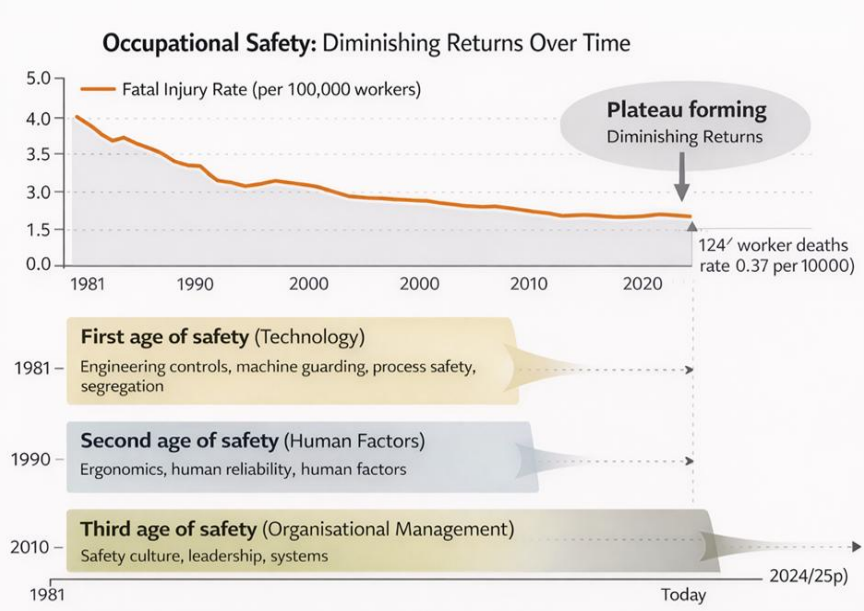

Are there limits to safety?

Background

The safety curve has not stopped improving because people stopped caring. If anything, the opposite is true: board-level commitment (#1 Priority?), Safety Management Systems, near-miss reporting, and an expanding industry of behavioural, cultural, and psychological interventions are now available. Yet the iconic curve appears to have bottomed out. That "flat tail" however, should not be a reason for resignation—it requires sharper diagnosis, because it raises an uncomfortable possibility that the field may be approaching the limits of current approaches, even while the limits of what is achievable remain open.

Three familiar explanations compete, and each implies a different next move. The first is the law of diminishing returns: once the highest-leverage controls have done their work, each additional reduction becomes progressively harder and more expensive. The second is cost benefit / ALARP economics: improvements slow not because risk is accepted, but because only a narrowing set of controls remains "reasonably practicable." The third is compositional: the remaining harm is concentrated in residual pockets of stubborn, hard-to-shift risk where general programmes barely bite. And hovering over all three is a more awkward question—whether some "improvements" increasingly function as professional reassurance rather than as interventions with observable effect on outcomes.

This is where the usual dismissal of behavioural adaptation (the "risk thermostat") goes wrong. A familiar critique assumes that if people adapt to safety, improvements should backfire and fatalities should rebound; because the long-run record shows sustained decline, adaptation is declared a myth. The hidden assumption is that adaptation must push harm back up. A learning-based account predicts something subtler and more operationally important: not rebound, but flattening—progress becomes stubborn when outcomes are rare, signals are sparse, and learning depends on proxies. The sections that follow set out the competing explanations for the plateau, and—crucially—what evidence would falsify each.

A familiar criticism of “risk thermostat” arguments goes like this: if people really adapt their behaviour to safety controls, then safety improvements should backfire and fatalities should rebound. But the long-run occupational safety record doesn’t show rebounds — it shows sustained decline. Therefore, the critique concludes, behavioural adaptation must be a myth, or at best a distraction.

The hidden assumption is that adaptation, if real, must push deaths back up. That is the caricature of homeostasis. A learning-based account predicts something subtler and operationally more important: not rebound, but a slowing of the improvement slope once the highest-leverage controls have already done their work. In other words, the signature is not “harm returns”, but “progress becomes stubborn”.

The Numbers

Take the UK’s long-run fatal injury record. The latest published full-year figures report 124 worker deaths in 2024/25, which corresponds to a fatal injury rate of 0.37 deaths per 100,000 workers (HSE, 2025). That is not “risk compensation causing harm”. It is an extraordinary success story. Yet the same body of reporting also makes the uncomfortable point: after decades of decline, recent years look broadly flat once you set aside pandemic distortions (HSE, 2025).

