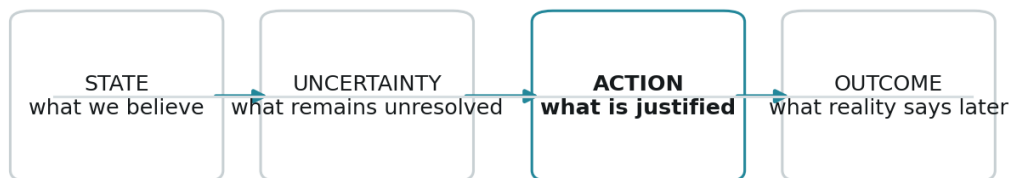


MODUS-HELIX 2

THE NEXT DECISION

From athlete state to bounded progression intelligence

THE NEXT DECISION



The intelligence is not in choosing more often. It is in choosing only when the evidence earns the choice.

Cover figure - state becomes useful only when it supports an action whose authority matches the evidence.

MODUS-HELIX 2 APPLIED SCIENCE SERIES / VOLUME III

Public Research Edition v1.0 | 3 October 2026 | BeepSync / MODUS Sports Tech, United Kingdom

Status: company-authored research synthesis and translational paper. Not independently peer reviewed.

READ THIS FIRST

A good athlete model is not yet a good decision system.



Volume I asked what progression means. Volume II asked how a system can build a defensible picture of an athlete whose true performance state is only partly observable. This final volume asks the question that determines whether any of that science is useful: given what we believe about the athlete now, what should happen next?

The obvious answer - choose the workout most likely to improve performance - hides the hardest scientific problem in the entire programme. A state estimate describes what may be true. A training decision changes what happens next. The decision must consider target, phase, uncertainty, alternative explanations, available actions, cost of error, burden, coach constraints and what evidence will be created by the action itself. A system can therefore estimate state reasonably well and still choose badly [1,5-7].

The MODUS-HELIX research programme rejects the idea that intelligence is the ability to produce a confident recommendation on every screen. The stronger definition is narrower: select the smallest action justified by the evidence; request information when it can materially improve the choice; abstain or defer when the consequence exceeds the system's authority; observe what was actually done; and learn from later outcomes without inventing causation [6-10].

This is why the final public volume is a decision book rather than an AI book. It is about constrained choices under uncertainty. The science underneath may eventually include probabilistic state models, individual response estimation, sequential experiments and adaptive policies. The consumer surface should still feel simple: what changed, what matters, what MODUS thinks, what to do next, why, and what would change that decision.

Executive abstract

A combat-athlete progression system should not move directly from measurement to prescription. The defensible chain is evidence -> state -> candidate explanations -> bounded action -> executed work -> response -> later outcome -> updated state. Each link has a different evidential requirement. A valid metric is not automatically actionable; a calibrated state estimate is not automatically a causal diagnosis; a recommendation is not the intervention unless the athlete actually follows it; and a later improvement cannot automatically be credited to the prior recommendation [6,11].

MODUS therefore operates over a bounded action vocabulary: REPEAT, CONSOLIDATE, PROGRESS, CHANGE STIMULUS, TEST, REDIRECT, DELOAD, MAINTAIN and HUMAN REVIEW. Stronger actions require stronger evidence. When two plausible explanations imply different actions, TEST can be more intelligent than guessing. When consequence or uncertainty exceeds authority, HUMAN REVIEW is a scientifically

legitimate output. This creates a selective decision system whose quality is judged by both coverage and risk, not by automation rate alone [6,7,25].

The central validation problem is actionability. Contemporary sport science increasingly distinguishes trustworthy measurement from evidence that the measurement actually changes a decision and, further downstream, improves an athlete-relevant outcome. MODUS must therefore earn authority in stages: deterministic replay, shadow operation, decision-impact studies, pragmatic live pilots, comparative outcome studies and eventually external replication. Complexity is required to outperform credible simpler rules rather than merely appear more sophisticated [6,11-16].

The frontier is individual counterfactual reasoning: for this athlete, under these conditions, which available action is likely to outperform the credible alternatives? Longitudinal history alone cannot identify that answer. Strong individual-response claims require repeated comparable contrasts or experimental variation. Accordingly, the programme defines an authority ladder from transparent deterministic rules through calibrated state estimation, athlete-specific response estimation, controlled individual experimentation and finally bounded adaptive policy selection [7,21,26,27,30-34].

The consumer translation is Progression Intelligence Compression: preserve as much defensible decision value as possible while reducing measurement burden, hardware dependence, cost, latency and cognitive load. The Interaction Governor decides whether additional evidence is worth asking for. The premium experience is therefore not a larger dashboard; it is less decision labour, fewer low-value prompts and a short route from a truthful READ to an executable COMMAND [8].

Contents

SECTION	QUESTION
01	A state is not yet a decision
02	Candidate gaps: plausible, not proven
03	The bounded action space
04	TEST is an action
05	Information has value - and cost
06	Abstention, consequence and human authority
07	The Decision Journal
08	Recommendation is not execution
09	Learning without causal fiction
10	Personalisation must earn authority
11	The decision-authority ladder
12	Decision utility and asymmetric error
13	Validating actionability
14	From replay to outcome evidence

SECTION	QUESTION
15	Human factors and implementation fidelity
16	The consumer decision surface
17	A worked MODUS trace
18	Progression Intelligence Compression
19	The frontier: minimum defensible decision regret
20	Research standard and conclusion
Glossary	
References	

READING RULE

Whenever this paper uses words such as likely, candidate, expected, causal or optimal, the evidential ceiling matters. The language is intentionally stronger only when the underlying research permits the stronger claim.

A STATE IS NOT YET A DECISION

HELIX describes what is currently defensible. MODUS must decide what to do with that belief.

One state can support different actions once target, phase, consequence and alternatives are considered.

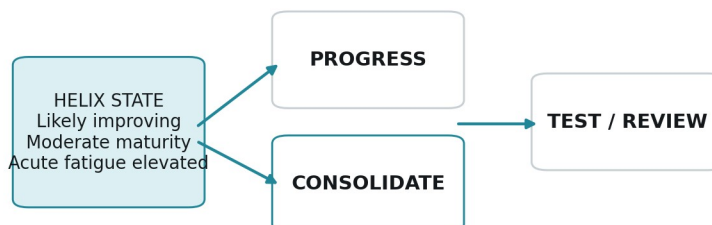


Figure 1.1 - The same athlete state can support different actions after target, phase, uncertainty and consequence are considered.

HELIX and MODUS have deliberately different jobs. HELIX integrates longitudinal evidence and produces domain-specific state envelopes: current estimate, baseline, trend, meaningful-change classification, uncertainty, evidence maturity, freshness, candidate-gap status and unresolved alternatives. MODUS reads that state and chooses among eligible actions. This separation prevents the state model from quietly becoming a prescription engine and prevents the decision engine from manufacturing athlete state to justify a preferred session [5,9,10].

Suppose a boxer has a repeatability benchmark that is probably improving, moderate evidence maturity, adequate recent exposure and elevated acute fatigue. The state is not a command. PROGRESS could be justified if the gain is robust and the next overload is low consequence. CONSOLIDATE could be preferable if the gain is recent and fight week is approaching. TEST could be better if the latest benchmark was collected under poor conditions. A slow-state improvement can therefore coexist with a conservative immediate decision.

This distinction also protects against the most seductive error in consumer training systems: turning a single signal into a prescription. Poor sleep does not automatically mean deload. A high readiness value does not automatically justify progression. A flat benchmark does not automatically mean change stimulus. The decision is a policy over several state components and constraints, not a lookup table attached to one metric [2,5,6].

A useful formalisation is to treat the action as a choice among eligible options conditioned on evidence E, current state X, target T, context C and governed knowledge K. The mathematics can later become probabilistic, but the conceptual requirement exists even in a deterministic Year 1 system.

$$a^* = \operatorname{argmax}(a \text{ in } A_{\text{eligible}}) E[\text{utility}(a) \mid E, X, T, C, K]$$

Conceptual decision rule only. Production authority depends on evidence and validation; the utility function is not yet empirically calibrated.

CORE DISTINCTION

HELIX answers: what is currently defensible about the athlete? MODUS answers: what is the smallest justified thing to do next?

CANDIDATE GAPS: PLAUSIBLE, NOT PROVEN

The largest deficit is not automatically the causal limiter.

Candidate gaps are hypotheses under evidence, not automatic “limiting factors”.

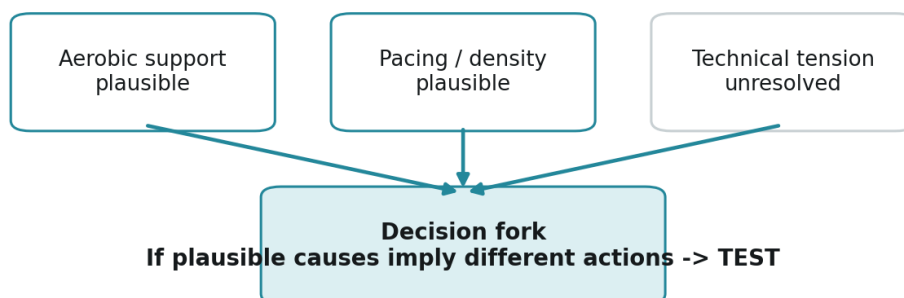


Figure 2.1 - Candidate-gap reasoning preserves competing explanations until the evidence can discriminate between them.

