Epistemic Certainty in Recruitment: The Gap Between What Football Knows and What It Claims to Know

Football's recruitment infrastructure now produces three distinct layers of confidence — the analyst's judgment, the model's probability score, and the market's price. Each looks like knowledge. They are not the same kind of claim, and conflating them is an error with eight-figure consequences.

Joel A. Adejola

KEY FINDING

> The football recruitment market has developed sophisticated tools for generating confidence — AI scouting models, prediction markets, crowdsourced valuations — but has not developed commensurate tools for interrogating that confidence. A scout's conviction, a model's probability score, and a market price all produce numbers that look like knowledge. They differ in epistemic status: in what they are conditioned on, what they exclude, and what would need to be true for them to be reliable. When these layers compound — models training on flawed measurement data, markets pricing in model outputs, agents anchoring negotiations to crowdsourced estimates — the original epistemic weakness propagates through the entire decision stack. We propose a framework for `bounding recruitment certainty` that requires each layer to disclose the warrant behind its claims.

01

The Hypothetical
Four data points, four kinds of claim

A striker is through on goal. They have beaten the offside trap, have no defenders in pursuit, and only the goalkeeper stands between them and the back of the net. They choose a finish. They miss. It is not the first time. A season of squandered one-on-ones is documented across social media and internal scouting databases alike.

You are the recruitment executive of a Champions League club. Your analytics department delivers a dossier: a super-cut of every clear one-on-one this striker has faced across the last three seasons. Your AI scouting platform delivers a probability score: `0.72` Premier League readiness. On Transfermarkt, the player is valued at `€35 million`. A prediction market contract on the player signing for a top-five league club by August trades at `$0.68`.

You now have four data points: the video, the model, the crowd, and the market. Each produces a number. Each number looks like an answer. But the dossier forces a series of questions that the numbers do not answer on their own:

What can we infer about this player's finishing ability? How strongly can we project their performance into a different league, tactical context, or coaching environment? How certain can we be about any of the above — and what kind of certainty are we claiming? What would increase or decrease that certainty? And which of these four signals should you trust, to what degree, for what purpose?

> This publication explores those questions and, more broadly, the problem of epistemic certainty in football recruitment — what we know, what we merely believe, and how the gap between the two has widened even as the tools for generating confidence have proliferated.

02

Two Kinds of Certainty

Epistemic warrant versus psychological conviction

In analytic epistemology, **epistemic certainty** describes the strength of justification for believing a proposition is true. It is a function of evidence, logical coherence, methodological soundness, and the absence of defeaters — reasons that would undermine the belief if they were known. A recruitment assessment that is epistemically well-warranted rests on evidence that is relevant, sufficient, and interpreted within a sound analytical framework.

Psychological certainty is a feeling of conviction, which may or may not track epistemic warrant. A coach convinced a striker will "come good" might possess high psychological certainty with low epistemic warrant if the data say otherwise. Conversely, a data analyst might report low confidence in a projection while the underlying evidence is, in fact, robust — the analyst is simply risk-averse in temperament.

The distinction matters because the modern recruitment environment has produced tools that amplify psychological certainty without necessarily improving epistemic warrant. A model that outputs `0.72` Premier League readiness feels more precise than a scout saying "I think he can play in England." The number creates an impression of rigour. But precision and warrant are different properties: a number can be precise and poorly warranted, vague and well-warranted, or any combination. The epistemic question is not "how confident is the system?" but "what would need to be true for that confidence to be justified?"

03

Three Layers of Confidence

Analyst, model, and market produce different kinds of claim

Recruitment decisions in 2026 are informed by three distinct categories of signal, each with its own epistemic character. We assess that the failure to distinguish between them — to treat their outputs as interchangeable — is a structural source of decision error in professional football.

LAYER	SOURCE	OUTPUT FORMAT	EPISTEMIC CHARACTER
Analyst	Scout, video analyst, coaching staff	Qualitative judgment, verbal assessment	Domain expertise applied to direct observation. Warrant bounded by sample, context, and cognitive bias.
Model	AI/ML scouting platform, statistical model	Probability score, composite rating	Pattern recognition across large datasets. Warrant bounded by training data, feature selection, and measurement infrastructure.
Market	Prediction market, Transfermarkt, bookmaker	Price, crowd-sourced valuation	Aggregated belief with financial skin in the game. Warrant bounded by participants' information access and the accuracy of underlying data.

