Athleet.Ai

PERFORMANCE INTELLIGENCE
WHITEPAPER SERIES

PAPER ONE OF SEVEN · GOVERNANCE AND ASSURANCE

The Responsible Performance System

Governance, explainability and trust in artificial intelligence for sport

Paul Forrest
Founder and Chief AI Officer, Athleet.AI
April 2025

About this series

This is the first of seven papers examining how artificial intelligence can be used responsibly in performance sport. Each paper stands alone. Together they set out a single argument, which is that the durable advantage in sport comes from better decisions rather than from more data, and that better decisions require systems designed around the person who is accountable for them.

This paper supplies the framework the other six draw on. Where a sport-specific paper says that a control belongs to the buyer rather than the supplier, this is the paper that says why.

Editorial balance. Roughly seven parts in ten of this document is evidence and sector analysis, two parts in ten is original framework, and one part in ten concerns the Athleet.AI platform itself. Readers who want the argument without the product can stop at Chapter 13.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim in this paper carries a tag stating what kind of evidence supports it. The tags are used consistently across the series.

PEER-REVIEWED A primary study in a peer-reviewed journal.

REVIEW A systematic review or meta-analysis.

CONSENSUS A consensus statement, governing-body position, regulator guidance or primary legislation.

COMMENTARY Methodological criticism, which carries weight without being primary data.

PRACTICE Practitioner opinion, including the author's own.

ATHLEET.AI The author's own framework or analysis, proposed rather than proven.

Where the evidence is thin, the paper says so. Where it points the other way, the paper says that too.

Legal and regulatory notice. This paper describes legislation, regulatory guidance and governing-body requirements throughout, and in particular in Chapters 4, 9 and 11. Nothing in this document constitutes legal advice. The provision of legal advice sits outside the terms of any engagement with the author or with Athleet.AI, and the material is presented to support discussion and further review by qualified advisers.

Statements about what a source says are attributed to that source. Where the author draws a practical implication the source does not itself state, the text says so. Legislative timetables in this area move, and the position described is the position the author was able to verify in April 2025.

ILLUSTRATIVE MATERIAL

Every athlete, club, programme and scenario in this paper is invented, and none describes a real athlete, coach or organisation. Cases are marked **illustrative scenario** wherever they appear. Published research is cited and numbered, and invented material is never presented as evidence.

CITATION

Forrest, P. (2025). *The Responsible Performance System: Governance, Explainability and Trust in Artificial Intelligence for Sport.* Athleet.AI Performance Intelligence Whitepaper Series, Paper One. Version 1.0, April 2025.

Comments and corrections are welcome and will be acknowledged in later versions.

ALSO IN THIS SERIES

- **Paper two.** *The Intelligent Track.* AI and event-specific expertise in elite athletics.
- **Paper three.** *Closing the Coaching Gap.* Responsible AI in grassroots athletics.
- **Paper four.** *Casting the Net Wider.* AI as a force multiplier in football pathways.
- **Paper five.** *The Connected Endurance Athlete.* Integrated preparation in triathlon.
- **Paper six.** *The Augmented Touchline.* Responsible AI in elite football.
- **Paper seven.** *The Augmented Golfer.* AI across technique, practice and strategy.

Contents

A NOTE FROM THE AUTHOR

The question nobody asks the supplier

I have sat on both sides of this table. I have built artificial intelligence products and sold them, and I have coached athletes who were on the receiving end of somebody else's. The two experiences agree on one thing. The question that decides whether a performance system helps or harms is almost never asked in the room where it is bought.

That question is not what the software can do. It is who will be answerable for what it says, on the Tuesday evening when it is confidently wrong and an athlete has already read it.

Sport has spent a decade being sold prediction. The published evidence for individual prediction is weak, and has been weak for as long as the products have been on sale. That is not a scandal. Sport is difficult, athletes are heterogeneous and the outcomes are rare, so the models were always going to struggle. What is a scandal is that the weakness has been so rarely disclosed to the people signing the contracts, and that the disclosure obligation has been so widely assumed to sit with somebody else.

This paper is my attempt to put the obligation back where it belongs. It is written for the person who has to answer for the decision, not for the person who built the model. It sets out what to insist on before deployment, what to record during it, and what to measure afterwards. Most of it costs nothing but attention.

I should say plainly that I have a commercial interest. I run a company that builds software in this area. I have therefore tried to write the paper that would make my own product harder to sell badly. Several of the recommendations here would rule out categories of product that are commercially straightforward to build. That is deliberate. A market in which nobody can tell a governed system from an ungoverned one is a market in which the governed one loses, and I would rather compete on ground where the difference is visible.

Test the citations rather than the author. Every one of them is listed, and every one has been checked twice.

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

Sport is buying artificial intelligence faster than it is governing it. The gap is not primarily an ethical problem, though it is that as well. It is a performance problem. Coaches do not act on advice they cannot interrogate, athletes do not trust systems that will not explain themselves, and a club that cannot say who approved a decision cannot learn from it. Ungoverned AI is not dangerous mainly because it is unsafe. It is wasteful because it is unused.

This paper sets out a governance and operating model that a federation, club or institute can adopt without creating a new department. It rests on classifying decisions rather than technologies, on stating what kind of evidence stands behind every number, and on putting a name and a date against every decision that affects a person.

1

Individual prediction does not currently work well enough to govern a career. Hamstring injury models tested across seasons returned a median area under the curve of 0.52, where 0.5 is a coin. The acute to chronic workload ratio adds nothing to an intercept-only model, at 0.574 against 0.5, within its own training sample.

2

A counted fact beats the models. Track and field athletes who completed more than 80 per cent of planned training weeks were seven times more likely to hit a performance goal, at an area under the curve of 0.72. That is a threshold anybody can audit, and it outperforms the black boxes on published numbers.

3

The law already gives athletes more than most clubs deliver. Access, portability, rectification, erasure, human intervention where automated decisions apply and a prior impact assessment are all existing rights. What is missing is an explanation on request and a record that travels with the athlete, and neither requires new law.

4

A human signature is not oversight. Automation bias is well documented in clinical decision support, and its mediators are workload, task complexity and time pressure, which describe a training ground exactly. The European legislature made the same judgement, obliging deployers to guard against over-reliance by name.

5

Sport is barely in scope, which is the problem. The word athlete appears nowhere in the European Union's Artificial Intelligence Act, and sport appears only in its recitals. No governing body had published a binding instrument on performance AI. Sport will be governed by whatever it writes for itself.

6

There is a hard line at health. A dashboard that reports what an athlete did sits outside the medical device regime. A tool that applies automated reasoning to output an individual's future risk of injury has described itself into the regulatory definition, which names prediction, prognosis and injury expressly.

What follows for a board

Adopt five controls and one habit. Classify every decision by consequence and reversibility before classifying the software. State the evidence grade behind every number a practitioner sees. Require that recommendations show provenance, uncertainty, assumptions and alternatives. Name who recommends, who approves, who acts, who reviews and who answers, and never let those collapse into one person. Ask the 12 procurement questions

in Chapter 11 before signing. The habit is writing decisions down, with their reasons, on the day they are made.

None of that is expensive. The register fits on one page, the three meetings already exist under other names, and the discipline of recording a reason costs a minute. What it buys is the ability to learn from a season rather than remember it, and the ability to answer for a decision after the people who made it have moved on.

Chapter 13 sets out the operating model and a 100-day sequence for adopting it. The call to action at the end of this paper is a readiness assessment, which is the cheapest useful thing an organisation can do before it buys anything else.

PART ONE

What is actually being governed

Three chapters on the difference between governing a technology and governing a decision, and on why the second is the only one that works.

CHAPTER ONE

Why responsible AI is a performance capability

Trustworthiness is usually filed under compliance. It belongs under performance, because a system nobody trusts is a system nobody uses.

Start with the failure that actually happens. A club buys a performance platform. It arrives with a readiness score, a risk flag and a dashboard. Within a season the medical team has stopped looking at the flag because it was wrong about somebody they knew was fine, the coaches have stopped looking at the score because it disagreed with what they saw on the grass, and the analyst is the only person who opens the software, chiefly to produce a report for a board that does not read it either. Nobody has been harmed. A six-figure sum has been spent to change nothing.

That is the modal outcome, and it is a governance failure rather than a technical one. The model may well have been competent. What was missing was any mechanism by which a practitioner could find out why it said what it said, disagree with it in a way the system recorded, and see whether their disagreement was vindicated. Without that mechanism, the first wrong answer is fatal to adoption, because there is no way to distinguish a system that is wrong occasionally from one that is wrong systematically.

1.1 What trust is actually made of

Trust in this setting is not a feeling. It is composed of four things a buyer can specify and test. The first is provenance, meaning that every input can be traced to a source and a date. The second is calibration, meaning that when the system says it is unsure it is in fact unsure. The third is contestability, meaning that a practitioner can disagree and have the disagreement recorded rather than merely ignored. The fourth is memory, meaning that the record of what was recommended, decided and observed survives the person who made it.

All four are engineering requirements as much as ethical ones, and all four can be written into a contract. None of them requires the model to be better. Most of them make the model's weaknesses more visible, which is precisely why they are resisted.

The argument of this paper is that these four properties are what convert a competent model into a used one, and that the conversion is where the return on the investment actually sits. A club with a mediocre model and excellent governance will outperform a club with an excellent model and none, because only the first club is capable of learning.

THE ADOPTION TEST

If a practitioner cannot register a disagreement in a form the system keeps, the system will be abandoned the first time it is wrong in front of somebody who knows better.

THE ARGUMENT IN ONE PARAGRAPH

Responsible artificial intelligence in sport is not a tax on performance work. It is the mechanism by which performance work compounds. Provenance lets a disputed number be checked. Calibration lets a coach know when to look harder. Contestability lets expertise enter the system rather than bounce off it. Memory lets a programme improve across a change of staff. Take those four away and what remains is an expensive opinion generator that nobody is obliged to believe.

1.2 Why the sector is unusually exposed

Three features of sport make the governance gap larger here than in most sectors. Decisions are made fast, by small teams, under public scrutiny, which compresses the time available to interrogate anything. The population is small and heterogeneous, which means models are trained on few examples of people who differ from one another in the ways that matter. And a large proportion of the subjects are young, which engages a body of law that most performance departments have never read.

To those, add a fourth that is specific to this moment. Sport is not named in the instrument most likely to regulate artificial intelligence in Europe. The word athlete does not appear anywhere in Regulation (EU) 2024/1689, and sport appears only in its recitals, never in an operative article and never in the high-risk annex.[33] CONSENSUS A sports performance system falls within the high-risk regime only derivatively, most plausibly where the athlete is an employee and the system monitors or evaluates their performance. My own reading is that this leaves most of the sector outside the regime, and a federation should take its own advice rather than rely on mine. The practical consequence is the same either way. Sport will be governed by what it writes for itself, or it will not be governed at all.

CHAPTER TWO

What sports AI actually is

Five different technologies travel under one word. They fail in different ways, and they need different controls.

Almost every governance failure I have seen begins with a category error. A board approves a policy on artificial intelligence as though it were one thing. Meanwhile the club is running five, each with its own failure mode, and the policy addresses none of them because it was written about a word rather than about a system.