Progression systems often use the language of a 'limiting factor' too casually. A measured deficit can be target-relevant without being the reason performance has stalled. Late-round output may be low because aerobic support is inadequate, because the athlete starts too fast, because technical tension raises energetic cost, because the task became more difficult, or because the most recent observation occurred under acute fatigue. Several explanations can fit the same surface pattern [5,7].

MODUS therefore treats gaps as evidence-qualified candidates. A domain enters the candidate set when it is relevant to the target, sufficiently observed, materially below the athlete's requirement or trajectory, and not already explained by a more immediate modifier. The state should retain the magnitude of the gap, evidence maturity, uncertainty and alternative explanations. The word 'limiter' is reserved for contexts where the design can actually support that stronger claim [7,10].

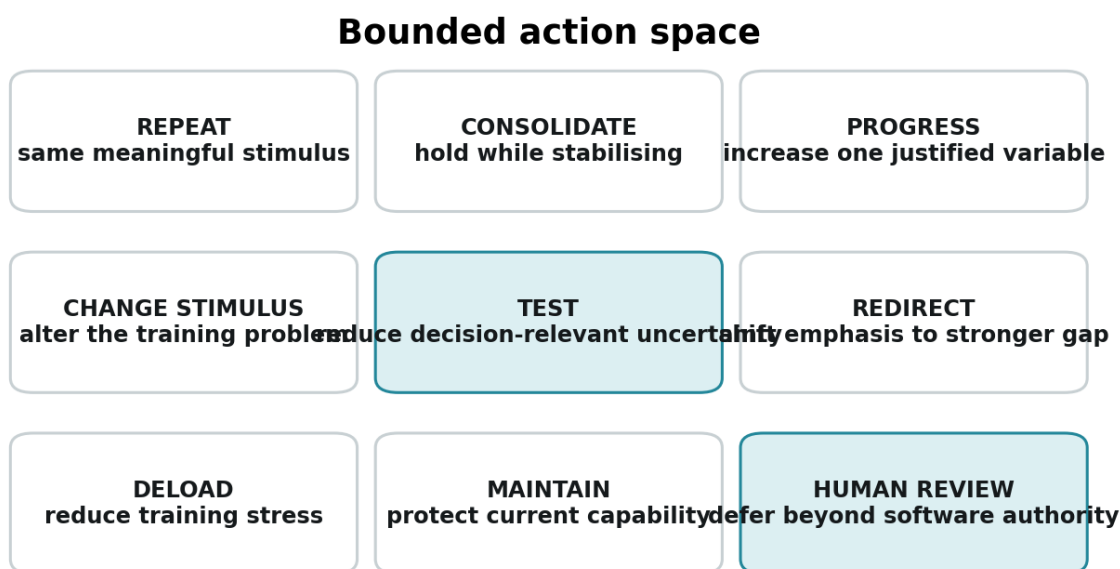
The decision consequence depends on whether competing explanations lead to the same action. If three plausible mechanisms all support one low-risk consolidation step, the system may act while preserving uncertainty. If aerobic limitation implies additional conditioning but technical inefficiency implies coached skill work, guessing would be an overreach. A discriminating benchmark, structured coach observation or repeated rested exposure may have more value than another training change.

This is a general scientific pattern: uncertainty becomes important when it changes the preferred action. The system does not need to resolve every unknown. It needs to resolve the unknowns that alter what should happen next. That principle later becomes the foundation of TEST and the Interaction Governor.

Candidate state	Decision implication	Appropriate MODUS behaviour
One plausible explanation; evidence mature	Action consequence relatively clear	Choose smallest eligible training action.
Several explanations; same low-risk action	Uncertainty does not alter action	Act, preserve alternatives, observe outcome.
Several explanations; different actions	Uncertainty is decision-relevant	TEST or request targeted human evidence.
High consequence + weak identification	Software authority exceeded	HUMAN REVIEW / defer strong claim.

THE BOUNDED ACTION SPACE

Progression intelligence begins with a small vocabulary of defensible changes.



MODUS chooses an action class first. Content routing comes second.

Figure 3.1 - MODUS action vocabulary. The engine selects an action class before it selects content.

A system that recommends from thousands of workouts before deciding what progression problem it is solving is difficult to audit. MODUS reverses the order. It first selects a bounded progression action; only then does the training router find eligible work that implements that action. This keeps the intelligence layer about progression rather than content recommendation [2,8,9].

REPEAT keeps the meaningful stimulus stable when evidence is insufficient, change remains within expected noise or the exposure is still productive. CONSOLIDATE protects a recent gain or stabilises response without adding more overload. PROGRESS increases one justified aspect of dose or complexity. CHANGE STIMULUS alters the training problem when a genuine plateau is sufficiently supported. REDIRECT shifts attention to a stronger target-relevant gap. DELOAD reduces training stress without pretending to diagnose injury or illness. MAINTAIN protects a capability that is not currently the priority. TEST gathers discriminating evidence. HUMAN REVIEW transfers authority to a coach or specialist when the question falls outside the system's validated envelope [2,6].

The vocabulary is intentionally small. It creates a decision record that can be compared across athletes and versions. It also makes error visible. If a later benchmark declines after repeated PROGRESS commands, researchers can inspect whether the state estimate was wrong, whether the action threshold was too aggressive, whether the training was actually completed, or whether the domain simply responded differently than expected. A content-only recommender cannot separate those failures cleanly.

Bounded action does not mean rigid training. Each action can route to many sessions, disciplines and modes. The restriction applies to the scientific claim: why the system changed the athlete's plan. A premium interface can feel varied while the underlying reason remains reproducible.

Action	Question it answers	Typical evidence pattern
REPEAT	Do we need more comparable evidence or consolidation?	Useful stimulus; no clear reason to alter it.
CONSOLIDATE	Should we stabilise rather than overload?	Recent gain, moderate uncertainty or competing priority.
PROGRESS	Has evidence earned a larger challenge?	Positive adaptation, tolerable response, target relevance.
CHANGE STIMULUS	Is the current intervention likely exhausted?	Plateau candidate after adequate exposure/time.
TEST	Would one observation materially change the decision?	Plausible states imply different actions.
REDIRECT	Does another domain matter more now?	Current domain sufficient/improving; stronger gap elsewhere.
DELOAD	Should immediate stress be reduced?	Short-term constraint; not a medical diagnosis.
MAINTAIN	Should capability be protected without further development?	Domain adequate but still target-relevant.
HUMAN REVIEW	Is this outside software authority?	Technical/clinical/safety or high-consequence ambiguity.

TEST IS AN ACTION

Sometimes the most useful intervention is to learn before changing training.



Testing is often treated as something that happens before the 'real' programme. In a longitudinal progression system, a test can itself be the correct next action. The reason is simple: when uncertainty is decision-relevant, reducing that uncertainty can have greater value than immediately applying another training stimulus [7,8].

The important word is discriminating. A test is not useful merely because it generates data. It should be chosen because plausible results would lead to different actions. A rested repeatability benchmark might separate persistent plateau from transient suppression. A coach rating under a defined technical rubric might distinguish exposure from proficiency. A short field benchmark might determine whether a Roadwork state is stale enough to block progression. If the result would not alter the action, the test is probably burden rather than intelligence.

This reframes uncertainty from a weakness to an operational object. MODUS can say: 'I cannot yet distinguish A from B; this specific observation is the smallest thing that would change the decision.' The result is more credible than a confident recommendation based on an unidentifiable state, and more useful than a vague message that simply says more data are needed.

TEST also creates a discipline against data hunger. The system should not ask for every scientifically interesting measurement. It should ask for the evidence with the highest expected decision value relative to burden, delay and cost. This is especially important in a consumer product where attention is finite.

TEST PRINCIPLE

Testing is justified by the decision fork it resolves, not by the number of metrics it can collect.

INFORMATION HAS VALUE - AND COST

Minimum sufficient evidence turns measurement burden into part of the scientific problem.

Ask only when new evidence can change the decision enough to justify the burden

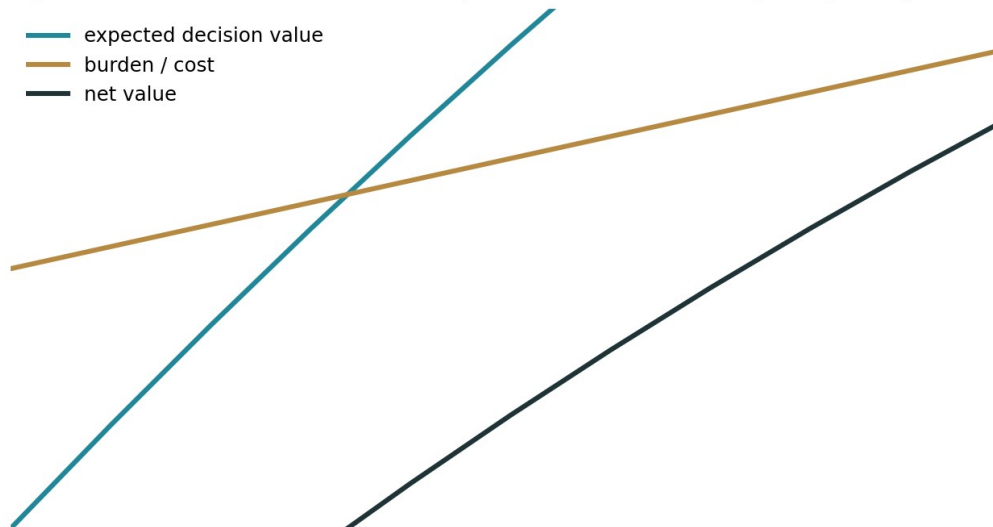


Figure 5.1 - Illustrative information-value curve. More evidence is useful only until its decision value is overtaken by burden and delay.

