Goal-Based Regulation: Building the Puzzle from Both Ends

I work as a consultant at NCC Group, supporting customers across a wide range of Cyber-Physical Systems—including critical national infrastructure and transport. Before moving into consultancy, I completed an Engineering Doctorate (EngD) at Warwick, focusing on on-board and off-board data platforms in automotive and related embedded technologies.

That blend of academic research and hands-on assurance work has shaped how I approach questions of regulation. Much of my work involves helping organisations make sense of structured assurance methods, building arguments and evidence for homologation, and tracing how these practices evolve when applied to emerging technologies and new regulatory landscapes.

It's from that position that I've been forming my perspective on **goal-based regulation**. I don't pretend to have the final word, but I hope I can offer insights from seeing how assurance plays out both in theory and in practice—insights that may be useful as we collectively work out what goal-based regulation really means, and how it could influence the bigger assurance picture beyond simple regulatory approval.

At London International Shipping Week (LISW), the phrase cropped up repeatedly. But I'm not convinced everyone is on the same page about what "goal-based regulation" actually means. Some treat it as a loosening of rules, others as a formalised assurance case, others still as a vehicle for embedding policy priorities. My view is that it's best understood as part of a wider movement—one that includes frameworks like the **NCSC's Principles-Based Assurance (PBA)**—toward outcome-focused assurance.

From Checklists to Principles

Traditional regulation often prescribes how to meet requirements: "install this piece of equipment," "apply this process," "comply with this standard in this way." Such approaches bring clarity but can also stifle innovation and invite box-ticking.

The alternative is a principles- or goal-based regime. Instead of dictating how to comply, regulators set high-level goals and expect industry to demonstrate that they are met. This mirrors the **Principles-Based Assurance (PBA)** approach emerging from the UK's National Cyber Security Centre, which advocates setting clear principles and requiring organisations to make structured **claims** about how those principles are satisfied, backed by evidence.

In other words: the regulator defines the "what," and the organisation shows the "how."

Not a Loosening, but a Reframing

One reason goal-based regulation is misunderstood is that it's often contrasted with axiomatic, checklist-driven assurance. That style can feel onerous: pages of prescribed steps, requirements, and tests, all of which must be completed whether or not they fit the specific context.

It's tempting, then, to see goal-based regulation as a loosening of the rules—as if regulators are stepping back and asking less. In reality, they are asking more. They are no longer satisfied with proof that every box has been ticked. They want to see the reasoning, the evidence, and the maturity of the assurance case.

Take witness testing under UNECE R155 as an example. Occasionally, this boils down to a regulator observing a third party perform a penetration test. But the real assurance value isn't in the act of watching the test run; it's in showing how the organisation's **Cybersecurity Management System (CSMS)**—already granted a Certificate of Compliance—has been applied to the vehicle type undergoing approval. That means looking at how the test was scoped, how findings were ingested and triaged, what decision-making processes were followed, and whether the organisation can demonstrate a before-and-after picture if patches or remedial steps were taken.

This makes R155 very different from an emissions regulation. Emissions regimes are largely about **measurement quality, repeatability, and preventing cheating**. They ask: was the test representative, reproducible, and free from manipulation? R155, by contrast, does not assume the absence of security findings. It asks instead: does the organisation have a live, functioning CSMS that manages risks systematically and applies it consistently to the product under approval?

And this does increase the workload. An effective scope to demonstrate these things often goes beyond traditional penetration testing. It can include considerations such as **Malicious Use of Intended Functionality**—where the very features, backdoors, or capabilities deliberately built into a system can be exploited. In these cases, the assurance question becomes: *what controls do you have in place to manage the risks associated with this functionality, and how have those controls been tested?* Demonstrating that rigorously is far more demanding than a one-off test; it requires process evidence, systematic risk management, and structured reasoning.

The Problem with Shallow Metrics

One of the traps of goal-based regulation is that it risks being captured by simple, shallow metrics. "Safe enough" becomes reduced to whatever is most easily measured. In the case of autonomous vehicles, this has often been the claim that a self-driving car should meet a safety standard of *"careful and competent human drivers."* On the surface, it's an appealing goal: intuitive, easy to communicate, and apparently measurable.