FIVE THINGS THE WORD COVERS

FAMILY	WHAT IT ACTUALLY DOES	WHERE IT EARNS ITS PLACE	WHAT IT SHOULD NOT BE ASKED TO DO
Prediction	Estimates a future value or the probability of an event from historical data.	Forecasting availability across a squad; projecting a season's competition load.	Individual injury forecasting. The published discrimination is close to chance.
Optimisation	Searches a space of options for the one that best satisfies a stated objective.	Scheduling, travel, fixture and facility planning, taper structures within stated constraints.	Any objective the club has not written down. It optimises what you specify, not what you meant.
Computer vision	Extracts positions, angles and events from video or optical tracking.	Counting, timing, tracking and flagging events a human would otherwise log by hand.	Transverse-plane joint angles and anything requiring millimetre precision in fast rotation.
Language models	Generates and summarises text, and answers questions over a document set.	Drafting, summarising, translating, retrieving from an approved knowledge base.	Anything asserted as fact without a retrievable source behind it.
Automation	Executes a defined sequence without a person present at each step.	Moving data between systems, producing recurring reports, chasing missing records.	Any step where the sequence itself constitutes a decision about a person.

The governance question is never whether a club uses artificial intelligence. It is which of these five is running, on whose data, and who signs the output.

FIGURE 1 The five families the word covers, with the work each one earns and the work it should be refused.

Athleet.AI framework. The failure modes in the fourth column are supported by the evidence in Chapters 3, 5 and 7. ATHLEET.AI

2.1 Prediction is the one that gets sold

Prediction estimates a future value from historical data. It is the family most heavily marketed into sport and the family with the weakest published record, which is a combination worth stating early. Chapter 5 gives the numbers. For now the point is structural. Prediction fails quietly. An optimisation error produces an obviously daft schedule. A vision error produces a joint angle a coach can see is wrong. A prediction error produces a number that looks exactly like a correct number, and it can only be evaluated by counting outcomes over a period longer than most staff tenures.

That asymmetry is why prediction deserves the strictest control of the five, and why the honest applications sit at squad level rather than individual level. Forecasting how many athletes are likely to be unavailable across a block is a planning question with a tolerable error. Forecasting which athlete will be injured on which day is a different question, and the published evidence does not support it.

2.2 The others fail more usefully

Optimisation searches for the option that best satisfies a stated objective, which makes it excellent at scheduling and travel and dangerous wherever the objective has not been written down. It will optimise what you specify rather than what you meant, and the gap between those two is where the trouble lives. Computer vision extracts positions and events from video, and the series' second paper sets out where its accuracy holds and where it does not.[17] COMMENTARY Language models generate and summarise text, which makes them genuinely useful for drafting and retrieval and unsuitable for asserting facts without a retrievable source behind them.

Automation is the family most often left out of AI policy altogether, on the grounds that a scheduled data transfer is not intelligent. It belongs in the policy, because the question is never how clever the step is. It is whether the step constitutes a decision about a person. A rule that automatically moves an athlete from

one training group to another is a decision, however simple the logic that produced it.

2.3 What follows for a policy document

An artificial intelligence policy that does not name which of the five families are in service, on which data, is not a policy. It is a statement of intent. The register described in Chapter 13 asks for four things per system, which are the family, the data it reads, the decisions it touches and the person accountable for it. Most organisations can complete that register in an afternoon, and most are surprised by what appears on it.

Zero

MENTIONS OF ATHLETES

The word athlete appears nowhere in the European Union's Artificial Intelligence Act. Sport appears only in its recitals, and never in the high-risk annex.

ILLUSTRATIVE SCENARIO

The register nobody expected

An invented national programme sets out to register its artificial intelligence and expects to list two systems, being the athlete management platform and a video analysis tool. The completed register runs to nine. It includes a scheduling optimiser inside the travel booking system, a language model embedded in the email client that drafts replies to athletes, a wearable whose sleep staging is a proprietary model, an automated rule that flags athletes for a welfare conversation, and a spreadsheet macro written by a departed analyst that nobody can explain but everybody uses. Only two of the nine had been through any approval. The macro turned out to be the one that touched selection.

CHAPTER THREE

The decision-risk spectrum

Govern the decision, not the technology. The same model may be safely automated in one setting and unacceptable in another.

The most useful move available to a board is also the cheapest. Stop asking whether a technology is risky and start asking what the decision does to the person on the receiving end, and how easily it can be undone. Those two questions produce a matrix that a performance director can complete in a morning and that survives every change of supplier, because it describes the club's decisions rather than anybody's software.

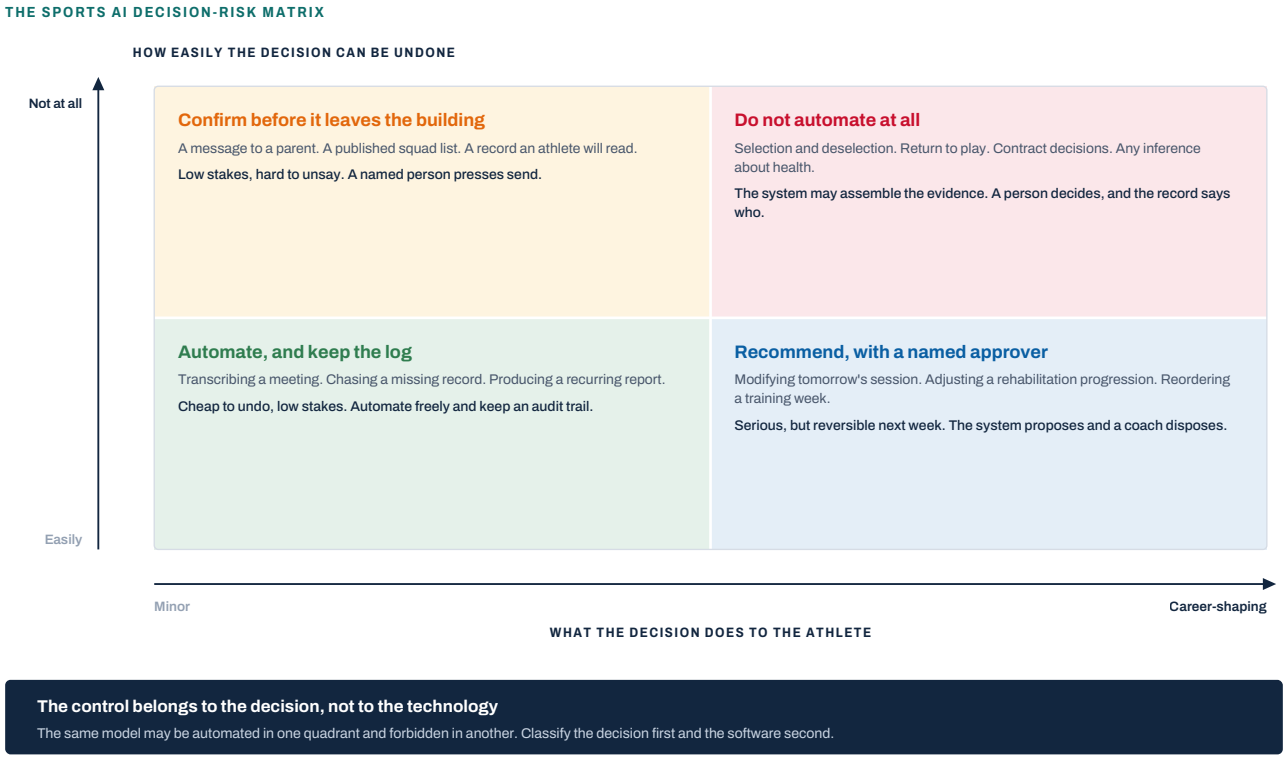


FIGURE 2 Consequence against reversibility, with the control each quadrant requires. The examples are illustrative and every organisation should place its own.

Athleet.AI framework. ATHLEET.AI

3.1 The four quadrants

Low consequence and easily undone is the automation quadrant. Transcribing a meeting, chasing a missing record and producing a recurring report belong here. Automate freely, keep a log, and stop treating this work as though it needed a policy debate. Most of the genuine time saving available to a performance department sits in this quadrant and is neglected because attention has gone to the glamorous end.

Low consequence but hard to undo is the confirmation quadrant. A message to a parent, a published squad list, a record an athlete will read. The stakes are modest and the words cannot be unsaid, so a named person presses send. This is the quadrant most often mishandled, because the low stakes make automation tempting and the irreversibility makes it costly.

High consequence but reversible within a week is where most performance work actually sits. Modifying tomorrow's session, adjusting a rehabilitation progression, reordering a training week. The system proposes and a coach disposes, with the approval recorded. This is the quadrant that justifies the whole investment, and it is the one that fails when the recommendation arrives without provenance.

High consequence and hard to undo is the refusal quadrant. Selection and deselection, return to play, contract decisions, and any inference about health. A system may assemble the evidence for these decisions and present it well. It may not make them, and the record should say who did.

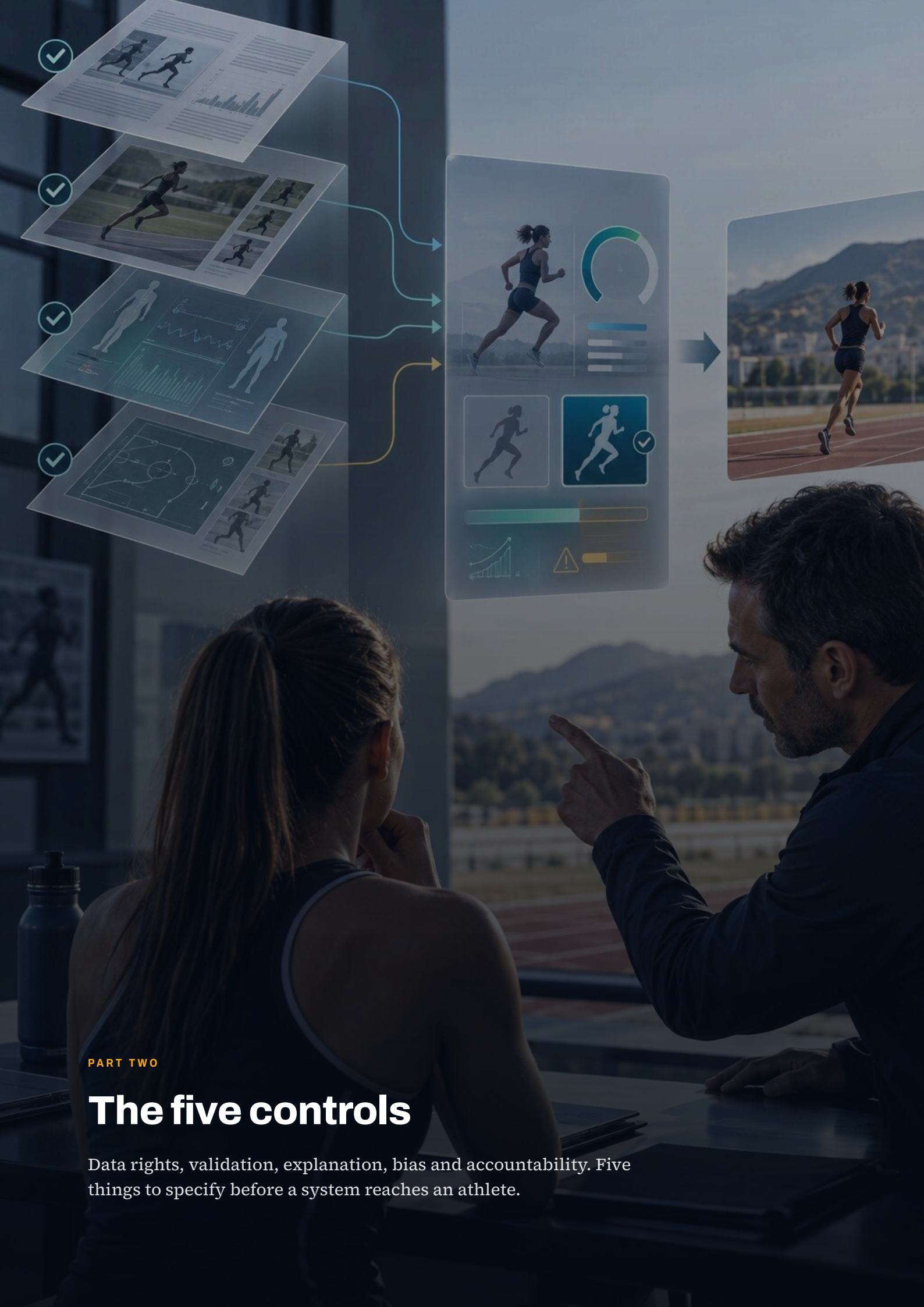
WHAT THIS RULES OUT

No automated selection or deselection of any athlete. No automated return-to-play clearance. No system that changes what an athlete is asked to do without a named person approving the change. No score attached to a person under 18. These are not aspirations. They are the four lines this paper asks a board to draw before it buys anything, and each of them is testable against a supplier's own documentation.