Value-of-information reasoning asks a practical question: how much better could the decision become if we knew the answer to one more question or test? The Frontier Thesis uses expected value of sample information as a conceptual tool for deciding whether to gather evidence before acting. Consumer translation adds the missing reality: evidence acquisition has a burden [7,8].

For a professional performance department, burden may mean staff time, laboratory access, equipment and athlete interruption. For a consumer athlete, it may mean a questionnaire after every session, a benchmark inserted into a busy week, device setup, a subscription to another platform or simple attention fatigue. An input that improves a model slightly but makes the product hard to live with can reduce the quality of the system in practice.

Minimum Sufficient Evidence therefore asks for the smallest evidence set capable of supporting the current decision at the required consequence level. The threshold should rise with consequence. A low-risk off-camp progression can tolerate more uncertainty than a major fight-week change. Conversely, no amount of low-value prompting should be used merely to make the interface look scientific.

The Interaction Governor described later is the product expression of this principle. It allocates the athlete's attention according to decision value. The athlete may experience long periods of nearly invisible monitoring followed by a very specific request when the state reaches a genuine fork.

$$NetVOI(q) = EVSI(q) - burden(q) - delay(q) - friction(q)$$

Conceptual design objective, not a validated utility function. Ask q only when the expected improvement in the decision is worth its cost.

ABSTENTION, CONSEQUENCE AND HUMAN AUTHORITY

The ability not to automate is part of scientific competence.

Abstention can improve reliability - but zero coverage is not useful

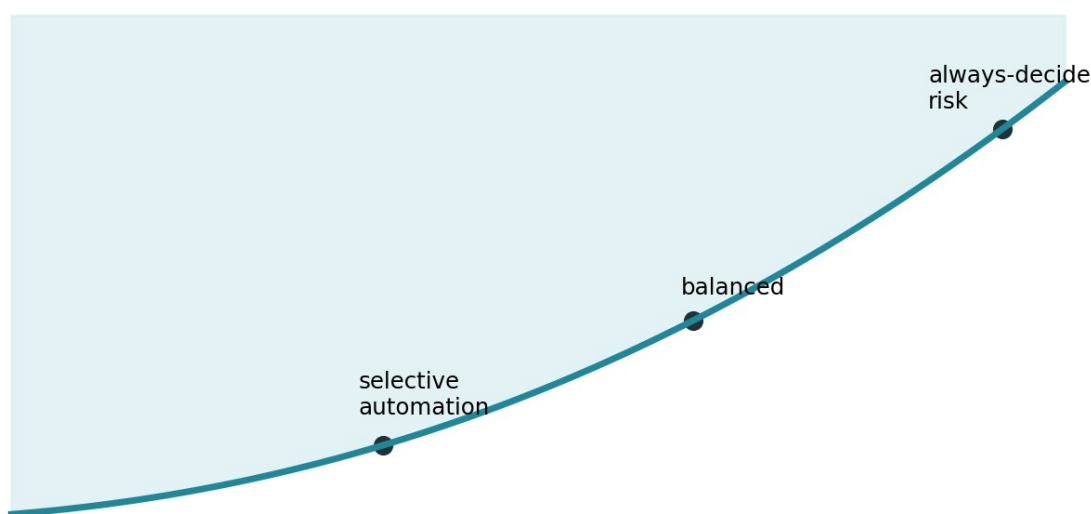


Figure 6.1 - Selective automation should be judged by both coverage and risk.

A system that always decides can look more capable than one that sometimes refuses. Statistically, the opposite can be true. Selective prediction formalises the idea that abstaining on difficult cases can reduce error among the cases that remain. In MODUS, TEST, REPEAT and HUMAN REVIEW provide operational forms of selective decision-making [6,25].

Validation should therefore report risk and coverage together. Coverage is the proportion of eligible decision points on which MODUS issues an automated progression action. Risk is the consequence-weighted error among those decided cases. A system that reduces error by abstaining on nearly everything has little practical value. A system that maximises coverage regardless of uncertainty may create confident mistakes. The operating threshold should be earned from decision consequences, not chosen to make automation statistics look impressive [6].

Human authority is not a fallback for an unfinished algorithm. Combat sport contains domains the phone cannot legitimately observe at sufficient resolution: technical quality, tactical interaction, injury

diagnosis, concussion clearance and other specialist questions. Coaches also possess context the software may not have. The architecture should therefore be coach-compatible, capable of structured coach augmentation and explicit about when a human has final authority [7,8].

Override is evidence rather than disobedience. If a coach repeatedly rejects a MODUS action because the athlete is preparing for a tactical opponent-specific need that the system cannot observe, the override exposes a missing contextual variable or a legitimate policy boundary. A mature system records that event and learns about its own limits instead of treating acceptance rate as the goal [6,16].

$$\text{Coverage}(\tau) = P(\text{confidence} \geq \tau) \quad \text{Risk}(\tau) = E[\text{loss} \mid \text{confidence} \geq \tau]$$

A useful selective policy seeks an operating region where decided cases remain valuable without forcing decisions in low-authority states.

THE DECISION JOURNAL

A progression system needs a record of what it believed, what it chose and what reality did next.

The Decision Journal separates recommendation, execution, response and later evidence.

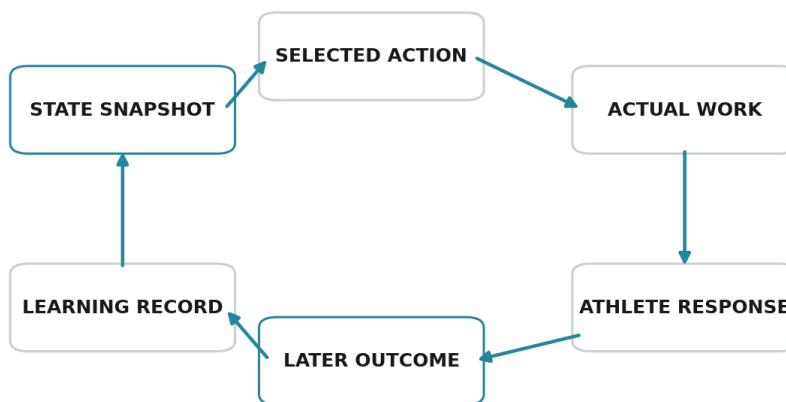


Figure 7.1 - Decision Journal. Recommendation, execution, response and later evidence remain separately recorded.

The Athlete Evidence Ledger records what happened. The HELIX State Envelope records the current interpretation. The Decision Journal records the policy event: what MODUS believed at that moment, which actions were eligible, which action was selected, why alternatives were rejected, whether a user overrode the recommendation, what work actually occurred and which future observation was expected to adjudicate the decision [6-10].

This separation is vital for scientific learning. Without it, a system may later observe improvement and attribute it to a recommendation that was never followed. Or it may see a poor outcome and blame a policy even though the athlete substantially altered the prescribed work. The decision record creates a chain of custody for intervention logic.

The journal should be versioned. State version, knowledge version, policy version and explanation version all matter. If a 2028 policy is replayed against a 2026 athlete history, researchers should be able to compare what the old system actually decided with what the new system would have decided. Historical truth and contemporary reinterpretation are different analyses.

For the consumer, almost none of this needs to be visible by default. The athlete may see only a short decision history: 'Consolidated after mixed evidence; rested retest later confirmed improvement.' The deep record exists so that the simple story can remain honest.

Decision Journal field	Scientific purpose
Target + phase	Defines what outcome matters and how costly an error may be.
State snapshot	Preserves the evidence available when the decision was made.
Eligible / rejected actions	Makes policy boundaries and alternatives auditable.
Selected action + reason	Records the actual recommendation claim.
Override / defer reason	Captures human context and authority boundaries.
Actual dose / content	Separates recommendation from execution.
Expected future evidence	Defines how the decision can later be evaluated.
Later outcome	Closes the loop without automatically claiming causation.

RECOMMENDATION IS NOT EXECUTION

The outcome chain breaks if the action was never delivered.



Outcome learning fails when these links are collapsed.



An unexecuted recommendation cannot be credited or blamed for a later outcome.

Figure 8.1 - Recommendation, execution, acute response and persistent adaptation are different objects.

Outcome learning sounds straightforward: recommend something, observe what happens, update the system. In practice, at least four objects must remain separate. The recommendation is the policy output. Execution is the work the athlete actually completed. Acute response is what happened during or immediately after that work. Adaptation is persistent change detected later under qualified measurement [1,2,6].

Each gap creates a different failure mode. A recommendation can be scientifically sensible but ignored. It can be accepted but poorly implemented. It can be executed exactly yet produce an adverse acute response. It can be well tolerated in the short term but fail to create the desired adaptation. A system that treats all four as one 'session outcome' cannot diagnose where its policy failed.