But as Professor Philip Koopman has argued, it's also dangerously misleading. A car could be statistically safer in aggregate while still making catastrophic mistakes no competent human would. Reducing "safety" to a single curve or benchmark creates perverse incentives: companies can tout mileage numbers or disengagement rates, while ignoring whether their system reliably avoids corner cases, protects vulnerable road users, or fails gracefully in the presence of uncertainty.

This isn't a hypothetical concern. In California, autonomous vehicle developers file "disengagement reports" tallying how often a human had to take over. These numbers have become a de facto scoreboard, with some firms boasting that their vehicles go thousands of miles between interventions. Koopman and others point out that this metric is almost meaningless. Disengagements depend on how aggressively a company chooses to test, how it defines a "safety-critical" handover, and what risks it is willing to tolerate. A vehicle could appear excellent on paper while masking brittle behaviours in real traffic.

The lesson for goal-based regulation is clear: a high-level claim like *"safer than a human driver"* cannot stand on its own. It must be unpacked into structured arguments and evidence — covering accident types, ethical failures, vulnerable populations, and rare but catastrophic risks. Left unqualified, such goals invite companies to cherry-pick data and declare victory. In the worst case, financial interests bend the meaning of "safe enough" until it serves the business model rather than the public interest.

This critique resonates far beyond road vehicles. Maritime operators might focus on average incident rates rather than near misses in congested waters. Aviation could declare autonomy successful if it passes aggregate reliability numbers, while overlooking failure modes in adverse weather. Energy systems might emphasise uptime while neglecting resilience against coordinated attack. In every domain, the temptation is the same: reduce the messy, multi-dimensional nature of safety and security to a single metric.

Goal-based regulation must resist this. The regulator's role is not just to state the goal, but to ensure the way sufficiency is measured cannot be gamed. Otherwise, what looks like flexibility becomes fragility.

Safety Principles in Practice: Reflections from the UK Consultation

This tension isn't just theoretical. In 2022, the UK Government ran a consultation on its draft Statement of Safety Principles (SoSP) for self-driving vehicles. As part of that process, our team at NCC Group responded to the call for evidence, drawing on work across the global automotive sector and lessons from other cyber-physical domains.

We welcomed the SoSP as a foundation for pre-deployment authorisation, post-deployment oversight, and annual performance monitoring. But we also highlighted areas where a purely principles-based framework could become hollow without structure and scope:

- **Design-phase assurance**: Safety principles should not only apply at the approval gate. They need to shape design-phase decisions, where architecture, risk trade-offs, and long-term maintainability are locked in. We argued that structured assurance cases should be mandated and reviewed across the vehicle lifecycle, making reasoning traceable and auditable.

- **Strategic outcomes**: Safety is not just absence of crashes. We urged the SoSP to align with national transport strategy — accessibility for disabled and elderly users, support for public transport and active travel, and consistency with decarbonisation goals. AVs should not be allowed to erode wider societal objectives under the cover of "being safe."

- **Business intent**: We argued that an operator's business model must form part of the safety picture. If a deployment model undermines public benefit, it is by definition unsafe at the societal level. Annual assessments are too limited; longitudinal accountability is needed.

- **Public reassurance**: Trust is built when citizens believe standards are enforced and behaviour is monitored. We recommended "black box" monitoring of AVs — recording near misses, performance deviations, and forensic data — so regulators can verify safety over time.

Most importantly, we cautioned against the government's proposal to use human driver comparison as the primary benchmark. Our reasoning aligned closely with Koopman's critique:

- Humans and AVs are categorically different — AVs should be engineered for predictability, explainability, and bounded failure, not simply "better averages."

- Aggregate statistics hide whether a system's safety reasoning is sound.

- The strength of automation is that its failure modes can be specified, tested, and controlled.

Instead, we urged that AVs be designed to **avoid unsafe states**, fail **predictably within their ODD**, and never introduce new hazards through erratic or ambiguous behaviour. The goal is not to engage in trolley-problem speculation, but to **design the trolley problem out of scope** through principled engineering and clear boundaries.

Finally, we flagged that this regulatory ambition requires regulator capability. The UK's certification agencies face a significant skills gap in cyber security. Without resourcing, expertise, and the ability to call in external specialists, the best-written safety principles risk being unenforceable.

Operational Design Domains: Who Defines the Boundaries?