3.2 Why the matrix beats a technology policy

A technology policy ages badly. It is written about the products available when it was drafted, and it is obsolete within two procurement cycles. A decision matrix ages well, because the decisions a club makes about athletes have been broadly stable for a century and will outlast every vendor in the market. When a new product arrives, the question is not whether it is an AI product. It is which cells of the matrix it touches, and whether the control for those cells is in place.

My own view, offered as practice rather than evidence, is that the matrix should be completed by the people who make the decisions rather than by the people who buy the software. The exercise surfaces disagreements about authority that were previously invisible, and those disagreements are more valuable than the finished grid.



PART TWO

The five controls

Data rights, validation, explanation, bias and accountability. Five things to specify before a system reaches an athlete.

CHAPTER FOUR

Data rights and athlete agency

Most of what athletes ask for, the law already grants. The gap is between the right and its delivery.

Conversations about athlete data usually begin in the wrong place, which is whether athletes should have rights over their performance data. In the United Kingdom and the European Union they already do, and have since 2018. The interesting question is why so little of it reaches them, and what a club would have to change to close the distance.

THE ATHLETE DATA RIGHTS MAP

What is collected	Who actually holds it	What the law already gives	What is still missing
Performance Distance, speed, power, output against plan	**The club or programme** As controller, usually without saying so	**Access and portability** Articles 15 and 20	**An explanation on request** No general right to one in law
Physiological Heart rate, sleep, wellness questionnaires	**The supplier** As processor, and sometimes beyond it	**Rectification and erasure** Articles 16 and 17	**A record that travels** The history stays with the club
Clinical Injury, illness, treatment, screening	**The device maker** Under its own consumer terms	**Human intervention** Article 22(3), where it applies	**A say in secondary use** Training a supplier's model
Judgement Coach notes, scout reports, selection reasoning	**The athlete** Rarely, and rarely in a usable form	**A prior impact assessment** Article 35, before deployment	**Anything after the career** Retention is rarely stated

The fourth column is where trust is won or lost, and none of it requires new law. It requires a decision, written down.

FIGURE 3 What is collected, who holds it, what the law already provides and what is still missing. The fourth column is the one that requires a decision rather than a statute.

Rights in the third column are set out in references 34 to 39. The fourth column is the author's own analysis. ATHLEET.AI

4.1 What the law gives

Under the UK General Data Protection Regulation an athlete has a right of access to their personal data and a right to receive it in a portable form, a right to have inaccurate data corrected and, in defined circumstances, erased.[34] CONSENSUS Where a decision is based solely on automated processing and produces a legal effect or similarly significantly affects them, they have a right to obtain human intervention, to express a point of view and to contest the decision.[36] CONSENSUS Before high-risk processing begins, the controller must complete a data protection impact assessment, and a systematic and extensive automated evaluation of personal aspects on which significant decisions rest is named in the regulation as exactly such a case.[39] CONSENSUS

Two features of that list deserve attention. The first is that the impact assessment is required before the processing, not after the pilot. A club that runs the assessment once the system is embedded has answered a different question from the one the regulation asks. The second is that Article 22 bites only on decisions taken solely by automated means, and almost no sporting decision is. That is why the article is invoked so rarely and why relying on it as the principal safeguard is a mistake. The protection that matters in practice comes from Article 35 and from the accountability principle in Article 5, not from Article 22.[35] CONSENSUS

4.2 What the sector has added, and what it has not

Professional football has gone furthest. A charter of player data rights developed by the players' union with the world governing body sets out eight rights over performance and personal data, covering information, access, revocation, restriction of processing, portability, rectification, complaint and erasure.[51] CONSENSUS It is a serious document and it is worth reading. It is also, on my reading, a restatement of rights the General Data Protection Regulation already provides. It contains nothing on explainability, nothing on model validation and nothing on artificial intelligence. The wider players' declaration adopted in 2017 is similar in character, asserting a right to privacy and protection in the collection, storage and transfer of personal data, and a right to have name, image and performance protected and commercially used only with consent.[52] CONSENSUS

Meanwhile a group action announced in 2020 asserted that third parties process professional footballers' performance and tracking data without a valid basis, and sought recovery of income going back six years. More than 400 players were reported to have joined at the time.[53] COMMENTARY I have found no reported judgment and no verifiable evidence that proceedings were determined, so it should be described as an announced claim rather than a decided case. Its significance is as a signal of appetite rather than as authority.

4.3 The four things a club can give that the law does not require

The interesting work sits in the fourth column of Figure 3, and none of it needs legislation. An explanation on request, meaning that an athlete can ask why a recommendation about them was made and receive an answer in the terms of Chapter 6. A record that travels, meaning that an athlete leaving a programme takes a usable version of their own history with them. A stated position on secondary use, meaning that the club has decided, and written down, whether a supplier may train its models on athlete data. And a retention position, meaning that somebody has decided what happens to a longitudinal record when a career ends.

My own view is that the last of these is the one to settle first, because it is cheap while the record is small and contentious once it is valuable. An athlete who cannot take their own history with them will reasonably conclude that the system was never built for them, and they will be right.

BEFORE, NOT AFTER

The impact assessment is required prior to the processing. An assessment completed after a pilot has already answered the question it was supposed to ask.

SPECIFICATION

Four positions to write down before a system goes live

- Who is the controller and who is the processor, named in the contract rather than assumed.
- Whether athlete data may be used to train models the club does not own, stated as a yes or a no.
- What an athlete receives on request, in what format, and within what period.
- What happens to the record when the athlete leaves and when the contract with the supplier ends.

CHAPTER FIVE

Evidence and validation

This is the chapter where a paper of this kind is expected to become enthusiastic about machine learning. The published evidence declines to cooperate.

The claim that a model can predict which athlete will be injured, and when, has been examined repeatedly and has not survived the examination. Stating that plainly is more useful to a performance director than any amount of qualified optimism, and it is the single most important thing this paper has to say about evidence.

FIGURE 4 What the published models achieve, on the discrimination measure their own authors report. A value of 0.5 is a coin toss. Bars show reported ranges where the study gives them.

References 1, 2, 3 and 58. Figures as reported by each study; the comparison across studies is the author's own and the designs are not equivalent.
PEER-REVIEWED

5.1 The numbers

Researchers built hamstring injury models on two seasons of data from a professional Australian football competition, covering 186 players in one year and 176 in another. Within a single year the models ranged widely, with median areas under the curve of 0.58 and 0.57. Tested across years, which is the only test that matters for a tool a club would actually deploy, the median fell to 0.52 with a range from 0.37 to 0.73.[1] PEER-REVIEWED The authors concluded that risk factor data cannot be used to identify athletes at increased risk with any consistency.

A second group modelled training load and injury in the same sport across three seasons, comparing regularised logistic regression, generalised estimating equations, random forests and support vector machines. For non-contact injuries the predictive performance was described as only marginally better than chance, below 0.65. The best single result, 0.76, was for hamstring-specific injury and

came from the simplest model in the set.[2] PEER-REVIEWED That detail is easy to skip and worth pausing on. The added complexity did not pay.

The acute to chronic workload ratio, which has done more than any other single idea to shape load management practice, was dismantled by re-analysis. Substituting contrived chronic loads for real ones reproduced the reported effect. Real data gave an odds ratio of 2.45, dividing acute load by a fixed value gave 1.95, and randomly generated chronic loads gave ratios averaging 1.53. Neither the ratio nor acute load alone conferred a meaningful predictive advantage over an intercept-only model, at a c-statistic of 0.574 against 0.5, within the training sample.[3] PEER-REVIEWED An earlier letter had already shown the mechanism, which is that the acute load forms a substantial part of the chronic load, so the two components of the ratio are not independent and the resulting association is mathematically coupled.[4] COMMENTARY

5.2 Why this was predictable

0.52

BETWEEN SEASONS

The median discrimination of hamstring injury models tested on a year they had not seen. A coin scores 0.5.

None of this should have surprised anyone, and a critical review had said so in advance. Validating a screening test requires three things, being a demonstrated prospective relationship between the marker and injury risk, examination of the test's properties in a relevant population using appropriate statistical tools, and evidence that an intervention targeted at the identified high-risk group outperforms the same intervention given to everybody. The review found no example of a screening test for sports injuries with adequate test properties, and no intervention study supporting screening for injury risk.[5] COMMENTARY The reason is structural. Test performance is measured on a continuous scale, and substantial overlap between players who go on to be injured and those who do not is exactly what should be expected.

The consensus statements of the period were appropriately hedged. The International Olympic Committee's 2016 statements on load and injury and on load and illness described the evidence as emerging, offered practical guidelines and expressly identified research priorities.[7,8] CONSENSUS A monitoring consensus the following year observed that the field had developed so quickly that it had given rise to industries aimed at developing new paradigms.[9] CONSENSUS The consensus statements were careful. The industry that grew around them was not.

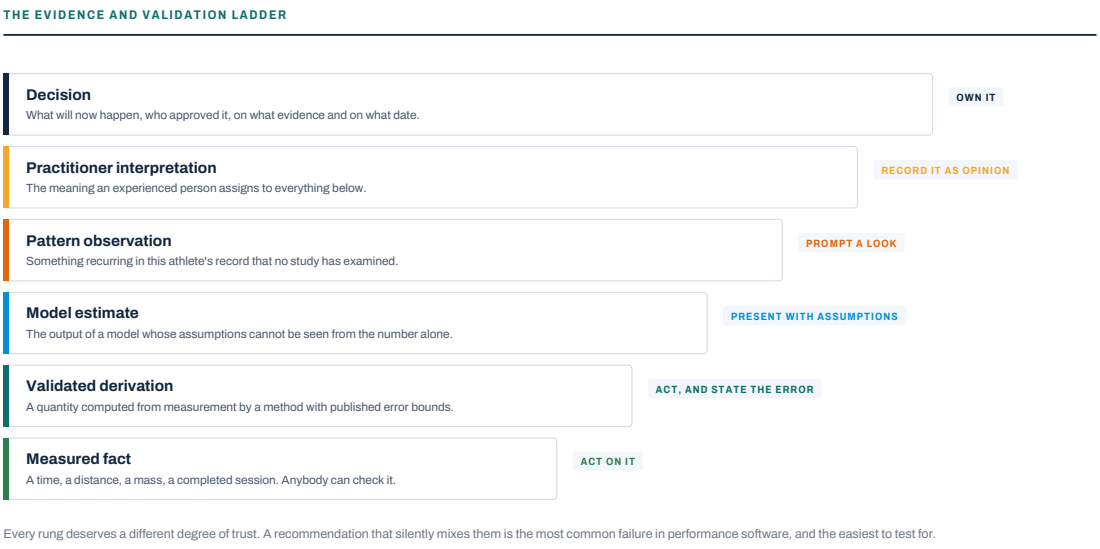


FIGURE 5 Six rungs, each deserving a different degree of trust. The recommendation that mixes them without saying so is the most common failure in performance software.