When you convert that rate into an individual annual probability,(Figure 1 -), 0.37 per 100,000 is about 3.7×10^{-6} per worker-year — in round terms, ~0.4×10^{-5}. This is the pattern worth staring at: the system has driven risk down into the few-in-a-million range, and then the curve starts to “stick”.

That steep-drop-then-tail shape is not unique to UK fatality rates. It shows up in other occupational and industrial metrics too. In the Australian oil and gas sector, APPEA’s 2011–12 report shows lost-time injuries per million hours worked falling to 0.8 in 2011 (from 1.0 in 2010 and 3.4 in 1996), while the total recordable injury rate fell from 13.4 (1996) to 5.1 (2010) to 4.7 (2011) per million hours (APPEA, 2011–12). Again: no rebound. But a recognisable story of big early gains and a flattening tail.

Figure 1 – The Individual occupational risk trends OBSERVED

So why does the tail flatten?

One answer is pure leverage. A useful shorthand is the "ages of safety": an early period dominated by technological controls, followed by increasing emphasis on human factors and then organisational and systemic management (ARPANSA, n.d.). The point is not that later approaches "don't work". It's that engineering interventions like guarding, segregation, enclosure, isolation, and intrinsically safer design can remove hazards directly and therefore produce large, immediate reductions. Once those wins are banked, what remains is more often about coordination under pressure, timing, workload, degraded margins, and the messy reality of work — domains where improvement is possible but slower, noisier, and harder to lock in.

A second answer is heterogeneity and residual pockets. A national average can "plateau" even while some sectors keep improving, if most of the remaining burden sits in a small set of high-hazard activities where elimination is technically or economically harder. In that world, the flat tail isn't a psychological set-point; it's a compositional effect: the easy hazards are gone, and the remaining ones are stubborn.

A third answer is governance economics. R2P2 (Figure 2 -), formalises the idea that, beyond a point, demanding ever-smaller risks regardless of cost is neither realistic nor socially optimal. Instead, we operate in a tolerable if ALARP space where further reduction is expected only where reasonably practicable (HSE, 2001). In that framing, flattening isn't mysterious: it's the empirical signature of diminishing marginal returns and bounded attention.

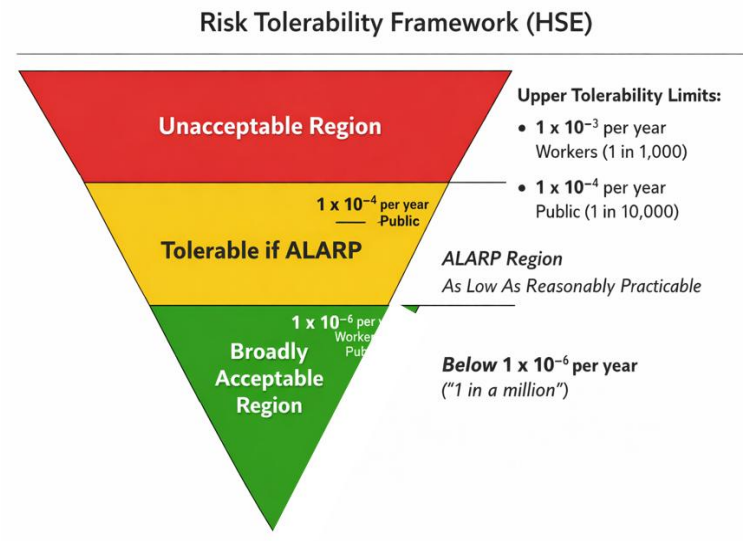

Figure 2 – The iconic R2P2 Tolerability of risk Triangle (of which I was partly responsible!)

But is there another explanation?

Those explanations are credible. But there is a deeper hypothesis that ties the plateau directly back to behavioural adaptation — without requiring rebounds — and it is exactly the one that "learning lens" implies.