A scout's **70%** conviction, a model's `0.72` probability score, and a prediction market's **$0.68** contract price are not the same kind of claim. They differ in what they are conditioned on, what kinds of evidence they can incorporate, what kinds of evidence they structurally exclude, and what would need to change for them to be wrong. A recruitment executive who treats them as interchangeable — who averages them, or defers to whichever produces the highest number — is committing a category error.

04

The Analyst's Confidence
What direct observation can and cannot tell you

Return to the striker. The dossier arrives: three seasons of one-on-one situations, tagged and annotated. The analyst's task is to extract a judgment about finishing ability that can be projected into a new environment.

Two analytical approaches are available. The first examines *observable execution* — the mechanics of the chosen technique, the biomechanical efficiency of the approach, the quality of contact, the ball trajectory. This is measurable, repeatable, and largely verifiable from video. The second examines *underlying decision architecture* — why the player chose that particular finish in that particular moment. This accesses intent, perceptual-cognitive processing, and situational awareness. It is useful for coaching and player development but relies on mental states that are not directly observable.

For forecasting purposes — where the question is "will this player perform in a different environment?" — the mechanically-grounded approach produces higher epistemic warrant. Observable execution provides evidence about physical capabilities that transfer across tactical systems. Decision architecture is context-sensitive: a player's shot selection depends on coaching instruction, team shape, league tempo, and psychological state, all of which change with a transfer. The scout who reports "his finishing mechanics are sound but his shot selection is poor under pressure" has separated a system-independent trait (mechanics) from a system-dependent output (decision-making under pressure). The scout who reports "he's a bad finisher" has conflated the two — and the conflation will propagate into every downstream model that ingests the assessment.

The analyst's confidence is bounded by several factors that are well-understood in principle but inconsistently managed in practice: sample size (how many one-on-ones?), context dependency (were these chances of equivalent difficulty?), selection bias (does the super-cut include the chances he *didn't* receive because defenders adjusted to his movement?), and the analyst's own cognitive biases — recency effects, narrative construction, anchoring to reputation.

There is a further complexity within the analyst layer that the three-layer model deliberately simplifies but should not obscure. The scout and the coach are different epistemic actors. The scout evaluates ability: "can this player finish?" The coach evaluates fit: "does this player solve *my* problem, in *my* system, alongside *my* existing squad?" Fit is the most context-dependent judgment in recruitment — it depends on tactical philosophy, squad composition, dressing-room dynamics, and the coach's own capacity to develop the player's weaknesses. It is also the judgment that ultimately matters most, because the coach decides whether the player plays. In practice, the analytics pipeline has often been built *around* the coach rather than *with* them: dashboards present multidimensional profiles across dozens of metrics when the coach's question may be narrower and more precise — "can he carry the ball under pressure in the half-space?" The analyst has answered a question the coach did not ask, in a format the coach cannot easily translate into selection decisions. When analytical complexity exceeds the coach's question, it does not enhance the decision. It obscures it. The epistemic discipline required here is not simplification for its own sake. It is the discipline of matching the analytical output to the decision it is supposed to inform.

A NOTE ON MECHANICAL EVIDENCE

STATSWING treats mechanics — the physical execution underlying technique — as the most reliable predictor of skill transfer between playing environments. Mechanical advantages are established before the primary action occurs, in what we term the *pre-contact phase*: positioning before the ball arrives, momentum manipulation through body contact, the spatial negotiation between competing players. A player who consistently establishes superior body position before aerial duels demonstrates a transferable skill that operates regardless of tactical system. A player whose numbers depend on service quality and system-specific delivery patterns presents a riskier recruitment proposition. This principle, articulated in our analysis of contested aerial opportunities [1], applies equally to the 1v1 finishing scenario: the striker's approach angle, body shape at the moment of contact, and balance through the shooting motion are mechanical properties. The choice to shoot early versus dribble past the goalkeeper is a decision property. The former transfers; the latter may not.