Athleet.AI framework, informed by references 11, 12 and 13. ATHLEET.AI

5.3 What good validation looks like

Clinical prediction has spent 20 years learning lessons that sport is now repeating, and the reporting standards it produced transfer without modification. A prediction model should be reported against an established checklist so that a reader can tell how it was built and on whom.[11] CONSENSUS Discrimination, the ability to rank one athlete above another, is not sufficient. Calibration, the correspondence between predicted risk and observed frequency, is what determines whether a number can be acted on, and a hierarchy exists for describing it. Strong calibration is unrealistic and encourages overly complex models; moderate calibration, meaning that among athletes given a predicted risk of a given value the observed event rate matches it, is achievable and has been shown to guarantee decisions that do no harm.[12,13] PEER-REVIEWED

The systematic review of covid-19 prediction models is the cautionary example every sports buyer should read, because the field it examined had every incentive to get this right. Of 66 models described in 51 studies, all were rated at high or unclear risk of bias, calibration was rarely assessed, and the reviewers recommended none of them for practice.[10] REVIEW That was medicine, under scrutiny, with resources. Sport should not assume it will do better unaided.

The commercial dimension makes it worse. Developers of proprietary prediction models are often unwilling to report or release how those models were developed and validated, and that reticence prevents any independent evaluation of performance, interpretability or generalisability before a system is used on athletes.[6] COMMENTARY A club cannot appraise what it cannot see, which is why Chapter 11 makes the validation study a condition of purchase rather than a courtesy.

SPECIFICATION

Three requirements for any predictive claim entering a performance department

- **Performance on unseen data, reported.** Not in-sample performance, and not performance on the population the model was fitted to. The between-year figure is the one that describes deployment.
- **Calibration, not only discrimination.** A model that ranks well but is systematically overconfident will produce a stream of interventions nobody needed.
- **Subgroup performance.** Stated separately for the athletes the club actually has, by sex and by level, or stated as unknown.

CHAPTER SIX

Explainability a coach can use

Most explainability in this market is decorative. A usable explanation names its inputs, states what it assumed and offers something to disagree with.

The word explainable has been attached to a great deal of software that explains nothing. The most common form is a chart of feature importances, which tells a coach that sleep contributed 23 per cent to a score without telling them which night, which measurement, or what the system would have said had the number been different. That is not an explanation. It is a decoration that satisfies a procurement checkbox.

6.1 Why post-hoc explanation is weaker than it looks

There is a serious technical literature behind this scepticism and a buyer should know it exists. The central argument is that building a black box and then producing a post-hoc explanation for it is the wrong approach for high-stakes decisions, because a post-hoc explanation is by construction not faithful to the model. If it were perfectly faithful it would be the model. For many structured-data problems there is little accuracy cost to using an inherently interpretable model in the first place.[17] COMMENTARY

Three empirical findings support the caution. Interpretability itself is underspecified, with the literature offering diverse and sometimes non-overlapping notions of what makes a model interpretable.[18] COMMENTARY Interpretations are fragile, in that two perceptually indistinguishable inputs with the same predicted label can be assigned very different explanations, and small perturbations can change feature importance dramatically without changing the prediction.[19] PEER-REVIEWED And some widely used saliency methods have been shown to be independent of both the model and the data-generating process, which makes them inadequate for exactly the tasks they are marketed for.[20] PEER-REVIEWED

Taken together these do not say that explanation is impossible. They say that an explanation generated after the fact by a separate mechanism is not evidence about the model, and should not be accepted as governance. The practical response is to require explanation by construction, which in sport usually means preferring a simpler model whose inputs a practitioner can see.

6.2 What a regulator asks for

Guidance produced jointly by the Information Commissioner's Office and the Alan Turing Institute sets out six types of explanation an organisation should be able to give about an AI-assisted decision, covering the rationale, who is responsible, what data was used, fairness, safety and performance, and impact.[21] CONSENSUS It remains the most practical checklist available to anyone specifying a system in this area. It is also, and this is worth knowing before it is cited in a board paper, not statutory guidance. The Data Protection Act 2018 requires the Commissioner to prepare four statutory codes, and this is not one of them.[42] CONSENSUS The United Kingdom's flagship explainability guidance therefore carries less legal force than the Children's Code.

A TEST FOR SUPPLIERS

Ask what the system says when it does not know. A product with no answer to that question has not been designed to be disagreed with.

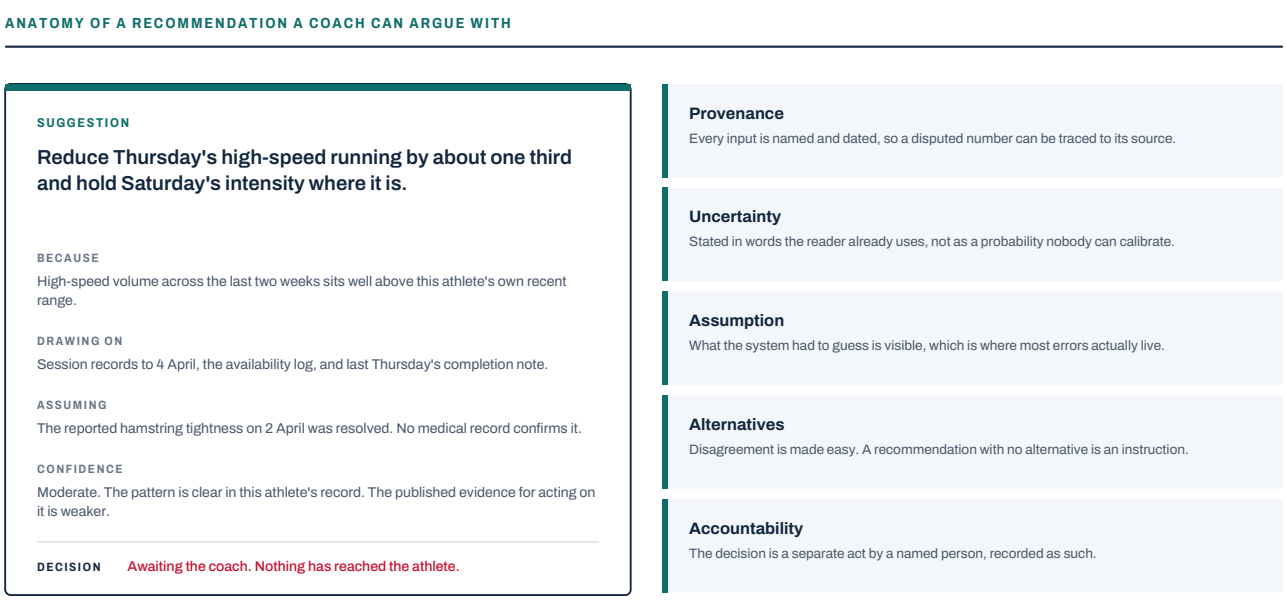


FIGURE 6 The template this series proposes. Every element exists so that a practitioner can disagree with the recommendation in a specific place rather than in general.

Athleet.AI framework, informed by reference 21. The scenario is invented. ATHLEET.AI

6.3 The template

A recommendation that reaches a practitioner should carry five things. It should name every input and its date, so that a disputed number can be traced. It should state its confidence in words the reader already uses rather than as a probability nobody can calibrate. It should say what it assumed, because assumptions are where most errors live and they are invisible in the output. It should offer at least one alternative, because a recommendation with no alternative is an instruction. And it should record the decision as a separate act by a named person, on a stated date.

The fifth of these is the one suppliers resist, because it makes the software's role smaller and the club's role larger. That is the correct division. My own view, offered as practice, is that a club should ask to see a single real recommendation from any product before buying it, and should judge the product on whether a coach could argue with what appears on the screen.

CHAPTER SEVEN

Bias, inclusion and uneven evidence

The most consequential bias in sports AI is not in the algorithm. It is in the literature the algorithm was trained to reproduce.

Discussions of algorithmic bias in sport tend to arrive by way of facial recognition and stay there. The canonical audit found that commercial gender classification systems misclassified darker-skinned women at error rates up to 34.7 per cent while the maximum error for lighter-skinned men was 0.8 per cent, and that the benchmark datasets were 79.6 and 86.2 per cent lighter-skinned subjects.[24]

PEER-REVIEWED That is a real and instructive finding, and it is not the mechanism most likely to harm an athlete.

7.1 The proxy label is the real risk

The mechanism to worry about is the choice of what to predict. A widely used healthcare algorithm affecting millions of patients was found to be racially biased because it predicted healthcare costs rather than illness, and unequal access to care meant less money was spent on Black patients. At a given risk score, Black patients were considerably sicker. Correcting the label would have increased the proportion of Black patients receiving additional help from 17.7 to 46.5 per cent.[23] PEER-REVIEWED The authors' generalisable conclusion is the sentence sport should paste above its desks, which is that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Sport is full of convenient proxies. A model of talent trained on who was previously selected learns the selection history rather than the talent. A model of readiness trained on who was withdrawn learns the caution of the medical staff. A model of return to play trained on who was cleared learns the risk appetite of the club that cleared them. In each case the model will be accurate against its label and wrong about the world, and it will be most wrong about the athletes who differ most from those in the training data.

7.2 The evidence base is unevenly built

Underneath the models sits a literature with a documented sex imbalance. An analysis of 1,382 articles across three leading journals covering 6,076,580 participants found 39 per cent were female, with the average per article ranging from 35 to 37 per cent.[25] PEER-REVIEWED A later analysis of 5,261 publications from 2014 to 2020, covering 12,511,386 participants, found 34 per cent female, with 31 per cent of publications studying males only against 6 per cent studying females only.[26] PEER-REVIEWED Read together, those two describe a gap that did not close.

The methodological consequence has been set out by a specialist working group, which observed that the majority of high-quality sport and exercise science data derive from studies with men as participants, reducing their application given known physiological differences, and identified a shortage of specialist knowledge on female physiology alongside a reluctance to adapt experimental designs.[27] CONSENSUS A model trained on that literature will be confident about a population it has barely seen, and it will not know that it is.

7.3 What a club can actually do

Three things, none of which requires new research. Require subgroup performance to be reported or declared unknown, which converts an invisible weakness into a stated one. Require the system to widen its stated uncertainty for athletes unlike its training population rather than answering with the same confidence for everybody. And record the composition of the population the club is itself accumulating, because a club that will not measure who its own data describes cannot claim to be widening anything.

Documentation practices from outside sport make this routine rather than heroic. Model cards accompany a trained model with benchmarked evaluation across demographic and intersectional groups and disclose the intended context of use.[14] CONSENSUS Datasheets accompany a dataset with its motivation, composition, collection process and recommended uses.[15] CONSENSUS Neither is difficult. Both

34%

FEMALE PARTICIPANTS

Across 5,261 sport and exercise science publications covering 12.5 million participants. The earlier figure, to 2014, was 39 per cent.

are rare in sports technology, and asking for them is a cheap way to find out how seriously a supplier takes the question.

CHAPTER EIGHT

Human accountability

A human in the loop is a design pattern, not a safeguard. What makes it a safeguard is the separation of five roles that usually collapse into one.

Every organisation I have worked with believes it has human oversight, and most of them have a signature. The two are not the same thing, and the difference is the single most useful distinction in this paper.