Implementation fidelity is therefore part of the scientific model. Prospective studies should record whether the user saw the recommendation in time, whether it was accepted, how it was modified, how closely the completed dose matched the intended action and whether the evidence needed for later evaluation was actually collected. This mirrors the broader actionability argument in contemporary applied sport science: the pathway, not the isolated metric, is the relevant unit of evaluation [6,11].

The same principle protects the athlete from false learning. If MODUS recommends PROGRESS but the athlete substitutes a much harder unscheduled session, a subsequent poor benchmark should not automatically reduce confidence in the original action threshold. Reality has to be represented before learning can be trusted.

LEARNING WITHOUT CAUSAL FICTION

Longitudinal memory is useful. It is not automatically evidence that one training choice caused the next outcome.

The phrase 'MODUS learns what works for you' is attractive but scientifically dangerous. Observational training history is heavily confounded by prior decisions. Athletes receive harder work when they are fitter, easier work when they are fatigued, different work when competition approaches and different work when coaches observe unrecorded problems. If later performance improves, the intervention and the context that caused the intervention are entangled [7,31,32].

Year 1 outcome learning should therefore be modest. The system can update descriptive response histories, calibration and context-conditioned patterns: this athlete has tended to report lower exertion after comparable exposures; this benchmark has remained stable across the last three consolidation blocks; this type of progression has often been overridden by the coach. These are useful facts without being individual treatment effects.

A stronger claim - action A is expected to outperform action B for this athlete - requires stronger identification. Repeated testing and individual-response methodology show why a single pre/post change cannot establish a stable response phenotype. The strongest individual claims generally require repeated comparable contrasts or experimental variation that separates treatment effect from day-to-day noise and regression to the mean [21,33,34].

Causal machine-learning methods may eventually help estimate heterogeneous treatment effects, but sophisticated modelling does not remove the need for identifying assumptions and adequate data. The research programme therefore treats causal response as an earned capability rather than a marketing default [7,32].

FRONTIER RULE

Longitudinal data are memory. They become stronger causal evidence only when the design creates informative contrasts and valid outcomes.

PERSONALISATION MUST EARN AUTHORITY

Knowing an athlete for a long time is not the same as knowing their counterfactual response.

Personalisation operates at several levels. The easiest level is contextual: the athlete's target, discipline, phase and available equipment alter what is relevant. The next is longitudinal: the athlete's own baseline and trajectory outrank generic population norms for claims of personal change. A stronger level is parameter personalisation: the model estimates athlete-specific response tendencies. Strongest of all is counterfactual personalisation: it can support the claim that one eligible action is expected to work better than another for this athlete now [3,5,7].

These levels should not be collapsed into a single 'personalised' label. A programme can be highly personalised in content while having no evidence about individual causal response. Conversely, a carefully designed N-of-1 comparison can provide strong evidence about one narrow response while saying nothing about the athlete's technical or tactical development [21].

Population knowledge remains useful at cold start. Hierarchical methods can supply conservative priors or plausible ranges until individual evidence becomes informative. The important governance rule is that population information should lose influence when robust athlete-specific evidence accumulates, while a small number of noisy sessions should not be allowed to manufacture a stable personal response phenotype [5,7].

For the consumer, this means the system can become more individually useful before it becomes individually causal. It can remember the athlete's history, preserve their baselines, learn their normal variability, adapt evidence requests and choose target-relevant actions without claiming to know their optimal dose. That distinction makes the product more credible, not less personal.

Personalisation level	Defensible claim	Evidence needed
Contextual	The recommendation reflects target, phase and constraints.	Known context + governed rules.
Longitudinal	This athlete changed relative to their own history.	Comparable repeated evidence + error model.
Response tendency	The athlete has tended to respond this way under similar recorded conditions.	Repeated comparable action-response records.
Individual counterfactual	A is expected to outperform B for this athlete.	Repeated/experimental contrasts or strong causal identification.
Adaptive policy	A decision rule improves outcomes across repeated states.	Prospective policy evaluation + transport + harms.

THE DECISION-AUTHORITY LADDER

The system should become more powerful only after simpler generations have been validated.

Authority must climb only after evidence and validation earn the next generation.

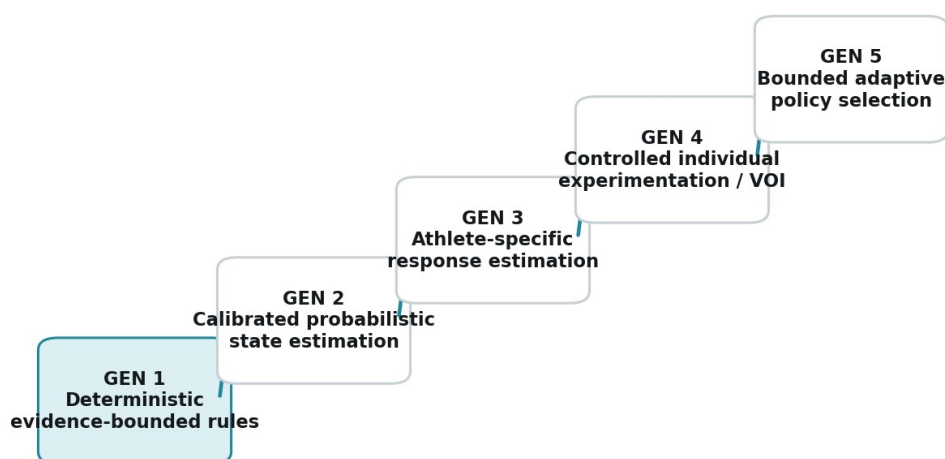


Figure 11.1 - MODUS authority ladder. Advanced modelling is a research privilege, not a launch requirement.

The Frontier Thesis converts ambition into a staged research programme. Generation 1 is deterministic evidence-bounded progression: transparent baselines, meaningful-change gates, explicit uncertainty rules and a small action vocabulary. Generation 2 adds calibrated probabilistic state estimation where data density and validation justify it. Generation 3 estimates athlete-specific response tendencies. Generation 4 uses controlled individual experimentation and value-of-information testing. Generation 5 supports bounded adaptive policy selection [7].

The order matters. A system should not jump from descriptive app history to an adaptive-policy claim simply because modern machine-learning tools exist. Every generation introduces new assumptions and new ways to be confidently wrong. Probabilities require calibration. Individual response requires identifiability. Experiments require ethical and practical control. Adaptive policies require evidence that the policy itself improves decisions or outcomes compared with credible simpler approaches [6,7].

This ladder also gives product development a useful escape from AI theatre. The first production version can be scientifically serious while deliberately simple. Transparent deterministic logic may outperform a more sophisticated model if the available data are sparse, heterogeneous or poorly observed. Complexity earns its place only when it changes the decision for the better.

A mature system can therefore have different generations in different domains. Aerobic progression may operate with calibrated probabilistic state while technical/tactical inference remains human-led. One validated domain should never confer authority on another.

GENERATION RULE

The next model generation is promoted only when it demonstrates incremental calibration or decision utility beyond the current transparent baseline.

DECISION UTILITY AND ASYMMETRIC ERROR

Not all wrong decisions cost the athlete the same amount.

The cost of the same wrong action changes with phase and consequence.

Off-camp	REPEAT error cost 1	PROGRESS error cost 1	TEST error cost 2	DELOAD error cost 0
Heavy block	REPEAT error cost 1	PROGRESS error cost 2	TEST error cost 1	DELOAD error cost 2
Fight week	REPEAT error cost 1	PROGRESS error cost 4	TEST error cost 2	DELOAD error cost 3

Figure 12.1 - Illustrative error asymmetry. The cost of a mistake changes with phase and consequence.

Accuracy is an incomplete measure of a progression policy because action errors have unequal consequences. An unnecessary repeat in an early off-camp block may cost little. An unjustified progression during fight week may be far more consequential. A burdensome test can be harmless in a development phase and disruptive in taper. A false deload can waste adaptation time; a false progress can impose avoidable stress [6,7].

Decision analysis provides a general language for this problem. A policy should weigh target benefit against burden, avoidable error, risk and irreversibility. The programme does not yet claim empirically calibrated numerical utilities. Year 1 should use conservative ordinal consequence classes and hard safety constraints rather than pretending that every trade-off can be reduced to a precise score [6,24].

This is another reason why the smallest justified change is often preferable. Reversible actions allow the system to collect more evidence with limited downside. Large irreversible changes require stronger evidence. A bounded policy can therefore be intelligent even when its state model is imperfect, because it limits the consequence of uncertainty.

The same logic applies to uncertainty thresholds. The acceptable uncertainty for PROGRESS should be lower when the consequence of error is high. The correct threshold is therefore domain- and phase-specific, not a universal confidence percentage.

$$\text{Decision utility} = \text{target benefit} - \lambda_b \cdot \text{burden} - \lambda_e \cdot \text{avoidable error} - \lambda_r \cdot \text{risk} \\ - \lambda_i \cdot \text{irreversibility}$$

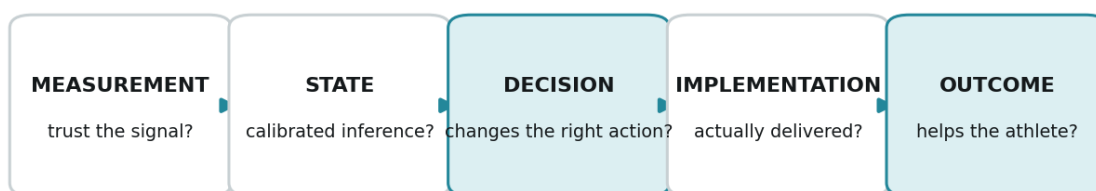
Conceptual governance equation. Hard safety boundaries remain non-tradeable.