If safety principles define *what outcomes must be achieved*, the **Operational Design Domain (ODD)** defines *where and when those outcomes apply*. For automated driving systems, the ODD specifies the exact conditions in which a system is designed to operate: road types, weather, lighting, traffic density, speed ranges, and more.

At first glance, the ODD seems like a technical detail. But in reality, it is a **regulatory linchpin**. Before you can ask whether a self-driving system is safe, you must know the boundaries of the domain in which it is expected to be safe. A shuttle geofenced to a business park in daylight fair-weather conditions presents a fundamentally different assurance challenge from a robotaxi claiming 24/7 operation on UK roads.

The **ISO 34503 standard** and BSI's PAS 1883 taxonomy now provide a structured vocabulary for defining ODDs. Siddartha Khastgir's group at Warwick University has been central in this work, emphasising that **ODD definition is the first step in a credible safety case**. By making scope explicit, regulators, developers, and the public can understand both the capabilities and the limitations of a system.

But there is a danger here: if left solely to developers, ODDs may be drawn according to *business convenience* rather than *societal value*. A company may select domains that are profitable or technically tractable, while neglecting contexts (rural roads, bad weather) where automation could deliver public good but is harder to achieve. Worse, companies have conflicting incentives: they may **downplay the limits of their ODD in marketing** to make products seem capable, then **emphasise those limits in liability discussions** after an incident.

That is why regulators must have a voice in ODD definition. They can set baseline expectations (for example, requiring that all systems handle common vulnerable road users in their declared domain) and align ODD categories with national strategy. Rather than passively accepting industry's chosen domains, regulators can signal the *ODDs society wants first* — urban shuttles that improve accessibility, freight automation on highways to reduce emissions, or rural services to tackle transport deserts.

Threats Within the ODD

Most ODD discussions focus on *passive factors*: the weather, the road type, the time of day. But in reality, these systems will not exist in a benign environment. They will encounter adversarial conditions — both accidental and deliberate. Just as **SOTIF (Safety of the Intended Functionality)** extends functional safety to cover performance limits and misuse, ODDs should be extended to consider *hostile actors*.

That means recognising that an ODD must include assumptions about threats:

- **Cyber-attacks**, where remote services or vehicle networks are targeted.
- **Sensor spoofing or jamming**, such as GPS denial, LiDAR blinding, or adversarial signage.
- **Human hostility**, where members of the public deliberately interfere with or obstruct vehicles.
- **Systemic disruption**, where AV services could be targeted precisely because they are public-facing and capable of degrading traffic flow or essential services.

If an ODD does not account for these factors, its boundaries are incomplete. For example, a claim that "the system is safe in daylight urban conditions" is insufficient unless it also shows how it withstands foreseeable threat conditions *within that environment*. A credible ODD must therefore include not just the physical world, but the threat world.

Seen this way, ODDs become even more important as assurance tools. They force developers and regulators to be explicit about **both environmental and adversarial assumptions** — and they make clear what evidence is needed to show resilience.

Cross-Sector Insights: The Learning is Everywhere

Although my reflections on goal-based regulation crystallised at London International Shipping Week, the underlying issues are not unique to shipping. Across every transport mode, and even beyond transport, regulators are wrestling with the same puzzle: how do you move from prescriptive checklists to outcome-focused assurance without letting industry quietly redefine sufficiency for itself?

Maritime: Goal-Based Standards on the Water

The maritime world already has experience with goal-based standards. The IMO's Goal-Based Standards (GBS) for ship construction were introduced more than a decade ago: rather than prescribing every technical detail, the IMO required that classification societies demonstrate their rules achieve high-level safety goals. The upcoming **MASS Code** for Maritime Autonomous Surface Ships follows the same template. Instead of rigid requirements, it sets functional goals — safe navigation, collision avoidance, operator accountability — and leaves it to designers and operators to demonstrate compliance.

But here too, the "operational domain" is not just about weather, waves, and sea state. Autonomous vessels are already attractive targets for piracy, GPS spoofing, and cyber attacks against control systems. A vessel's ODD equivalent must therefore include assumptions about adversarial threats: *can it withstand signal spoofing in congested ports, cyber intrusion during cargo operations, or deliberate obstruction in choke points?* If those are not defined up front, the assurance case is incomplete.