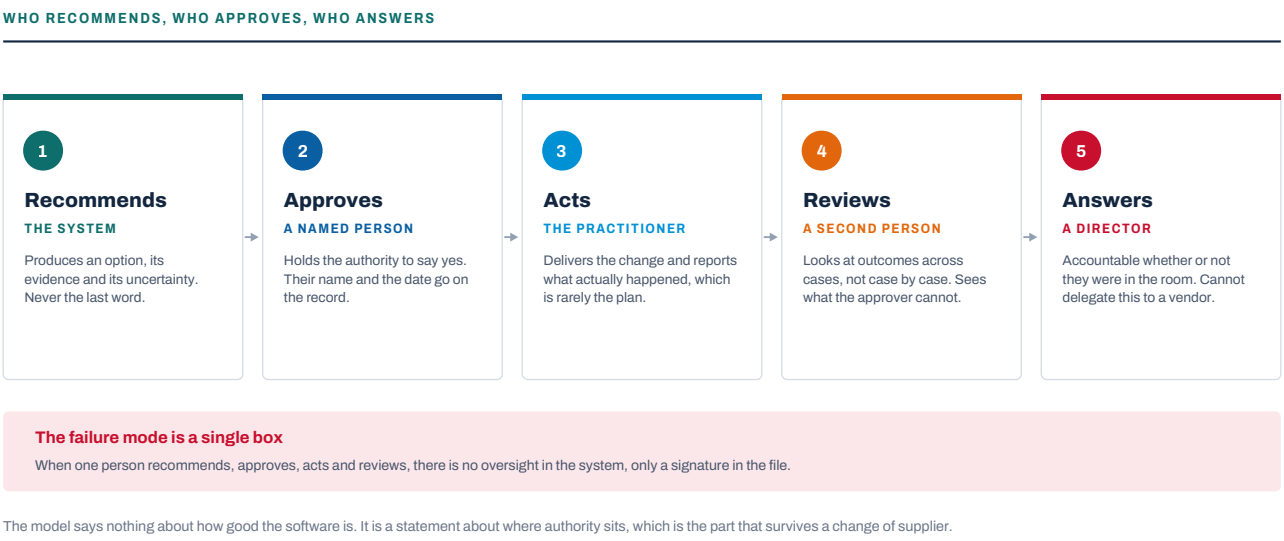


FIGURE 7 Five roles that must not collapse into one person. The model describes where authority sits, which is the part that survives a change of supplier.

Athleet.AI framework, informed by references 29 and 33. ATHLEET.AI

8.1 What the evidence says about oversight

Automation bias is well documented. A systematic review of 74 studies drawn from 13,821 retrieved papers found that although most research shows overall improved performance with clinical decision support, there is often a failure to recognise the new errors that such systems introduce. The mediators it identified were user cognitive style, system characteristics, task-specific experience, trust and confidence, together with environmental mediators of workload, task complexity and time constraint, which pressurise cognitive resources.[29] REVIEW Those three environmental mediators describe a training ground and a match day exactly. The mitigators the same review identified were organisational, being training and emphasising user accountability, alongside design choices such as presenting information rather than a recommendation.

The older human factors literature frames the same problem as misuse, meaning over-reliance on automation, alongside disuse, meaning neglect, and abuse, meaning automation applied without regard to human consequences.[30,31]

PEER-REVIEWED Sport manages to exhibit all three at once, often within the same department.

The European legislature reached the same conclusion. The Artificial Intelligence Act obliges providers of high-risk systems to enable overseers to remain aware of the possible tendency of automatically relying or over-relying on the output, and names automation bias in the legislative text. It requires that the overseer be able to decide not to use the system, or to disregard, override or reverse its output. For remote biometric identification it goes further and requires separate verification by at least two people.[33] CONSENSUS The legislature evidently did not think a single signature was sufficient oversight either.

I should be honest about the limit of this argument. I could not find a study quantifying how often human override fails in a deployed sporting system, and I am not going to invent one. The case rests on mechanism and on the legislature's own concession, which is weaker than a number and still strong enough to act on.

8.2 The five roles

The system recommends. A named person approves. A practitioner acts and reports what actually happened, which is rarely the plan. A second person reviews outcomes across cases rather than case by case, because patterns are invisible from inside a single decision. And a director answers, whether or not they were in the room, and cannot delegate that to a vendor.

The failure mode is a single box. Where one person recommends, approves, acts and reviews, there is no oversight in the system, only a signature in the file. That configuration is extremely common in small performance departments, and the honest response is not to pretend otherwise but to name a reviewer from outside the department, which costs an hour a month.

ILLUSTRATIVE SCENARIO

The compliance posture that reads well and does nothing

An invented institute publishes a data protection protocol stating that it profiles athletes for pathway identification and funding recommendations, and that decisions are made with human input rather than being automated or determined by algorithms. Every word is true and the document is well drafted. What it does not say is who the human is, what they see before deciding, what proportion of the model's recommendations they have ever changed, or what happens to the record of a decision they did change. The protocol establishes that a person is present. It establishes nothing about whether that presence is doing any work.

A real document of this kind exists in the United Kingdom high-performance system, and it is worth reading precisely because it is careful and public.[32] CONSENSUS The point of the invented version above is not that such statements are dishonest. It is that a statement of human involvement is untestable unless the organisation also records the override rate, and almost nobody does.

8.3 Making oversight testable

Three measurements convert a claim of human oversight into something a board can audit. The proportion of recommendations that were changed before action, which should never be zero and rarely above half. The proportion of decisions where the recorded reason differs from the system's stated reason, which is where practitioner expertise is actually entering. And the time between recommendation and decision, because a median measured in seconds is a rubber stamp regardless of what the policy says.

None of those three requires new software. All three require the club to record the decision alongside the recommendation, which is the habit this whole paper reduces to.

CHAPTER NINE

Safeguarding, health and the medical boundary

There is a line in law where a performance tool becomes a regulated medical device. A good deal of what is being sold sits on the wrong side of it.

This chapter draws two boundaries. The first is around children, where the applicable law is stronger than most performance departments realise and the sport-specific guidance is weaker. The second is around health, where the regulatory definition is more expansive than most suppliers admit.

9.1 Children

The Age Appropriate Design Code applies to information society services likely to be accessed by children in the United Kingdom and sets 15 standards, among them the best interests of the child, data protection impact assessments, transparency, detrimental use of data, default settings, data minimisation, data sharing, restrictions on profiling and restrictions on nudge techniques.[41] CONSENSUS It came into force in September 2020 following a transition period and is a statutory code rather than optional guidance. Its profiling standard requires that options using profiling are off by default unless a compelling reason exists, and that profiling is allowed only where measures are in place to protect the child from harmful effects.

Whether the code applies directly to a club's internal player-management system is arguable, since such a system may not be an information society service likely to be accessed by children. A club relying on it should say that it is applying the standards by analogy, as best practice, and should lead on the impact assessment obligation in Article 35, which applies to any controller regardless.[39] CONSENSUS

What does not exist is sport-specific guidance. The standards for safeguarding and protecting children in sport, published in 2002 and revised in 2018, run to ten headings covering policy and procedures, operating systems, prevention, codes of conduct, equity, communication, education and training, access to advice, implementation and influencing.[50] CONSENSUS Not one of them addresses data, technology, profiling or automated decision-making. I searched for published guidance on artificial intelligence and children's data in sport and found none, and a bounded search of the peer-reviewed literature on children's data in sports technology returned almost nothing on point. That absence is the strongest single argument for a club writing its own position rather than waiting.

WHAT THIS RULES OUT

No score attached to a child. No ranking of children by predicted adult potential. No system that communicates with a child without an adult in the loop. No profiling of a young athlete without a completed impact assessment that considers the child's interests specifically. The series' third and fourth papers develop these into practical requirements for grassroots athletics and academy football.

9.2 The medical boundary

The second boundary is drawn by medical device regulation, and it is drawn further out than the sports technology market behaves as though it is. In the European Union, a medical device is defined to include software intended for one or more specified medical purposes, and the list of purposes expressly names prediction and prognosis, and expressly names injury.[44] CONSENSUS Software intended to provide information used to take decisions with diagnostic or therapeutic purposes is classified as at least class IIa under Rule 11, rising with the severity of the consequence.

United Kingdom guidance draws the same line in more practical language. Monitoring general fitness, health and wellbeing is not usually considered a medical purpose, and apps monitoring sport or fitness measures such as heart rate are not considered medical devices, although in specific cases where the intention is to investigate physiological processes they may be. Software is most likely to be a device where it results in a diagnosis or prognosis, providing future risk of disease, and decision support software is usually considered a medical device when it applies automated reasoning such as a calculation or an algorithm.[45] CONSENSUS

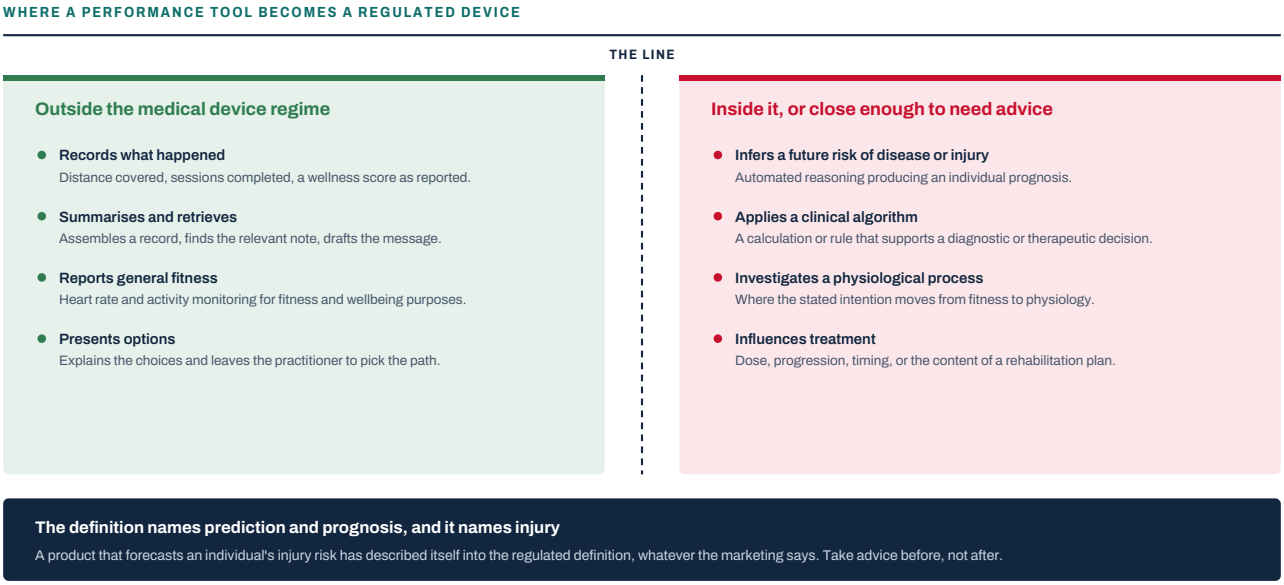


FIGURE 8 The line as drawn by the regulator's own tests. The right-hand column is not a list of bad products. It is a list of products that need advice before deployment.

References 44 and 45. The allocation of examples to each side is the author's own reading and is not a regulatory determination. PRACTICE

Read those two sources together and the implication for the market is uncomfortable. A product that forecasts an individual athlete's future injury risk has described itself into the regulatory definition, whatever the marketing says. My own view, which a club should verify with its own advisers rather than accept from this paper, is that a substantial share of the injury-prediction products sold

into sport have never tested themselves against Rule 11, and that the question is more likely to be asked by a regulator than by a buyer.