Risk thermostat and target-risk accounts (Adams; Wilde) do not require safety to "fail". They argue that when conditions change, behaviour and practice adjust until perceived difficulty, reward, and danger settle into a workable balance. Translate that into a TD/prediction-error

idiom and you get something like this: systems update policies and habits when experience generates sufficiently informative prediction errors. When safety controls reduce consequences and smooth experience, they may also reduce the frequency and clarity of the "margin signals" that drive updating. Near misses become less visible; degraded-mode cues are engineered away; everyday work feels more controllable; and the behavioural envelope quietly expands into the newly available safety margin. The net outcome can remain overwhelmingly positive — fatalities stay low — but the *rate of further improvement slows* because learning converges on a new steady state.

This is where the plateau around ~0.4×10^{-5} per worker-year becomes provocative. It may be telling us not only about engineering limits or ALARP economics, but about a natural prediction-error tolerability level.; i.e., a region where catastrophic outcomes are rare enough that, without deliberately engineered feedback, they stop providing a usable learning signal for the system as a whole. The "tail" then reflects a kind of informational floor. Below a few-in-a-million per year, the system can no longer rely on harm events to calibrate behaviour, and must rely on proxies — audits, near misses, weak signals, process drift indicators — to keep learning alive.

If that is even partly true, it sheds light on why R2P2-style tolerability bands have felt surprisingly durable. R2P2's "one in a million per annum" is explicitly presented as a guideline boundary between broadly acceptable and tolerable regions for individual risk of death (HSE, 2001). And it also recognises asymmetry between worker and public risk at the upper end, reflecting voluntariness/consent and social expectations (HSE, 2001). A "prediction-error tolerability" interpretation doesn't replace those ethical and political arguments — it complements them: it says these numbers may be governable partly because they sit near a natural transition where raw outcomes stop being informative and engineered feedback has to take over. At that boundary, it is no longer reasonable to demand endless improvement through the same mechanisms that worked earlier, because the learning substrate has changed.

Implications for Safety

This reframes the practical challenge for the next decade. The question is not whether safety controls work — they clearly do. The question is whether we are designing and governing modern high-control systems in a way that preserves learning without requiring harm. In Safety-II language, this means learning from everyday performance, adaptation, and "what goes right," rather than waiting for rare catastrophes to teach us (Hollnagel, 2015). In TD language, it means increasing the density and quality of prediction-error signals without increasing injury: strong near-miss visibility, operational "friction" metrics, degraded-mode cues, workload and time-pressure indicators, and explicit margin-to-failure dashboards.

Crucially, this hypothesis is testable. If the plateau is mostly "residual hazard pockets," then only structural hazard elimination in those pockets should move the national mean. If it is mostly "ALARP economics," then shifts in what counts as reasonably practicable (technology cost curves, enforcement intensity) should predict changes in slope. If it is partly "prediction-error tolerability," then organisations and sectors that systematically amplify non-harm feedback — turning weak signals into strong learning — should continue to drive the curve downward when others stall.

Conclusion

So the occupational curves don't disprove behavioural adaptation. They show why the debate won't go away. Controls drove huge early gains. Then the system adapted. Now the slope is harder to move. The next gains depend less on adding another barrier and more on designing an environment where the learning signal remains strong even when the harm signal has become mercifully rare.

References

- The plateau claim is grounded in the most recent published GB fatality statistics context (HSE/UKATA summaries). HSE, (2025) Work-related fatal injuries in Great Britain,
- The "tail shape" cross-check uses APPEA's reported LTIFR/TRIFR improvements. (https://www.appea.com.au/wp-content/uploads/2013/04/APPEA_HSE_2011-12.pdf?utm_source=chatgpt.com)
- The link to tolerability bands uses HSE's own R2P2 wording on the "one in a million" guideline boundary and ALARP framing. (https://assets.publishing.service.gov.uk/media/6693ad9e49b9c0597fdafc36/IQ8.10.J_Document_9_Health_and_Safety_Executive__Reducing_risks__protecting_people__HSE_s_decision-making_process__2001.pdf?utm_source=chatgpt.com)
- The behavioural-adaptation mechanism is anchored in classic risk thermostat / target-risk formulations (Adams; Wilde), but applied here as a hypothesis about information/learning. (https://www.john-adams.co.uk/wp-content/uploads/2017/01/RISK-BOOK.pdf?utm_source=chatgpt.com)