VALIDATING ACTIONABILITY

A trustworthy metric does not become useful decision support merely because it appears on a dashboard.



Validation authority is lost if any link is unsupported.



Measurement -> state -> decision -> implementation -> athlete-relevant outcome

Figure 13.1 - Measurement validity, state calibration, decision impact, implementation and outcome benefit form a chain.

The strongest recent shift in applied monitoring science is the move from asking whether a signal is valid to asking whether using that signal changes a decision and improves or protects an athlete-relevant outcome. Besson and Stucky describe actionability as a claim about decision utility rather than an inherent property of a metric. MODUS adopts that distinction directly [11].

The validation chain has five links. Measurement asks whether the observation is valid, reliable and responsive in the intended protocol. State inference asks whether HELIX estimates are calibrated and appropriately uncertain. Decision impact asks whether MODUS changes the action that would otherwise have occurred. Implementation asks whether the recommendation was actually delivered with acceptable burden and fidelity. Outcome asks whether the full pathway improves or protects a valid athlete-relevant endpoint [6,11-16].

A system can pass one link and fail the next. A reliable benchmark may be irrelevant to the next decision. A calibrated state may produce the same action a coach already intended. A recommendation may change behaviour but worsen the target outcome. A good policy may fail in practice because the athlete

ignores burdensome prompts. The product claim should therefore stop at the strongest link that has actually been demonstrated.

This creates an important discipline for MODUS language. 'Recorded' is a data claim. 'Likely improving' is an inferential claim. 'Can inform next-session decisions' is a decision-impact claim. 'Improves decision quality' is a comparative utility claim. 'Demonstrated performance benefit' is an outcome claim. Those statements are not interchangeable.

Authority level	Strongest permitted claim	Evidence requirement
A0 - Descriptive	Recorded / observed qualified evidence.	Data integrity and replay.
A1 - Inferential	Estimated / likely / uncertain state.	Prospective state calibration.
A2 - Decision-impact	Can inform a defined training decision.	Paired pre/post decision study.
A3 - Decision-utility	Improves decision quality or reduces avoidable error/burden.	Prospective comparative decision study.
A4 - Outcome-supported	Improves/protects athlete-relevant outcome in tested scope.	Randomised/pragmatic causal evaluation + harms.

FROM REPLAY TO OUTCOME EVIDENCE

A decision engine should be observed before it is trusted, and compared before it is celebrated.

A recommendation should become visible before it becomes authoritative.

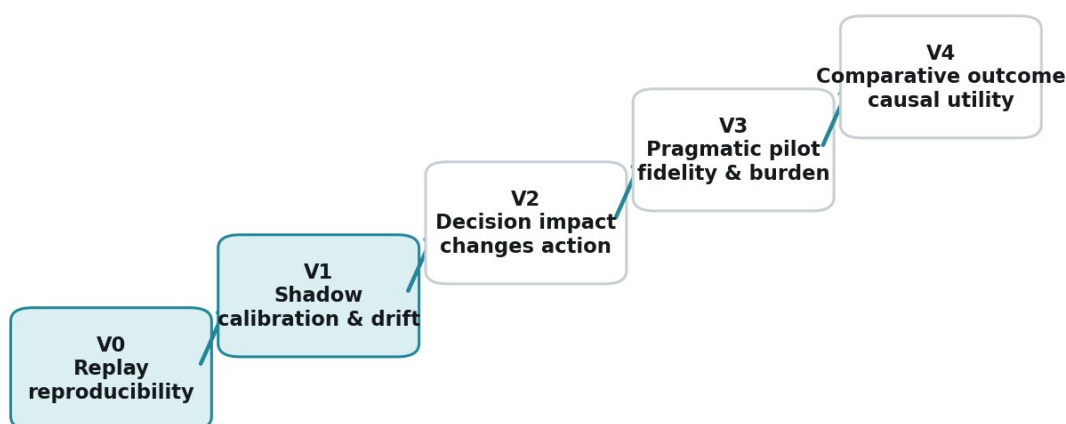


Figure 14.1 - Validation ladder from deterministic replay to comparative outcome evidence.

Retrospective replay is the first gate because it can expose impossible actions, versioning failures and contradictions. Existing athlete histories can be processed through fixed knowledge and policy versions to test deterministic reproduction and later-outcome calibration. Replay cannot show that an athlete would have improved under a recommendation they never received. Counterfactual benefit should not be invented from historical logs [6].

Shadow mode is the critical next step. MODUS receives live evidence and produces timestamped recommendations, but users do not see them. Researchers can then observe calibration, missingness, instability, disagreement with real coaching decisions, abstention frequency and model drift without allowing experimental policy to alter training. Shadow operation also prevents hindsight leakage because the recommendation exists before the later outcome is known.

Decision-impact studies then reveal MODUS only after the coach or athlete has committed to an initial plan. The post-reveal decision can be accepted, modified, rejected or deferred. This creates a direct record of how the information changed action. High action-change rate is not automatically desirable; the goal is change concentrated in cases where the additional evidence genuinely matters [6].

Pragmatic pilots expose a narrow validated domain to live use. They test the whole operational pathway: evidence collection, recommendation latency, explanation, override, adherence, completed dose and later benchmark. Only after these stages should comparative outcome designs ask whether MODUS-supported progression improves a target outcome relative to a credible baseline. Parallel randomised trials, cluster designs, SMARTs, microrandomized trials and N-of-1 methods each answer different questions; none is a universal validation template [6,21,26,27,30].

Stage	What it can establish	What it cannot establish
V0 Replay	Reproducibility, impossible paths, policy coverage.	Counterfactual athlete benefit.
V1 Shadow	Calibration, drift, abstention, disagreement map.	Effect of visible advice on behaviour.
V2 Decision impact	Whether information changes action and why.	That changed action improves outcome.
V3 Pragmatic pilot	Fidelity, burden, near-term operational value, harms.	Broad causal superiority.
V4 Comparative outcome	Causal utility in tested domain/population.	Automatic transport to other domains or users.

HUMAN FACTORS AND IMPLEMENTATION FIDELITY

The pathway fails if the athlete cannot understand, trust or execute it appropriately.

Athlete monitoring research repeatedly shows that implementation depends on more than metric quality. Feedback, coach support, athlete beliefs, time burden and monitoring culture affect whether evidence is reported honestly and whether recommendations are acted upon. These are not cosmetic product issues; they change the data-generating process and therefore the scientific validity of the system [15,16,29].

MODUS explanations should expose the target, dominant evidence, uncertainty, action boundary and the condition that would change the recommendation. Too little explanation creates opaque authority. Too much can move the cognitive burden back onto the athlete. Human-AI research in other decision-support domains warns that explanations can improve appropriate reliance but can also create automation bias when confidence is misplaced [35-37].

The intended relationship is collaborative rather than autonomous. The athlete can accept, override, defer or ask a coach. The coach can contribute structured evidence or impose constraints. The system retains a visible reason and records the override. This makes disagreement scientifically useful: it can reveal missing variables, poor translation, policy error or a legitimate human authority boundary.

Implementation fidelity should therefore be measured in validation: was the recommendation displayed in time, understood, accepted or overridden for a recorded reason, translated into the intended dose and followed by the evidence needed to evaluate it? A recommendation that never reached the athlete is not part of an intervention.

Implementation variable	Why it matters	Example measure
Exposure to recommendation	No exposure means no causal pathway.	Display success, latency, acknowledgement.
Comprehension	Misunderstood advice can invert the intended action.	Paired interpretation / recall check.
Acceptance / override	Separates software advice from human choice.	Accept, modify, reject, defer.
Dose fidelity	Outcome cannot be credited to unexecuted work.	Planned vs completed dose distance.
Evidence adherence	Missingness may become selective and biased.	Response rate, pattern of omissions.

Implementation variable	Why it matters	Example measure
Burden	Useful intelligence can still be impractical.	Prompts, seconds/session, dropout, satisfaction.

THE CONSUMER DECISION SURFACE

Elite-grade reasoning should arrive as a small, truthful next step.

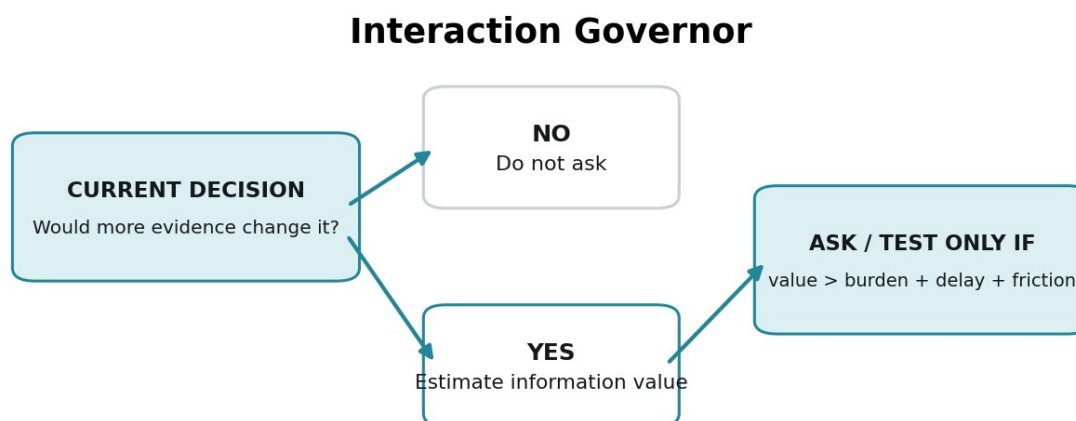


Figure 16.1 - Interaction Governor. The app asks only when additional evidence can materially change the current decision.