The practical guidance is simple. If a system outputs a future risk of injury or illness for a named person, treat it as needing advice before deployment rather than after. If it reports what happened and leaves the inference to a clinician, it is on the other side of the line. Chapter 3's refusal quadrant and this boundary are the same boundary seen from two directions.

CHAPTER TEN

Knowledge ownership and intellectual property

The most valuable asset in a performance department is a coaching method held in five or six heads. Nobody has decided who owns it.

Three kinds of knowledge meet inside a performance system and they are governed by different rules that almost nobody separates. There is athlete data, which is personal data and belongs to a regime described in Chapter 4. There is the supplier's model, which is the supplier's intellectual property and normally not disclosed. And there is coaching methodology, which is the reason the club is any good, which has usually never been written down, and which is silently transferred to the supplier the moment the club starts entering it into their software.

That third category is where the value leaks. A club that spends a decade developing a way of preparing athletes, and then enters that method into a platform under terms permitting the supplier to improve its models using customer data, has given away the asset it was trying to protect. The transaction is rarely deliberate and rarely noticed, because it looks like data entry rather than like a transfer.

10.1 What to settle in the contract

THE ONE CLAUSE

Whether the supplier may train its own models on your data is the single clause with the largest long-term consequence in most performance software contracts.

Four positions, stated plainly. Whether the supplier may use club or athlete data to develop or improve models the club does not own. Whether coaching content entered into the system is licensed to the supplier and on what terms. What the club receives on exit, in what format, and whether that includes the derived record rather than only the raw inputs. And whether the supplier may reference the club in marketing, which sounds trivial and is the clause most often used to establish that a relationship was closer than it was.

Because a processor may act only on the controller's documented instructions, a club that has not addressed model training in its contract has not established that it does not happen.[38] CONSENSUS That is a legal observation, not legal advice, and a club should take its own.

10.2 The case for writing the method down

There is a second reason to take this seriously that has nothing to do with contracts. A method that lives only in people leaves when they do. Every performance director I have spoken to can name a period when a change of staff cost the organisation years, and none of them can produce the document that would have

prevented it. Artificial intelligence is genuinely useful here, and the use is unglamorous. A system that captures why a decision was made, in the practitioner's own words, at the time it was made, produces the institutional memory that no amount of prediction ever will.

My own view is that this is the most undervalued application in the sector. It does not require a model at all. It requires a habit and a place to put the record, and it is the application least likely to be sold to you because there is no algorithm to demonstrate.

CHAPTER ELEVEN

Procurement and assurance

Twelve questions asked before signature will tell a buyer more than any demonstration.

Sport has no procurement framework of its own for artificial intelligence. I looked, and found none published before this paper was written. What does exist is public-sector guidance that transfers without modification, and a governing body in receipt of public funding can adopt it directly and defensibly.

United Kingdom government guidelines for procuring artificial intelligence set out ten considerations, and the eighth of them is to avoid black box algorithms and vendor lock-in.[48] CONSENSUS That is government procurement policy, published in 2020, and it says in four words what the methodological literature takes a paper to say. A parallel multi-stakeholder toolkit published the same month covers the same ground for public buyers internationally.[49] CONSENSUS Neither was written for sport. Both work.

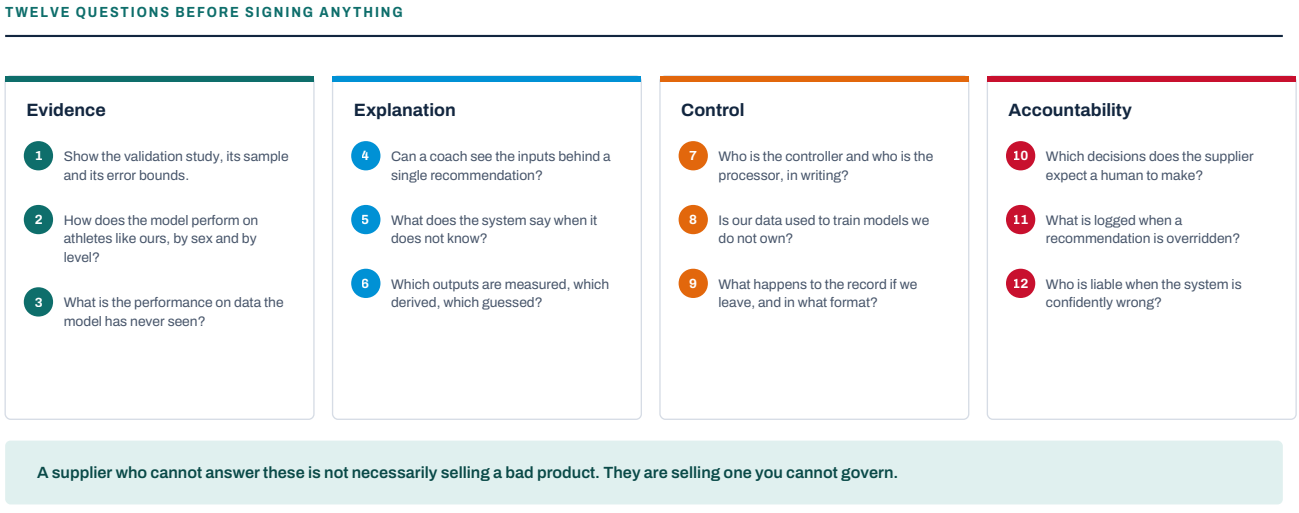


FIGURE 9 The responsible sports AI procurement scorecard. A supplier who cannot answer these is not necessarily selling a bad product. They are selling one that cannot be governed.

Athleet.AI framework, informed by references 48 and 49. ATHLEET.AI

11.1 How to use the twelve

Send them in writing before the demonstration rather than after it. A demonstration is designed to answer questions the supplier chose, and the 12 are designed to answer questions the club needs. Score the responses as answered, partially answered or not answered, and treat the pattern of non-answers as the finding. In my experience the questions that go unanswered cluster, and the cluster tells you what kind of company you are dealing with.

Two of the 12 deserve particular weight. The question about performance on data the model has never seen separates suppliers who have done external validation from those who have reported training performance and hoped nobody asked. The question about what is logged when a recommendation is overridden separates products designed for governed use from products designed for demonstration.

11.2 Standards a buyer can point at

Three instruments give a board something external to reference without commissioning anything. A national risk management framework for artificial intelligence organises the work into four functions, being govern, map, measure and manage.[46] CONSENSUS An international management system standard for artificial intelligence exists and is certifiable, which matters because it lets a club ask a supplier for evidence rather than assurance.[47] CONSENSUS And an internal auditing framework from the research literature sets out five stages of scoping, mapping, artifact collection, testing and reflection, each producing documents that together form an audit report.[16] CONSENSUS

None of these is a certification a sports body needs. All three are useful as a vocabulary, because they let a performance director describe what they want in terms a technology supplier already recognises.

SPECIFICATION

What to require in the contract, beyond the usual

- The validation study, its population and its error bounds, supplied before signature rather than on request.
- A statement of which decisions the supplier expects a human to make, in the supplier's own words.
- Export of the derived record, not only the raw inputs, in a documented format, on 30 days' notice.
- An explicit position on whether club or athlete data may be used to train models the club does not own.
- Notification when the model changes materially, because a system that is revalidated silently is a different system.

CHAPTER TWELVE

Proving value without overclaiming

Availability is the honest value metric. It is measurable, auditable, prospectively validated, and it beats the models on their own ground.

Every organisation that buys a performance system eventually has to justify it, and the temptation at that moment is to claim a causal effect the data cannot carry. There is a better argument available, and it has the advantage of being true.

FIGURE 10 The chain a club can audit at every link. The two figures are from prospective cohort studies, and the chain is an association rather than a demonstrated causal path.

References 57 and 58. The chain itself is the author's own framework. ATHLEET.AI

12.1 What the availability evidence supports

Researchers followed 24 European football clubs prospectively for 11 seasons, covering 155 team-seasons, 7,792 injuries and 1,026,104 hours of exposure. Total injury incidence was 7.7 injuries per 1,000 hours, injury burden was 130 days lost per 1,000 hours and player match availability was 86 per cent. Lower injury burden and higher match availability tracked a higher final league position; lower incidence, lower burden and higher availability tracked more points per match; and lower burden and higher availability tracked an increase in the club's European coefficient.[57] PEER-REVIEWED

The same relationship holds in an individual sport, where 33 international track and field athletes were monitored weekly across 76 athlete-seasons over five years. Athletes who completed more than 80 per cent of planned training weeks were seven times more likely to achieve a performance goal, at an area under the curve of 0.72. Training availability accounted for 86 per cent of successful seasons, and for every modified training week the chance of success fell.[58] PEER-REVIEWED

Set that 0.72 beside the 0.52 from Chapter 5. A counted threshold, requiring no model at all, outperforms machine-learned injury prediction on the discrimination measure the modellers themselves chose. That comparison is not quite like for like, since the studies ask different questions in different sports, and I say so

rather than let the rhetoric run. It remains the most useful single fact in this paper for anyone deciding what to measure.

12.2 What may and may not be claimed

Both studies are prospective cohorts, not trials. They establish association between availability and success, in populations that differ from most clubs. A club may reasonably act on them. A supplier may not advertise them as proof that software caused a result, and a club should refuse to repeat such a claim on a supplier's behalf.

The defensible claim is narrower and more useful. A system that improves the quality and the recording of decisions should, if the association holds, show up first as more completed planned work, then as fewer interrupted preparations, and only then as competitive outcomes, which are influenced by a great many things software does not touch. Measure the first, expect the second and stop short of promising the third.

THE ARGUMENT IN ONE PARAGRAPH

Do not try to prove that artificial intelligence made the team better. Prove that more athletes completed more of what was planned, that more decisions have a recorded reason, and that the reasons stand up when reviewed. Those are countable, they are auditable, and the published evidence links them to outcomes in both a team sport and an individual one. Everything beyond that is a claim the data will not carry, and making it is how credibility is lost.

CHAPTER THIRTEEN

The responsible performance system

Three meetings, one register and one habit. The operating model is deliberately small, because a large one will not be adopted.

Everything in this paper reduces to an operating model that fits into meetings a club already holds. I have deliberately not proposed an ethics committee, a new head of anything, or a policy document with a version number. Those exist in this sector and they do not work, because they sit beside the decisions rather than inside them.