Automotive: Cybersecurity and Safety Cases

Automotive regulation is already grappling with goal-based logic. UNECE R155 and R156 do not prescribe one right way to secure vehicles or update software. Instead, they require manufacturers to establish a **Cybersecurity Management System (CSMS)** and a **Software Update Management System (SUMS)**, then show how those systems are applied to each vehicle type. This is goal-based regulation in action: demonstrate that risks are managed systematically.

As noted earlier, this is fundamentally different from emissions regulation. Where emissions focuses on repeatable measurements and avoiding cheating, cybersecurity requires showing that a **process is alive and effective**. And that process cannot ignore hostile actors: if an ODD claims safety only in fair weather, but does not consider that a vehicle might be targeted with a CAN-injection attack or GNSS jamming, then the assurance case is blind to the real risks.

Aviation: The Conservative Gold Standard

Aviation has long used assurance case-like methods. Certification requires manufacturers to demonstrate airworthiness, typically with probabilistic arguments about redundancy and failure rates. Safety goals like "no catastrophic failure at the fleet level" are unpacked into detailed analyses.

But aviation also knows adversarial conditions. From deliberate laser dazzling of cockpits to cyber threats against avionics, regulators increasingly expect that operational envelopes include hostile conditions. A drone delivery service, for instance, cannot claim an ODD of "suburban airspace in daylight" without also considering the inevitability of interference — malicious RF jamming, spoofed navigation signals, or physical attacks on small drones.

Rail: From Prescription to Safety Cases

Rail safety in the UK shifted in the 1990s from prescriptive rules to a **safety case regime**, where each operator had to demonstrate risks are managed ALARP (as low as reasonably practicable). This is a goal-based principle. The shift gave operators flexibility but imposed a burden: transparent safety cases regulators could inspect.

And like other domains, rail has learned to account for hostile conditions. TS 50701, the cybersecurity standard for rail, explicitly integrates cyber threats into assurance. Train control systems cannot simply claim "safe in normal operating conditions"; they must also prove resilience against intrusion, denial-of-service, or manipulation of signalling. The "ODD" for rail is not just track and weather, but threat context.

Energy and Cyber: Principles-Based Assurance

Critical infrastructure such as power and telecoms has moved decisively toward outcome-based regimes. The UK's NIS Regulations and the upcoming NIS2 directive require operators of essential services to **demonstrate resilience against foreseeable risks**, including cyber attack. Here too, the "operational design domain" is not just technical boundaries of a grid but assumptions about adversaries: insiders, state actors, criminal ransomware groups.

The NCSC's **Principles-Based Assurance (PBA)** is another direct articulation of this idea: claims and evidence must be evaluated not only against benign operating assumptions but against the spectrum of adversarial threats. A system that is safe in ideal conditions but brittle under attack is not assured.

The Common Thread

Across these sectors, the pattern is clear. Goal-based and principles-based frameworks create adaptability and allow innovation. But they also **increase the assurance burden**: you must show your reasoning, not just your results, and you must scope your claims against both *environmental* and *adversarial* conditions. Regulators need the skill and resources to interrogate those arguments, and organisations must adopt structured methods that expose assumptions rather than conceal them.

The learning is everywhere. Shipping, cars, planes, trains, grids, and code: all are converging on the same challenge. And in every case, an ODD (or its sectoral equivalent) that ignores threats is an ODD that cannot support a credible assurance case.

Structured Assurance: The Grammar of Goal-Based Regulation

If goal-based regulation defines *what must be shown*, structured assurance provides the **language for showing it**. Without a clear grammar, claims risk becoming hand-wavy, arguments stay implicit, and evidence is cherry-picked to suit the story. With structure, regulators, developers, and the public can actually see how reasoning flows from principle to practice.

The two most widely used notations are **CAE (Claims–Arguments–Evidence)** and **GSN (Goal Structuring Notation)**. Both take the same idea:

- **Claims**: what you are asserting (e.g. "the vehicle is acceptably secure in its ODD").
- **Arguments**: how you justify that claim (e.g. by decomposing it into subclaims about network security, sensor integrity, update mechanisms, etc.).
- **Evidence**: the concrete artefacts that back up the arguments (test results, analysis, process audits, code review records).