The consumer product should not expose the internal decision system as a sports-science dashboard. The athlete should not have to analyse state vectors, confidence decomposition or competing hypotheses before training. The internal complexity is valuable precisely because it allows the external interaction to remain small [8].

The Consumer Decision Surface answers five questions. What changed? summarises only qualified new evidence. What matters? identifies the target-relevant domain rather than every available metric. What MODUS thinks expresses state with calibrated uncertainty. What to do next exposes one bounded action, test or escalation. Why shows the dominant evidence and the condition that would change the decision. This is semantic compression, not simplification by overclaim [8].

The Interaction Governor controls when the product interrupts the athlete. A fixed daily questionnaire is rejected as a default architecture because the same question does not have the same decision value every day. The governor asks whether uncertainty is currently action-relevant and, if so, which question or test offers the largest improvement relative to burden. This makes user effort reciprocal: the system can explain why it needs an input and what decision fork that input will resolve.

Premium, in this model, means reduced decision labour. The system remembers the work, knows which evidence is relevant, suppresses low-value prompts, preserves uncertainty internally and brings the athlete directly to executable training. More science underneath should result in less work for the user, not a larger screen of numbers.

Consumer surface: the explanation is smaller than the science, not stronger than it.

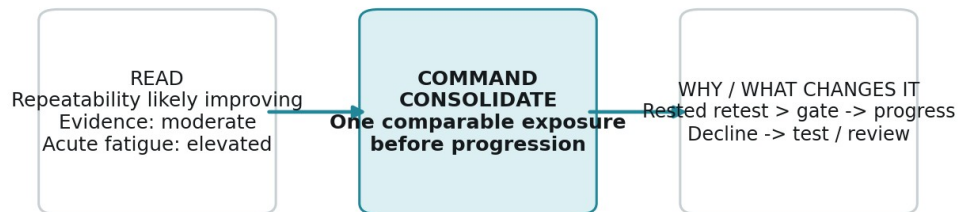


Figure 16.2 - The athlete-facing surface is deliberately smaller than the internal decision record.

TRANSLATION FIDELITY

The consumer statement may be shorter than the internal reasoning, but it must not become more certain, broader in construct or stronger in causal meaning.

A WORKED MODUS TRACE

What a conservative next decision looks like when the evidence is good but not complete.

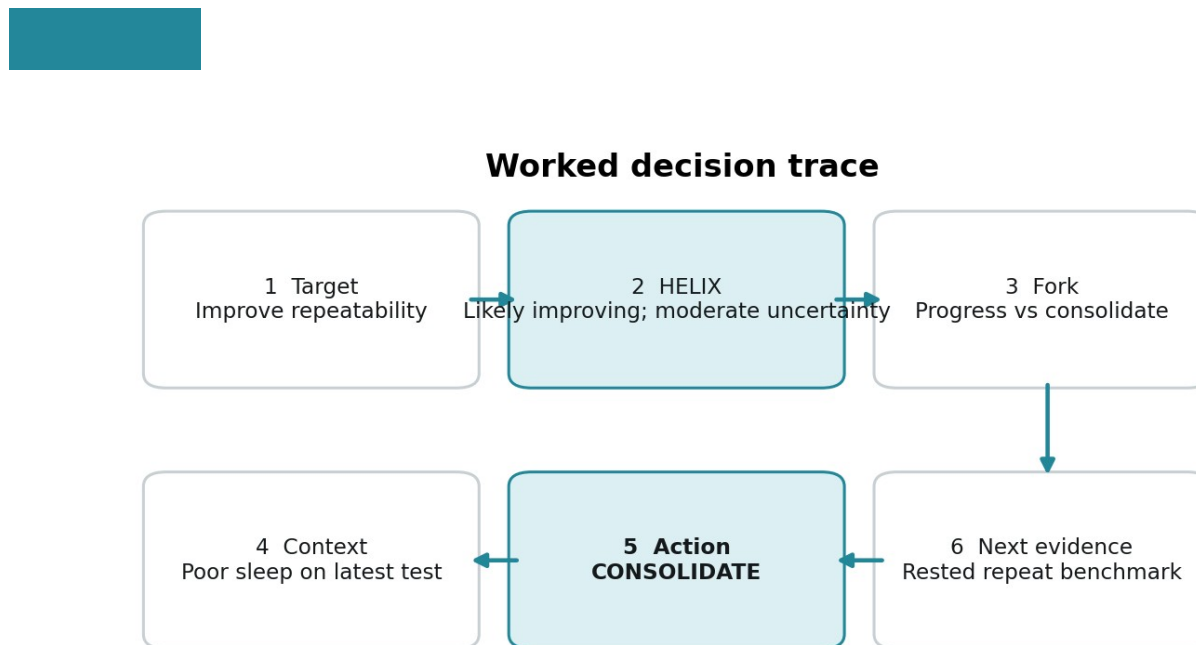


Figure 17.1 - Hypothetical decision trace. The purpose is to demonstrate reasoning discipline, not a validated product output.

Consider an amateur boxer in an eight-week development block targeting improved repeatability. HELIX has six standardised combat-conditioning benchmark results, eighteen session-response records and adequate exposure under a stable taxonomy. The latest benchmark is 3.2% better than the previous block reference, but the uncertainty interval overlaps the practical-change threshold. Comparable sessions feel easier, yet the athlete slept poorly before the latest test. The slow state is probably improving; evidence maturity is moderate; acute context is unfavourable [5].

A simplistic system could report '+3.2% conditioning' and progress density immediately. MODUS should not. The latest test provides positive evidence, but it cannot cleanly separate persistent gain from context and ordinary variability. PROGRESS is eligible but not dominant. CONSOLIDATE preserves the productive stimulus while protecting against a false positive. TEST - one rested repeat benchmark - has high information value because it could change the next decision.

The command is therefore CONSOLIDATE one comparable exposure, then repeat the benchmark under rested standardised conditions. The reason is not 'you are tired'. It is: evidence points upward, but the measurement that would justify a stronger progression was collected under a condition that increases ambiguity. If the rested retest exceeds the expected-error and worthwhile-change gates, PROGRESS becomes stronger. If it returns to baseline, the state remains stable or uncertain. If it declines despite adequate exposure and comparable conditions, a plateau or redirect investigation becomes more plausible.

Notice what MODUS does not claim. It does not call the athlete fitter overall. It does not infer technical improvement. It does not claim that consolidation will cause better performance. It selects the smallest action consistent with current evidence and names the next observation that can increase authority.

Internal object	Hypothetical value	Meaning
Target	Repeatability over current block	Defines the decision objective.
Slow state	Likely improving	Positive trend, not yet decisive practical change.
Fast modifier	Poor sleep before latest test	Raises ambiguity; does not erase slow state.
Candidate actions	CONSOLIDATE / PROGRESS / TEST	Multiple actions remain eligible.
Selected action	CONSOLIDATE	Smallest justified training change.
Information need	Rested repeat benchmark	Could move decision to PROGRESS or back toward TEST.
Claim ceiling	Adaptation evidence only	No causal optimal-dose or technical claim.

PROGRESSION INTELLIGENCE COMPRESSION

The scientific challenge is to preserve decision value while stripping away institutional overhead.

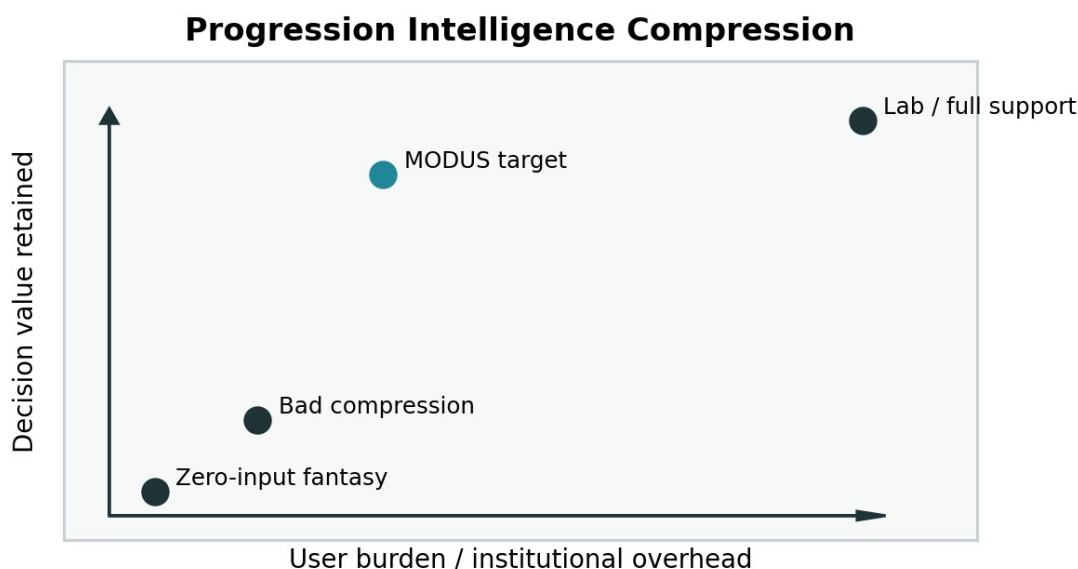


Figure 18.1 - Progression Intelligence Compression seeks high retained decision value at consumer-level burden.