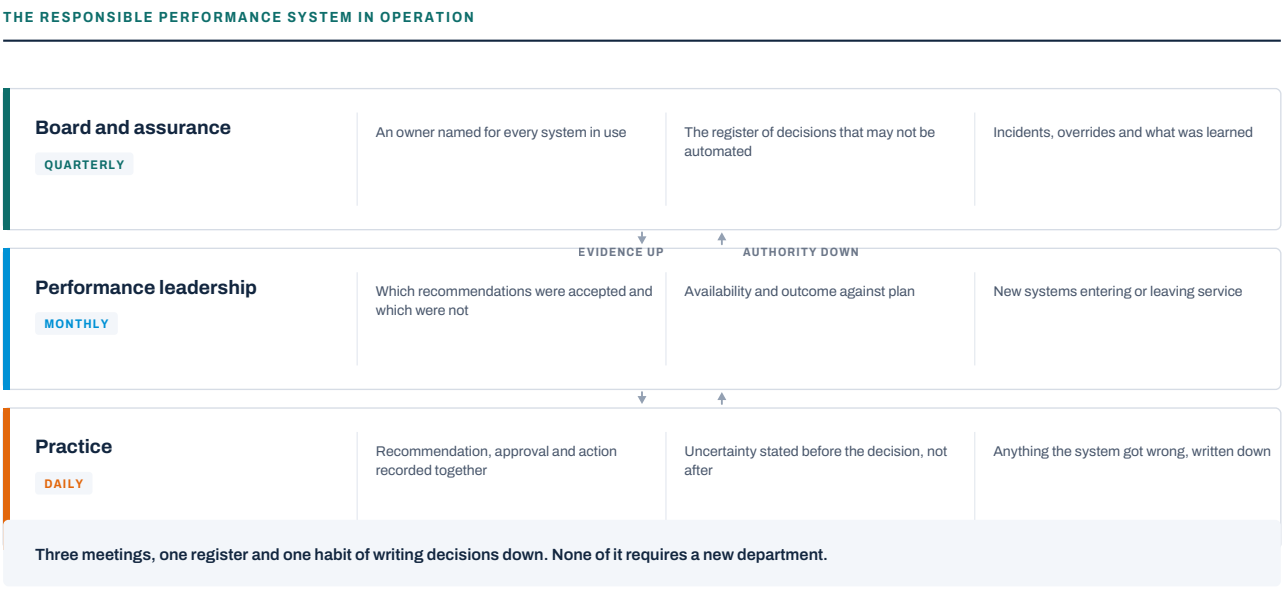


FIGURE 11 Three tiers, three cadences and the artefacts each produces. Evidence travels upward and authority travels downward, which is the opposite of how most governance in sport is drawn.

Athleet.AI framework. ATHLEET.AI

13.1 The register

One page, four columns. Every system in service, the family it belongs to from Chapter 2, the decisions it touches from Chapter 3, and the person accountable for it. Alongside it, the shorter and more important list of decisions this organisation will not automate. That second list is the one to publish internally, because it tells practitioners what the software will never do to them, which is the fastest way to earn their attention for what it does.

13.2 The three cadences

Daily, at the point of practice, the recommendation, the approval and the action are recorded together, uncertainty is stated before the decision rather than after, and anything the system got wrong is written down. Monthly, performance leadership looks at which recommendations were accepted and which were not, at availability and outcome against plan, and at systems entering or leaving service. Quarterly, the board sees the register, the list of decisions that may not be automated, and the incidents, overrides and what was learned from them.

None of that requires a new meeting. It requires three existing meetings to have one additional standing item, and it requires somebody to own the register.

13.3 The first 100 days

In the first month, complete the register and the decision matrix, and expect both to surprise you. In the second, write the four contract positions from Chapter 4 and the four from Chapter 10, and send the 12 procurement questions to every supplier already in service rather than only to new ones. In the third, start recording decisions alongside recommendations for one squad, measure the override rate, and review it. Do not begin with a technology decision. Beginning with a technology decision is how organisations end up governing a product instead of a practice.

ILLUSTRATIVE SCENARIO

What the first quarterly review usually finds

An invented federation completes its first cycle. The register lists 11 systems, four of which nobody had approved. The override rate on the load management tool is zero, which the performance director initially reports as evidence that the model is good and which the review correctly reads as evidence that nobody is looking. Two decisions in the quarter were recorded with a reason that differed from the system's, and both turned out well, which is the first genuine sign that expertise is entering the system. The most useful item on the agenda is a single line noting that the wearable's sleep staging has never been validated against anything, and that three decisions were made on it anyway.

13.4 What good looks like in three years

A programme that has done this properly will not look dramatically different from outside. The same coaches will be coaching, the same arguments will be had, and the athletes will still do the work. What will have changed is that a decision made in September can be found in April with its reasoning intact, that a new performance director inherits a method rather than folklore, that an athlete leaving takes their own record with them, and that nobody is being asked to trust a number whose provenance they cannot establish.

That is a modest description of success and a demanding one to achieve. It is also, on the evidence assembled here, considerably more valuable than the prediction the sector keeps being offered instead.

A club with a mediocre model and excellent governance will outperform a club with an excellent model and none, because only the first club is capable of learning.

CHAPTER ONE

END MATTER

Board readiness checklist

Twenty items a board can work through in one session. The first six cost nothing and find most of the problems.

- [ ] **The register exists**
  Every system in service is listed with its family, the decisions it touches and a named owner.
- [ ] **The refusal list is published internally**
  Practitioners can see which decisions this organisation will never automate.
- [ ] **The decision matrix is complete**
  Filled in by the people who make the decisions, not by the people who buy the software.
- [ ] **Impact assessments precede deployment**
  Not after a pilot, which answers a different question.
- [ ] **Controller and processor are named in writing**
  For every supplier, in the contract rather than in an assumption.
- [ ] **The model training position is stated**
  Whether a supplier may train its own models on club or athlete data, as a yes or a no.
- [ ] **Validation studies are held on file**
  With population, sample and error bounds, for every predictive claim in service.
- [ ] **Performance on unseen data is known**
  Or the absence of that figure is recorded as a known gap.
- [ ] **Subgroup performance is reported or declared unknown**
  By sex and by level, for the athletes this organisation actually has.
- [ ] **Recommendations carry provenance**
  Every input named and dated, so a disputed number can be traced.
- [ ] **Uncertainty is stated in usable words**
  Before the decision, not in a footnote afterwards.
- [ ] **Alternatives are offered**
  A recommendation with no alternative is an instruction.
- [ ] **The five roles are separated**
  Nobody recommends, approves, acts and reviews the same decision.
- [ ] **The override rate is measured**
  Zero is a finding, not a success.
- [ ] **Time to decision is measured**
  A median in seconds is a rubber stamp whatever the policy says.
- [ ] **No score is attached to anybody under 18**
  And no ranking of children by predicted adult potential.
- [ ] **Health inference is identified and advised on**
  Any system outputting individual future injury or illness risk has been through legal advice.
- [ ] **Exit terms include the derived record**
  Not only the raw inputs, in a documented format.
- [ ] **Athletes can obtain their own record**
  In a usable form, within a stated period, on request.
- [ ] **Availability against plan is the headline measure**
  Reported to the board alongside, and ahead of, any predictive output.

END MATTER

Glossary

Area under the curve
A measure of how well a model separates the people to whom an event happens from those to whom it does not. A value of 0.5 is chance and 1.0 is perfect separation.

Automation bias
The documented tendency to rely on an automated output more than the evidence warrants, and to miss errors the system introduces.

Calibration
Whether predicted risks match observed frequencies. A model can rank athletes well and still be systematically overconfident.

Controller and processor
The organisation that decides why and how personal data is processed, and the organisation that processes it on the controller's documented instructions.

Data protection impact assessment
An assessment required before high-risk processing begins, covering the processing, its necessity, its risks and the measures addressing them.

Discrimination
A model's ability to rank one person above another. Not sufficient on its own for a decision about an individual.

Explainability by construction
Designing a system so that its reasoning is visible, rather than generating an explanation afterwards by a separate mechanism.

External validation
Testing a model on data it has never seen, ideally from a different population, season or organisation.

Override rate
The proportion of system recommendations changed by a human before action. The most useful single measure of whether oversight is real.

Proxy label
Something measurable, standing in for the thing actually of interest. The most common source of consequential bias in applied machine learning.

Provenance
The traceable origin of every input behind an output, with its source and its date.

Special category data
Data requiring a further condition beyond a lawful basis, including data concerning health and biometric data used to identify a person.

END MATTER

Method and evidence grading

How the sources in this paper were selected, checked and graded, and what the reader should distrust.

Sources were gathered from the peer-reviewed literature, primary legislation, regulator guidance, national standards bodies and governing-body publications, with priority given to systematic reviews, methodological criticism and primary legal instruments over single primary studies. Every reference in the list that follows was checked against the publishing journal, publisher, issuing body or the legislation itself, and then checked a second time by an independent process against the same sources. Point-in-time legislative text was used where the position has since changed.

What could not be verified

Five things were searched for and not found, and their absence shapes this paper. No sport-specific guidance on artificial intelligence and children's data was located from any safeguarding body. No artificial intelligence procurement frame-

work written for sport was found. No binding governance instrument on performance AI was found from any international federation, and the strategy documents that exist are aspirational rather than operative. No peer-reviewed application of contextual integrity to sport was found, so where this paper reasons in that direction it does so on its own account. And no study quantifying the failure rate of human override in a deployed sporting or clinical system was found, so Chapter 8 argues from mechanism rather than from a number.

Two sources are cited with an explicit limitation. The operating guidelines for the athlete biological passport are quoted from the version the author was able to obtain rather than the most recent, and the reference says which. The strategy document published by the international body in 2024 is characterised from its own launch release rather than from the document text, which could not be obtained.

What to distrust in this paper

Three things. The decision-risk matrix in Chapter 3, the accountability model in Chapter 8 and the operating model in Chapter 13 are the author's own frameworks, proposed rather than validated, and they are published in a form designed to invite disagreement. The comparison in Chapter 12 between an availability threshold and machine-learned injury prediction sets side by side studies that asked different questions in different sports, and the text says so where the comparison is made. And the readings of the medical device boundary and of the scope of the European regulation are the author's, offered to prompt proper advice rather than to substitute for it.

Conflict of interest

The author is the founder of Athleet.AI, which builds software in the area this paper examines. The paper recommends against several categories of product that would be commercially straightforward to build and sell, and it proposes controls that would make the author's own product harder to sell badly. Readers should weigh the argument accordingly. Test the citations rather than the author.

END MATTER

About the author

Paul Forrest is the founder and Chief AI Officer of Athleet.AI. He works with organisations on the deployment of artificial intelligence, having begun in business process reengineering in the 1990s and moved into natural language and machine learning work from 2014, building small, domain-specific models in preference to general-purpose ones.

He is a qualified athletics coach, holding level three qualifications in sprints, hurdles and endurance, and he coaches as a volunteer at club level. That combination of enterprise systems work and time on a track is the reason this series exists, and it is also the reason the papers decline to recommend several things that would be straightforward to sell.

He is the author of *Tales from AI Development Hell* and *AI Made the Decision: Who Answers When Nobody Owns the Outcome?*. This paper is the flagship of the series and the one the other six draw their assurance model from.

END MATTER

References

Fifty-eight sources, each verified twice against the publishing journal, publisher, issuing body or the legislation itself.