The visual structure of GSN or the tabular rigour of CAE forces clarity. You cannot wave at "safety" as a single box; you must break it down, show dependencies, and expose gaps.

This is especially important in domains where regulators are setting high-level goals. If a maritime code says "the vessel must remain safely operable under foreseeable cyber attack," then a structured assurance case makes it explicit how that broad claim is supported: what scenarios are covered, what assumptions are made, what evidence is considered sufficient. Without that structure, industry is tempted to provide glossy narratives rather than traceable reasoning.

Structured assurance also supports **fragment reuse**. Regulators can publish partial arguments or common fragments — for example, mapping ISO 21434 requirements into security claims for vehicles, or mapping IEC 62443 controls into claims for maritime systems. Developers then build on these fragments, adding system-specific evidence. This avoids everyone reinventing the wheel, while still keeping responsibility with the regulated entity.

But perhaps the greatest strength of structured assurance is that it **makes manipulation harder**. By exposing claims, arguments, and evidence explicitly, it becomes easier for a regulator to see where assumptions are too optimistic, where evidence is thin, or where arguments don't really flow. It is the opposite of compliance theatre.

There are limitations. Today's tools often feel static — a snapshot rather than a living case. That's where the next step comes in: recognising assurance as a **temporal process**, not just a document frozen in time.

Assurance Over Time and Assurance Debt

One of the weaknesses of many current assurance practices is that they are treated as **static artefacts**. A safety case is written for approval, then quietly shelved; a cyber certification is granted once, then left to gather dust. This makes sense if technology is static. But cyber-physical systems are anything but: software evolves, attack surfaces shift, and operational contexts change. Assurance that does not evolve is soon irrelevant.

That is why goal-based regulation must be tied to a **temporal lens**. We already use maturity models elsewhere. **Technology Readiness Levels (TRLs)** describe how far a system has progressed from concept to deployment. **CMMI and other maturity frameworks** assess organisational process capability. The same logic should apply to assurance. A TRL-3 prototype will not — and should not — carry the same weight of evidence as a TRL-9 production system. What matters is that assurance *matures in step with the technology*. Claims are refined, arguments deepen, evidence accumulates.

This temporal perspective becomes even more important where public funding and trials are involved. A prototype that makes bold safety claims but does not build an assurance case is not just unconvincing; it is laying down **assurance debt**.

What is Assurance Debt?

The analogy to **technical debt** is intentional. Technical debt accrues when developers take shortcuts, pushing messy fixes into the future until the cost of change explodes. **Assurance debt accrues when systems advance without the reasoning and evidence that regulators (and society) will ultimately demand.**

At first, the debt feels manageable. A trial proceeds with limited documentation, a regulator gives provisional approval, a team promises to "fill in the case later." But as the system matures, the gap widens. By the time the product is ready for mass deployment, the backlog of missing assurance is insurmountable: the evidence cannot be retrofitted, the design decisions cannot be justified, the arguments cannot be reconstructed.

Why Debt Visibility Matters

Goal-based regulation makes this debt visible in a way checklists do not. Because the **top-level claims are known from the outset**, everyone can see how far the current evidence falls short. If a team is not incrementally building toward those claims, the shortfall is obvious. That visibility can be uncomfortable — but it is exactly the pressure needed to keep assurance in lockstep with development.

This is not just about compliance. Assurance debt undermines confidence for all stakeholders:

- **Regulators** face pressure to approve systems without sufficient grounding.
- **Investors** see risk that products will stall at approval gates.
- **Operators** inherit systems whose trustworthiness they cannot demonstrate.
- **The public** encounters technologies that arrive with slick marketing but shaky credibility.

Managing Assurance Debt

Some debt is inevitable. Early-stage prototypes cannot be expected to carry full safety cases. The question is whether organisations have a plan to pay it down. Are claims being identified early? Are arguments being drafted in lightweight form? Is evidence being accumulated as part of development, not bolted on afterward?

This lifecycle approach to assurance requires regulators to adjust as well. Approval processes should not only look at today's evidence, but at how convincingly an organisation can demonstrate that assurance maturity is growing. A promising technology with a weak assurance trajectory is not on track.

In short: assurance is not a document. It is a living process, and one that can go bankrupt if debt is left unmanaged.