High-performance environments distribute progression decisions across coaches, physiologists, strength-and-conditioning staff, analysts, laboratories and specialist software. A consumer app cannot reproduce those institutions by shrinking their dashboards onto a phone. It must compress functions rather than professions [8].

The Progression Intelligence Compression Problem therefore asks how much defensible decision value can survive when measurement burden, hardware dependence, cost, latency and cognitive load are constrained. Compression succeeds when the system removes operational overhead while preserving scientific meaning. It fails when fewer measurements become fabricated certainty, when simplified wording becomes a stronger claim or when an opaque score replaces a domain-specific inference [8].

The design mechanisms are now visible across all three volumes. Minimum Sufficient Evidence controls what the system asks for. Progressive evidence resolution lets optional sensors and experts increase resolution without creating a second athlete model. HELIX protects construct meaning and uncertainty across time. MODUS selects a bounded action rather than a workout first. The Interaction Governor spends attention only when information can change the decision. The Decision Journal allows later outcomes to update the system without rewriting history.

This architecture is especially important for serious amateur and semi-professional combat athletes. They may have intermittent coaching, no laboratory, limited specialist testing and only optional wearables. The product should still be useful at low evidence resolution, becoming more precise when qualified evidence is available rather than making hardware a prerequisite for intelligence.

CONSUMERISATION RULE

The science may be institutional in depth; the experience must be consumer in burden. Complexity should disappear from the interaction without disappearing from the reasoning.

THE FRONTIER: MINIMUM DEFENSIBLE DECISION REGRET

The highest standard is not knowing everything. It is choosing well despite not knowing everything.

The frontier is disciplined decision-making under bounded uncertainty - not omniscience.

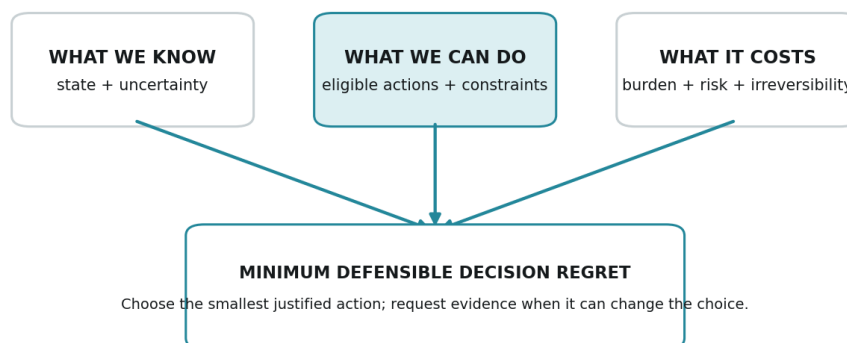


Figure 19.1 - The frontier objective: minimise avoidable decision regret while respecting uncertainty, burden and authority.

The Frontier Thesis asks what a benchmark combat-athlete progression system would have to know, infer, test and prove. Its conclusion is deliberately less dramatic than the language often attached to AI: the frontier is sequential decision-making under partial observability and causal uncertainty. The athlete's true state is hidden; training is both intervention and evidence; adaptations occur on different time scales; technical and tactical performance depend on an opponent; and historical training choices are confounded by why those choices were made [7].

The strongest long-term question is counterfactual: for this athlete, at this point in time, under these constraints, which available action is most likely to improve the target relative to credible alternatives? State estimation is necessary but not sufficient. Knowing that repeatability is stagnant does not prove that more intervals, a density change, a technical redistribution or simple consolidation is the best causal intervention.

The programme therefore frames the practical frontier as minimum defensible decision regret under bounded uncertainty. The objective is not to eliminate uncertainty or maximise intervention frequency. It

is to choose the action whose expected value is strongest after accounting for burden, risk, irreversibility and what the evidence does not identify - while preserving the option to TEST or defer. This is an original research framing, not a validated scalar performance metric [7].

The one-sentence standard is: know which action is justified now, what evidence would change that action, how uncertain the decision is, and whether repeated outcomes demonstrate that the policy deserves greater authority. A system reaches the frontier not by claiming benchmark status, but by prospectively beating credible simpler policies, transporting across intended populations and remaining able to lose authority when calibration fails.

$$a^*_t = \operatorname{argmax}_a E[V_T(x_{\{t+h\}}) - \lambda_b B(a) - \lambda_r R(a) - \lambda_u U(a) - \lambda_i I(a) \mid \text{evidence}]$$

Frontier conceptual objective: choose among eligible actions while penalising burden, risk, uncertainty and irreversibility. It is a research specification, not a validated clinical or sport score.

RESEARCH STANDARD AND CONCLUSION

Authority is something the system earns, domain by domain, and can lose.



Across the research programme, one principle recurs: no component should inherit authority from a more impressive neighbouring component. A validated sensor does not validate the state model. A calibrated state model does not validate the decision policy. A policy that changes decisions does not prove benefit. A successful aerobic domain does not validate technical inference. The claim ceiling follows the strongest demonstrated link in the exact domain, population, protocol and consequence range tested [3-7].

The research standard for MODUS is therefore deliberately adversarial. Compare complex policies with transparent safe rules, the current production system, conservative repeat/no-change strategies and human decisions where appropriate. Track calibration as well as discrimination. Report abstention and coverage. Record burden and implementation. Use athlete-relevant outcomes independent of proprietary scores. Stratify by intended populations and phases. Requalify after material rule, model or protocol changes. External replication is required before benchmark claims become credible [6,7].

This standard also protects the product from its own success. As more athlete histories accumulate, it will become tempting to promote correlations into personal causal claims. As probabilistic models become more accurate, it will become tempting to hide residual uncertainty. As recommendations become easier to deliver, it will become tempting to optimise engagement or acceptance. The governing objective should remain the athlete's best interest and the smallest scientifically justified action.

The three public volumes therefore end where the product loop begins. Volume I defined progression as an inference problem. Volume II showed how qualified evidence can become a longitudinal athlete state without collapsing the athlete into a score. Volume III shows how that state can support a bounded decision, how the decision can request evidence, how reality can test it and how authority can expand only after validation. The resulting loop is simple to say and difficult to deserve: WORK -> EVIDENCE -> READ -> COMMAND -> WORK.

The scientific ambition is substantial, but the consumer promise should remain disciplined. MODUS-HELIX is not an omniscient coach, a universal fighter score, a medical system or an automatic judge of technique. Its credible destination is a progression operating system that knows what it can observe, what it can infer, what it cannot yet distinguish, what action is justified now and what evidence would earn a stronger decision next time.

MODUS-HELIX 2 Applied Science Series



Evidence -> state -> decision -> outcome -> learning

Figure 20.1 - The public trilogy: progression science -> athlete state -> bounded decision.

FINAL THESIS

The highest form of progression intelligence is not confident prediction. It is disciplined action under uncertainty, with authority restricted to what the evidence has earned.

G /

GLOSSARY

Core terms used across the MODUS-HELIX 2 Applied Science Series.

TERM	WORKING DEFINITION
Abstention	A deliberate refusal to automate a strong action when evidence or authority is insufficient.
Actionability	Evidence that information belongs to a defined decision pathway and, at stronger levels, improves decision utility or athlete-relevant outcomes.
Candidate gap	A target-relevant deficit or state discrepancy supported strongly enough to consider, but not automatically a proven causal limiter.
Claim ceiling	The strongest statement or action permitted by the current evidence and validation status.
Decision impact	Whether MODUS changes the action that would otherwise have been selected.
Decision Journal	Versioned record linking state, eligible actions, selected action, override, executed dose and later outcome.
Decision utility	Consequence-aware value of an action after accounting for target benefit, burden, avoidable error, risk and irreversibility.
HELIX	The longitudinal athlete-state inference layer beneath MODUS.
HUMAN REVIEW	Action class that defers to a coach, clinician or specialist when software authority is exceeded.
Interaction Governor	Mechanism deciding whether an additional question, test or prompt is worth its expected decision value.
Minimum Sufficient Evidence	Smallest evidence set required to improve a meaningful decision at the required consequence level.
MODUS	The bounded progression-decision layer that consumes HELIX state and chooses the next action.
Outcome learning	Use of later execution, response and outcome records to improve calibration or future models without automatically inferring causation.
Progression Intelligence Compression	Preserving maximum defensible progression decision value while minimising burden, cost, hardware dependence, latency and epistemic loss.