What the prediction evidence shows

1. **Ruddy, J. D., Shield, A. J., Maniar, N., Williams, M. D., Duhig, S., Timmins, R. G., Hickey, J., Bourne, M. N., & Opar, D. A.** (2018). Predictive modeling of hamstring strain injuries in elite Australian footballers. *Medicine & Science in Sports & Exercise*, 50(5), 906–914. doi:10.1249/MSS.0000000000001527
2. **Carey, D. L., Ong, K., Whiteley, R., Crossley, K. M., Crow, J., & Morris, M. E.** (2018). Predictive modelling of training loads and injury in Australian football. *International Journal of Computer Science in Sport*, 17(1), 49–66. doi:10.2478/ijcss-2018-0002
3. **Impellizzeri, F. M., Woodcock, S., Coutts, A. J., Fanchini, M., McCall, A., & Vigotsky, A. D.** (2021). What role do chronic workloads play in the acute to chronic workload ratio? Time to dismiss ACWR and its underlying theory. *Sports Medicine*, 51(3), 581–592. doi:10.1007/s40279-020-01378-6
4. **Lolli, L., Batterham, A. M., Hawkins, R., Kelly, D. M., Strudwick, A. J., Thorpe, R., Gregson, W., & Atkinson, G.** (2019). Mathematical coupling causes spurious correlation within the conventional acute-to-chronic workload ratio calculations. *British Journal of Sports Medicine*, 53(15), 921–922. First published online 3 November 2017. doi:10.1136/bjsports-2017-098110
5. **Bahr, R.** (2016). Why screening tests to predict injury do not work, and probably never will. A critical review. *British Journal of Sports Medicine*, 50(13), 776–780. doi:10.1136/bjsports-2016-096256
6. **Bullock, G. S., Hughes, T., Arundale, A. H., Ward, P., Collins, G. S., & Kluzek, S.** (2022). Black box prediction methods in sports medicine deserve a red card for reckless practice. A change of tactics is needed to advance athlete care. *Sports Medicine*, 52(8), 1729–1735. doi:10.1007/s40279-022-01655-6
7. **Soligard, T., Schwellnus, M., Alonso, J.-M., Bahr, R., Clarsen, B., Dijkstra, H. P., et al.** (2016). How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury. *British Journal of Sports Medicine*, 50(17), 1030–1041. doi:10.1136/bjsports-2016-096581
8. **Schwellnus, M., Soligard, T., Alonso, J.-M., Bahr, R., Clarsen, B., Dijkstra, H. P., et al.** (2016). How much is too much? (Part 2) International Olympic Committee consensus statement on load in sport and risk of illness. *British Journal of Sports Medicine*, 50(17), 1043–1052. doi:10.1136/bjsports-2016-096572
9. **Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., Gabbett, T. J., Coutts, A. J., Burgess, D. J., Gregson, W., & Cable, N. T.** (2017). Monitoring athlete training loads. Consensus statement. *International Journal of Sports Physiology and Performance*, 12(s2), S2-161–S2-170. doi:10.1123/IJSPP.2017-0208
10. **Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., et al.** (2020). Prediction models for diagnosis and prognosis of covid-19. Systematic review and critical appraisal. *BMJ*, 369, m1328. Figures quoted are from the version searched to 7 April 2020. doi:10.1136/bmj.m1328

Validation, calibration and documentation

1. **Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M.** (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). The TRIPOD statement. *BMC Medicine*, 13, 1. doi:10.1186/s12916-014-0241-z
2. **Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W.** (2016). A calibration hierarchy for risk models was defined. From utopia to empirical data. *Journal of Clinical Epidemiology*, 74, 167–176. doi:10.1016/j.jclinepi.2015.12.005
3. **Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W.** (2019). Calibration. The Achilles heel of predictive analytics. *BMC Medicine*, 17, 230. doi:10.1186/s12916-019-1466-7
4. **Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T.** (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. doi:10.1145/3287560.3287596
5. **Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K.** (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. doi:10.1145/3458723
6. **Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P.** (2020). Closing the AI accountability gap. Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. doi:10.1145/3351095.3372873

Explainability and interpretability

1. **Rudin, C.** (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. doi:10.1038/s42256-019-0048-x
2. **Lipton, Z. C.** (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. doi:10.1145/3233231
3. **Ghorbani, A., Abid, A., & Zou, J.** (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 3681–3688. doi:10.1609/aaai.v33i01.33013681
4. **Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B.** (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems 31*. No DOI or page range is assigned by the proceedings. arXiv: 1810.03292
5. **Information Commissioner's Office & The Alan Turing Institute** (2020). *Explaining Decisions Made with AI*. Wilmslow, ICO. Draft issued for consultation 2 December 2019; final guidance published 20 May 2020. Non-statutory guidance.
6. **Gunning, D., & Aha, D. W.** (2019). DARPA's explainable artificial intelligence program. *AI Magazine*, 40(2), 44–58. doi:10.1609/aimag.v40i2.2850

Bias, inclusion and uneven evidence

1. **Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S.** (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. doi:10.1126/science.aax2342
2. **Buolamwini, J., & Gebru, T.** (2018). Gender shades. Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81, 77–91.
3. **Costello, J. T., Bieuzen, F., & Bleakley, C. M.** (2014). Where are all the female participants in sports and exercise medicine research? *European Journal of Sport Science*, 14(8), 847–851. doi:10.1080/17461391.2014.911354
4. **Cowley, E. S., Olenick, A. A., McNulty, K. L., & Ross, E. Z.** (2021). "Invisible sportswomen". The sex data gap in sport and exercise science research. *Women in Sport and Physical Activity Journal*, 29(2), 146–151. doi:10.1123/wspaj.2021-0028
5. **Elliott-Sale, K. J., Minahan, C. L., de Jonge, X. A. K. J., Ackerman, K. E., Sipilä, S., Constantini, N. W., Lebrun, C. M., & Hackney, A. C.** (2021). Methodological considerations for studies in sport and exercise science with women as participants. A working guide for standards of practice for research on women. *Sports Medicine*, 51(5), 843–861. doi:10.1007/s40279-021-01435-8
6. **Karkazis, K., & Fishman, J. R.** (2017). Tracking U.S. professional athletes. The ethics of biometric technologies. *The American Journal of Bioethics*, 17(1), 45–60. doi:10.1080/15265161.2016.1251633

Human oversight and automation bias

1. **Goddard, K., Roudsari, A., & Wyatt, J. C.** (2012). Automation bias. A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. doi:10.1136/amiajnl-2011-000089
2. **Parasuraman, R., & Riley, V.** (1997). Humans and automation. Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. doi:10.1518/001872097778543886

3. **Skitka, L. J., Mosier, K. L., & Burdick, M.** (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006. doi:10.1006/ijhc.1999.0252

4. **UK Sport** (2024). *UK Sport Data Protection Protocol for Athletes*, version 04. London, UK Sport. Initial version 15 October 2019; version 04 issued July 2024. Section F addresses profiling in pathway and funding decisions.

Law, regulation and statutory codes

1. **European Parliament and Council of the European Union** (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, OJ L, 2024/1689, 12 July 2024. In force 1 August 2024, applying in stages under Article 113. Articles 5(1)(f), 14(4) and Annex III point 4 referenced in the text.

2. **United Kingdom General Data Protection Regulation**, Articles 15, 16, 17 and 20. Rights of access, rectification, erasure and data portability. Text as in force 31 March 2025.

3. **United Kingdom General Data Protection Regulation**, Article 5. Principles relating to processing, including the accountability principle at Article 5(2). Text as in force 31 March 2025.

4. **United Kingdom General Data Protection Regulation**, Article 22. Automated individual decision-making, including profiling, with the United Kingdom insertion at Article 22(3A). Text as in force 31 March 2025.

5. **United Kingdom General Data Protection Regulation**, Article 25. Data protection by design and by default. Text as in force 31 March 2025.

6. **United Kingdom General Data Protection Regulation**, Article 28. Processor obligations, including the requirement to process only on documented instructions from the controller. Text as in force 31 March 2025.

7. **United Kingdom General Data Protection Regulation**, Article 35. Data protection impact assessment, including Article 35(3)(a) on systematic and extensive automated evaluation. Text as in force 31 March 2025.

8. **Data (Use and Access) Bill** (2024–25). Introduced in the House of Lords 23 October 2024. At 31 March 2025 the Bill had completed all Lords stages and Commons Committee stage and awaited Commons Report. It had not received Royal Assent at the date of this paper.

9. **Information Commissioner's Office** (2020). *Age Appropriate Design. A Code of Practice for Online Services.* Wilmslow, ICO. Laid before Parliament 11 June 2020; issued 12 August 2020; in force 2 September 2020, transition ended 2 September 2021.

10. **Data Protection Act 2018**, sections 121 to 125 and section 127. The four statutory codes the Commissioner must prepare, and the effect of a failure to act in accordance with a code.

11. **Department for Science, Innovation and Technology** (2023, 2024). *A Pro-Innovation Approach to AI Regulation*, CP 815, 29 March 2023, and the government response, CP 1019, 6 February 2024. Five non-statutory cross-sectoral principles issued to existing regulators.

12. **European Parliament and Council of the European Union** (2017). Regulation (EU) 2017/745 on medical devices. *Official Journal of the European Union*, OJ L 117/1, 5 May 2017. Article 2(1) definition and Annex VIII Rule 11. Applied from 26 May 2021 following Regulation (EU) 2020/561.

13. **Medicines and Healthcare products Regulatory Agency** (2023). *Guidance. Medical Device Stand-Alone Software Including Apps*, version 1.10f. London, MHRA. First published 8 August 2014, last substantively updated 1 July 2023.

Standards, procurement and safeguarding

1. **Tabassi, E.** (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1. Gaithersburg, National Institute of Standards and Technology. Published 26 January 2023. doi:10.6028/NIST.AI.100-1

2. **International Organization for Standardization & International Electrotechnical Commission** (2023). *ISO/IEC 42001:2023 Information Technology. Artificial Intelligence. Management System.* Edition 1, published December 2023.

3. **Office for Artificial Intelligence** (2020). *Guidelines for AI Procurement.* London, UK Government. Published 8 June 2020. Consideration eight is to avoid black box algorithms and vendor lock-in.

4. **World Economic Forum** (2020). *AI Procurement in a Box.* Geneva, World Economic Forum. Published 11 June 2020.

5. **NSPCC Child Protection in Sport Unit** (2018). *Standards for Safeguarding and Protecting Children in Sport.* Second edition; first published 2002, revised 2018.

Athlete rights, integrity and value

1. **FIFPRO** (2022). *Charter of Player Data Rights.* Hoofddorp, FIFPRO. Launched 19 September 2022 and developed with FIFA, setting out eight rights over performance and personal data.

2. **World Players Association** (2017). *Universal Declaration of Player Rights.* Adopted by the Executive Committee in Paris, 7 April 2017. Seventeen articles; Articles 11 and 12 concern data and image.

3. **Skelton, S. K.** (2020). Footballers to take legal action over use of performance and tracking data. *Computer Weekly*, 6 August 2020. An announced group claim, not an adjudicated case.

4. **Osborne, B., & Cunningham, J. L.** (2017). Legal and ethical implications of athletes' biometric data collection in professional sport. *Marquette Sports Law Review*, 28(1), 37.

5. **World Anti-Doping Agency** (2019). *Athlete Biological Passport Operating Guidelines*, version 7.1. Montreal, WADA. Quoted from version 7.1; a later version 8.0 exists and was not obtainable.

6. **International Olympic Committee** (2024). *Olympic AI Agenda.* Launched 19 April 2024, London. Characterised in the text from the IOC's own launch release rather than the document text.

7. **Hägglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H., & Ekstrand, J.** (2013). Injuries affect team performance negatively in professional football. An 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, 47(12), 738–742. doi:10.1136/bjsports-2013-092215

8. **Raysmith, B. P., & Drew, M. K.** (2016). Performance success or failure is influenced by weeks lost to injury and illness in elite Australian track and field athletes. A 5-year prospective study. *Journal of Science and Medicine in Sport*, 19(10), 778–783. doi:10.1016/j.jsams.2015.12.515

WHERE THIS GOES NEXT

A responsible sports AI readiness assessment

The argument in this paper is testable, and the cheapest way to test it is to complete the register and the decision matrix before buying anything else. Most organisations find something on the register they did not know was running, and most find that the decision they most want help with is one this paper says should never be automated.

Athleet.AI runs readiness assessments with federations, clubs, institutes and technology companies. The output is a register, a completed decision matrix, a procurement position and a list of the gaps that matter, delivered as a document the organisation owns and can act on without us.

What an assessment involves

- Two workshops with the people who make the decisions, not only the people who buy the software.
- A completed register of systems in service, with families, decisions touched and named owners.
- The decision-risk matrix filled in for this organisation, with its own refusal list.
- A written procurement position and the twelve questions issued to existing suppliers.
- A measurement baseline covering availability against plan and the current override rate.