Different Eyes, Different Evidence

Assurance is never one-size-fits-all. The same claim — *"this system contributes to road safety"* or *"this vessel remains secure under cyber attack"* — will mean different things to different audiences. If goal-based regulation is to function, it must recognise this plurality and structure evidence accordingly.

- **Regulators** need depth. Their role is to interrogate claims, challenge arguments, and demand robust evidence. For them, assurance must expose assumptions and show traceability: how specific tests, analyses, and processes ladder up to top-level goals. They need to see the black-box data recorders, the audit logs, the penetration test scoping, the decision-making processes around triage and remediation.
- **Operators** need operational assurance. They are less concerned with how a piece of software was verified in the lab, and more with whether the system will function reliably in service, under cost and performance constraints. Evidence for them must speak to maintainability, resilience under real-world stressors, and clarity of boundaries (ODDs, service guarantees, fallbacks).
- **Funders and investors** need trajectory. Their question is not only "is the system safe today?" but "is the assurance case keeping pace with development?" A company with strong technology but mounting assurance debt is a risky bet. For them, evidence includes roadmaps, maturity models, and demonstration that assurance reasoning matures with each iteration.
- **The public** needs reassurance, not raw artefacts. Citizens do not need to pore over a code review log or a probabilistic risk assessment. They need credible, accessible signals: that the regulator has interrogated the case, that independent scrutiny exists, that performance is being monitored, and that systems are not quietly extending beyond their authorised scope. Public trust is fragile, and it depends less on the presence of evidence than on the *credibility of the assurance process itself*.

Structured assurance helps here. One claim can be connected to multiple arguments and multiple forms of evidence, each tuned to its audience. The regulator may receive a detailed CAE case with hundreds of supporting artefacts. The operator may see a dashboard of resilience metrics. The public may see a plain-language summary validated by an independent regulator. All are valid, all stem from the same structured root.

This is why assurance must be treated as a **multi-layered conversation** rather than a monolithic report. The danger of goal-based regulation is that it becomes captured by one audience — usually regulators and industry insiders. The opportunity is to design assurance that works for all, and in doing so, strengthens both legitimacy and trust.

Beyond Approval: Assurance in Society

It's tempting to treat assurance as a hurdle: pass the test, win the approval, move on. But that mindset misses the bigger picture. Regulatory sign-off may be a critical milestone, but in complex, cyber-physical domains it is not the end of the story.

Assurance is fundamentally about **legitimacy in society**. It is the means by which a technology earns and maintains its right to operate. The regulator's perspective is one piece of that, but not the only one.

A system that meets minimum safety thresholds but undermines public transport, worsens congestion, or concentrates benefits in narrow economic interests is not assured in any meaningful sense. A system that passes a security audit but then erodes trust through opaque behaviour or inconsistent accountability is not truly safe in the eyes of the public. Assurance in this wider view is not just *can you do it legally*, but *should you do it, and are you doing it responsibly?*

This is where goal-based regulation shows its full potential. By framing outcomes in terms that link to national strategies — decarbonisation, accessibility, resilience, security — regulators can ensure that safety cases are not just about product survival, but about public good. And by requiring structured assurance that matures over time, they can make visible whether industry is paying down its **assurance debt**, not just meeting today's baseline.

It also means that the **interface between regulators and organisations** becomes more than a gatekeeping moment. It is a site of ongoing scrutiny, where the top-down puzzle pieces (societal goals, safety principles, regulatory objectives) meet the bottom-up puzzle pieces (industry evidence, assurance cases, operating data). The fit between them must be continually checked, tested, and adjusted.

Seen this way, assurance is not a cost centre or a compliance exercise. It is the connective tissue between technology, regulators, and society. It is the way we decide which innovations deserve trust, which risks are acceptable, and which futures we want to build.

The work is heavy, yes. Checklists may feel more straightforward, but they hide debt, disguise assumptions, and lull us into false security. Goal-based frameworks, structured assurance, and adversarially-aware ODDs demand more — of regulators, of industry, of ourselves. But the reward is greater too: technologies that are not only allowed, but welcomed, because they can demonstrate, transparently and robustly, that they serve society's goals.

That is the puzzle worth building. Not from the top down alone, not from the bottom up alone, but from both ends at once, until the picture is clear.