TERM	WORKING DEFINITION
Selective decision system	System that can choose not to automate certain cases so reliability can be protected where it does decide.
TEST	Action that acquires discriminating evidence because plausible outcomes would change the next decision.
Translation Fidelity	Degree to which athlete-facing wording preserves the construct, uncertainty and action meaning of the internal record.
Validation authority	Strongest product claim supported by demonstrated measurement, state, decision, implementation and outcome evidence.

REFERENCES

Selected programme sources and external scientific literature supporting this public research edition.

1. BeepSync / MODUS Sports Tech. Inferring Progression Under Uncertainty in Combat Sport: A Scientific and Computational Architecture for Longitudinal Athlete State and Bounded Training Decisions. Independent Thesis. 2026.
2. BeepSync / MODUS Sports Tech. From Training Evidence to Progression Decisions in Combat Sports. MODUS-HELIX 2 Scientific Foundation, Volume I, Version 1.1. 2026.
3. BeepSync / MODUS Sports Tech. What It Takes to Progress: Combat Performance Domains and the MODUS Benchmark Registry. MODUS-HELIX 2 Scientific Foundation, Volume II, Version 1.1. 2026.
4. BeepSync / MODUS Sports Tech. From Field Test to Sensor: Measurement Validity, External Evidence and Data Admission in Combat Sports. MODUS-HELIX 2 Scientific Foundation, Volume III. 2026.
5. BeepSync / MODUS Sports Tech. Longitudinal Athlete State: Mathematics for HELIX 2. MODUS-HELIX 2 Scientific Foundation, Volume IV, Version 1.1. 2026.
6. BeepSync / MODUS Sports Tech. Decision Validation in Practice: Coaches, Athletes and the MODUS Progression Engine. MODUS-HELIX 2 Scientific Foundation, Volume V. 2026.
7. BeepSync / MODUS Sports Tech. The Combat Athlete Progression Frontier: What a Benchmark Longitudinal Decision System Must Know, Infer, Test and Prove. Frontier Thesis, Version 1.0. 2026.
8. BeepSync / MODUS Sports Tech. Consumer Progression Intelligence: Compressing High-Performance Combat-Athlete Science into an Affordable, Low-Burden Mobile Decision System. Translational Thesis, Version 1.0. 2026.
9. BeepSync / MODUS Sports Tech. Canonical Computational, Scientific and Consumer Interaction Architecture for MODUS OS2 / HELIX 2. Official Appendix A, Version 1.0. 2026.
10. BeepSync / MODUS Sports Tech. HELIX 2: Canonical Athlete-State Inference Architecture. Official Appendix B, Version 1.0. 2026.
11. Besson C, Stucky F. From Monitoring Metrics to Decision Evidence: Validating Actionability in Applied Sport Science. *Front Sports Act Living*. 2026;8:1983979. doi:10.3389/fspor.2026.1983979.
12. Rebelo A, Bishop C, Thorpe RT, Turner AN, Gabbett TJ. Monitoring Training Effects in Athletes: A Multidimensional Framework for Decision-Making. *Sports Med*. 2026;56:1603-1624. doi:10.1007/s40279-026-02417-4.
13. Gabbett TJ, Nassis GP, Oetter E, et al. The athlete monitoring cycle: a practical guide to interpreting and applying training monitoring data. *Br J Sports Med*. 2017;51:1451-1452. doi:10.1136/bjsports-2016-097298.
14. Bourdon PC, Cardinale M, Murray A, et al. Monitoring Athlete Training Loads: Consensus Statement. *Int J Sports Physiol Perform*. 2017;12(Suppl 2):S2-161-S2-170.
15. Neupert EC, Gupta L, Holder T, Jobson SA. Athlete monitoring practices in elite sport in the United Kingdom. *J Sports Sci*. 2022;40(13):1450-1457. doi:10.1080/02640414.2022.2085435.
16. Neupert EC, Holder T, Gupta L, Jobson SA. More than metrics: the role of socio-environmental factors in determining the success of athlete monitoring. *J Sports Sci*. 2024;42(4):323-332. doi:10.1080/02640414.2024.2330178.
17. Paul D, Read P, Farooq A, et al. Factors Influencing the Association Between Coach and Athlete Rating of Exertion: a Systematic Review and Meta-analysis. *Sports Med Open*. 2021;7:1. doi:10.1186/s40798-020-00287-2.

18. Figueiredo DH, Bellenger C, Machado FA. Individually guided training prescription by heart rate variability and self-reported stress tolerance in recreational runners: Effects on endurance performance. *J Sports Sci.* 2022;40(24):2732-2740. doi:10.1080/02640414.2023.2191082.
19. Vesterinen V, Nummela A, Heikura I, et al. Individual Endurance Training Prescription with Heart Rate Variability. *Med Sci Sports Exerc.* 2016;48(7):1347-1354. doi:10.1249/MSS.0000000000000910.
20. Ranieri LE, Casado A, Martin D, et al. Performance and Physiological Effects of Race Pace-Based Versus Heart Rate Variability-Guided Training Prescription in Runners. *Med Sci Sports Exerc.* 2025;57(7):1510-1522. doi:10.1249/MSS.0000000000003671.
21. Robinson ZP, Helms ER, Trexler ET, et al. N of 1: Optimizing Methodology for the Detection of Individual Response Variation in Resistance Training. *Sports Med.* 2024;54:1979-1990. doi:10.1007/s40279-024-02050-z.
22. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *J Am Stat Assoc.* 2007;102(477):359-378. doi:10.1198/016214506000001437.
23. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. *J R Stat Soc Series B.* 2007;69:243-268. doi:10.1111/j.1467-9868.2007.00587.x.
24. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565-574. doi:10.1177/0272989X06295361.
25. Feng J, Sondhi A, Perry J, Simon N. Selective prediction-set models with coverage rate guarantees. *Biometrics.* 2023;79(2):811-825. doi:10.1111/biom.13612.
26. Kidwell KM, Almirall D. Sequential, Multiple Assignment, Randomized Trial Designs. *JAMA.* 2023;329(4):336-337. doi:10.1001/jama.2022.24324.
27. Zhang S, Dziak JJ, Tomlinson DC, Nahum-Shani I. Experimental designs with repeated randomizations for optimizing adaptive interventions: guidelines and practical considerations. *Ann Behav Med.* 2026;60(1):kaag006. doi:10.1093/abm/kaag006.
28. Soligard T, Schwellnus M, Alonso JM, et al. How much is too much? Part 1: International Olympic Committee consensus statement on load in sport and risk of injury. *Br J Sports Med.* 2016;50:1030-1041. doi:10.1136/bjsports-2016-096581.
29. Saw AE, Main LC, Gastin PB. Monitoring the athlete training response: subjective self-reported measures trump commonly used objective measures: a systematic review. *Br J Sports Med.* 2016;50:281-291. doi:10.1136/bjsports-2015-094758.
30. Whiston A, Kidwell KM, O'Reilly S, et al. The use of sequential multiple assignment randomized trials (SMARTs) in physical activity interventions: a systematic review. *BMC Med Res Methodol.* 2024;24:308. doi:10.1186/s12874-024-02439-4.
31. Wang C, Kaufman JS, Steele RJ, Shrier I. Target trial framework for determining the effect of changes in training load on injury risk using observational data: a methodological commentary. *BMJ Open Sport Exerc Med.* 2024;10(3):e002037. doi:10.1136/bmjsem-2024-002037.
32. Feuerriegel S, Frauen D, Melnychuk V, et al. Causal machine learning for predicting treatment outcomes. *Nat Med.* 2024;30:958-968. doi:10.1038/s41591-024-02902-1.
33. Hecksteden A, Kraushaar J, Scharhag-Rosenberger F, Theisen D, Senn S, Meyer T. Individual response to exercise training: a statistical perspective. *J Appl Physiol.* 2015;118:1450-1459. doi:10.1152/jappphysiol.00714.2014.
34. Hecksteden A, Pitsch W, Rosenberger F, Meyer T. Repeated testing for the assessment of individual response to exercise training. *J Appl Physiol.* 2018;124:1567-1579. doi:10.1152/jappphysiol.00896.2017.
35. Strong J, Rogers H, Sun E, et al. Human-AI Collaboration in Healthcare: A Scoping Review. *npj Digit Med.* 2026. doi:10.1038/s41746-026-02918-6.
36. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc.* 2012;19(1):121-127. doi:10.1136/amiajnl-2011-000089.
37. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc.* 2017;24(2):423-431. doi:10.1093/jamia/ocw105.

38. International Organization for Standardization. ISO 9241-210:2019 Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. Confirmed current 2025.

Publication integrity note

This public research edition is a BeepSync / MODUS Sports Tech company-authored synthesis of the MODUS-HELIX 2 internal research programme. It is not an independent systematic review, clinical guideline or peer-reviewed validation study. The bounded action vocabulary, candidate-gap logic, Interaction Governor, Decision Journal, Progression Intelligence Compression framework, authority ladder and minimum-defensible-decision-regret framing are proposed programme constructs. They should not be represented as empirically validated product efficacy until prospective studies support the relevant claims.

The external literature cited here supports the underlying principles of measurement validity, longitudinal monitoring, calibration, selective prediction, individual-response methodology, adaptive-intervention design, human factors and decision validation. Transfer to combat-sport progression and to the specific MODUS policy requires product- and domain-specific testing. Medical diagnosis, injury prediction, concussion clearance and unvalidated technical/tactical judgement remain outside the intended authority of this system.