind of project?"*
    - *"What new information would need to be true for us to change our minds?"*

4.3.3 Make your decision-making environment more like weather-forecasting

Calibration isn't just about personal skill; it can be encouraged by the environment you operate in. Morgan (2014) starkly illustrated this by comparing two expert groups: doctors diagnosing pneumonia and weather forecasters predicting rain. When 80% confident, doctors were only correct in their judgement 20% of time – but weather-forecasters were right 80% of the time, perfectly calibrated.

Why the gap? Weather forecasters make thousands of predictions annually and get rapid, reliable feedback – often within hours. The doctors operated in slow-feedback environments, with outcomes unfolding over days or weeks. They don't get to recalibrate their confidence nearly as often or as quickly.

Many real-world decisions resemble the doctors' world more than the weather-forecasters'. Outcomes can take months or years to materialise, feedback can be ambiguous, and the environment may not encourage rigorous self-assessment. That doesn't mean we have to accept this as inevitable. We can reshape decision-making environments to be more like weather forecasting systems – feedback-rich, probabilistic, and transparent. This means:

- Setting clear, measurable delivery outcomes alongside project goals, focused on concrete, near-term markers of success.
- Making explicit forecasts on those outcomes, including confidence levels.
- Using those forecasts to inform planning and adjust expectations realistically.

- Reviewing actual results promptly, comparing them against forecasts, and tracking accuracy over time.

This four-step cycle – forecast, plan, act, review – turns calibration from an isolated individual skill into an organisational property. It was put into practice in the UK government in the early 2020s via its internal *Cosmic Bazaar* platform, where thousands of civil servants anonymously submit forecasts on geopolitical risks, track their accuracy, and adjust predictions based on new information.

5. Stay calibrated – and teach others to do the same

Most people are never taught to track their confidence. We're taught to chase correct answers, not to ask how well our certainty lines up with reality. Calibration offers a simple, scalable way to change this by asking us to routinely pause, ask "*How sure am I?*", and observe our track record over time.

This may feel effortful at first, but the power of this habit is that it makes cognitive distortions harder to sustain. It doesn't eliminate error at the source, but it makes our overconfidence and blind spots more visible, and that visibility invites adjustment. Over time, if you persist and if the environment provides regular, meaningful feedback, you will begin to develop accurate intuitive expertise (Kahneman & Klein, 2009).

Learning the correct form for an exercise, like a sprint or a bench press, won't on its own make you a weightlifting champion – that requires innate talent and thousands of hours of work. But everyone can learn good form, and everyone can use it to improve, whatever their starting point.

The same is true of decision-making. Calibration is 'good form for thinking'. This report has shown you the fundamentals of that form. Learning it won't instantly turn you into a world-class decision-maker. But it will make you a better thinker.

So stay calibrated – and teach others to do the same.

References

Armstrong, B. A., Dutescu, I. A., Tung, A., Carter, D. N., Trbovich, P. L., Wong, S., Saposnik, G., & Grantcharov, T. (2023). Cognitive biases in surgery: Systematic review. *British Journal of Surgery, 110*(6), 645–654. https://doi.org/10.1093/bjs/znad004

Belton, I. K., & Dhami, M. K. (2021). Cognitive biases and debiasing in intelligence analysis. In R. Viale (Ed.), *Routledge handbook of bounded rationality* (pp. 548–569). Routledge/Taylor & Francis Group.

Egan, M. (2024) The UK public is overconfident. BIT Working paper No. 007 https://www.bi.team/publications/working-paper-no-6-the-uk-public-is-overconfident/

Egan, M., Tran, C., & Whitwell-Mak, J. (2025). Overconfidence is the norm.

Gutierrez de Blume, A. P. (2022). Calibrating calibration: A meta-analysis of learning strategy instruction interventions to improve metacognitive monitoring accuracy. *Journal of Educational Psychology, 114*(4), 681–700. https://doi.org/10.1037/edu0000692

Hallsworth, M., Egan, M., Rutter, J., & McCrae, J. (2018). Behavioural Government: Using behavioural science to improve how governments make decisions. The Behavioural Insights Team.11 Jul 2018

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. The American psychologist, 64(6), 515–526. https://doi.org/10.1037/a0016755

Kelly, M. O., & Mandel, D. R. (2024). The effect of calibration training on the calibration of intelligence analysts' judgments. *Applied Cognitive Psychology*. Advance online publication. https://doi.org/10.1002/acp.4236

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences of the United States of America, 111*(20), 7176–7184. https://doi.org/10.1073/pnas.1319946111

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., ... & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552-3565.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown Publishing Group.

BIT

58 Victoria Embankment
London EC4Y 0DS
+44 (0)20 7438 2500
info@bi.team
𝕏 @B_I_Team
www.bi.team