05

The Model's Confidence

What a probability score epistemically warrants

The analytics department returns a second opinion: the AI scouting platform rates the striker at `0.72` Premier League readiness. The number is derived from a machine learning model trained on historical data — past players who transferred between comparable leagues, their pre-transfer statistical profiles, and their post-transfer performance. The model has identified features that correlate with successful adaptation: progressive carrying distance, pressing intensity, aerial win rate, shot-on-target percentage, and dozens of other metrics, weighted by the algorithm.

The output looks authoritative. It is a probability, expressed to two decimal places, produced by a system that has processed more data than any human scout could review in a career. Clubs across Europe are deploying comparable systems — from SkillCorner's AI-powered physical analytics to Sevilla's Scout Advisor built on IBM's watsonx platform, from Comparisonator's league-adjusted comparison engine to StatsBomb's composite models. The infrastructure is real, the investment is substantial, and the results, in aggregate, are useful.

But the epistemic status of `0.72` is not what it appears to be. The number packages a dense web of assumptions into a clean output:

Training data completeness. The model learned from players who transferred in the past. If the measurement infrastructure that produced their statistics carried systematic blind spots — as our analysis of the aerial duel metric demonstrated [1] — then the model inherited those blind spots. It does not know what was not measured. It cannot weight a variable that does not exist in its feature set.

League-comparability adjustment. Projecting performance from one league to another requires assumptions about how statistical outputs translate across competitive environments. A player's pressing numbers in a league with a slower average tempo may overstate or understate their likely output in the Premier League, depending on how the model handles context. The adjustment is necessarily approximate — and the model's confidence score does not disclose the uncertainty introduced by this approximation.

Feature selection. The model's designers chose which variables to include. Variables that are difficult to quantify — tactical intelligence, dressing-room leadership, adaptability to coaching instruction — are typically excluded. The model optimises on what it can measure, not on what matters. This is the Description Trap applied to computation: the output describes a statistical pattern, but the pattern is not insight. The model's confidence is confidence about the data it was given, not confidence about the player.

Automation Bias and the Rubber Stamp

The risk is not that models are useless. They are not. The risk is that their outputs are consumed in ways that exceed their epistemic warrant.

Research on human–AI collaboration across high-stakes industries has converged on a consistent finding: decision support systems that are correct **80–90%** of the time still introduce new errors, because users reverse correct judgments to match incorrect machine output [2]. This phenomenon — **automation bias** — was first studied in aviation, where pilots deferred to faulty autopilot recommendations despite contradictory instrument readings. It has since been replicated in medicine, military command systems, personnel selection, and process control [3][4]. The pattern is domain-independent: when a system produces a confident recommendation in a format that looks authoritative, humans tend to defer to it even when their own judgment is better.

Two aviation accidents illustrate the failure mode. In 2013, a Boeing 777 crashed while landing at San Francisco International Airport because the flight crew believed the autothrottle would maintain safe airspeed — but they had inadvertently deactivated it. In 2009, an Air France Airbus A330 crashed into the Atlantic after ice crystals disabled airspeed sensors and the autopilot disconnected; the pilots, suddenly handed manual control, failed to recognise a stall condition because their mental model of the automated system was wrong [5]. In both cases, "human oversight" existed on paper. The pilots were in the loop. But being in the loop is not the same as being at the helm.

A recent clinical AI framework distinguishes these precisely [6]: **oversight** means reviewing the output — a check-box exercise at the end of a workflow. **Agency** means controlling the question, setting the constraints, verifying the evidence, and making the final decision. The football translation is direct: if a sporting director does not understand what a model's confidence score is conditioned on — where the training data has gaps, which variables were excluded, how the league adjustment works — then reviewing the model's shortlist is oversight, not agency. And the errors that escape oversight are, by definition, the errors that matter most.

> The better a system performs on average, the more dangerous the exceptions become. High accuracy breeds high trust, which breeds low vigilance, which means the errors that do occur are the ones that go undetected longest.

What Models Should and Should Not Do

The question is not whether to use AI in recruitment. The infrastructure exists, it is improving, and it offers genuine advantages in scope and speed. The question is how to situate models within a decision architecture that preserves epistemic discipline.

We assess that AI models are well-suited to three categories of recruitment task. First, *filtering at scale*: reducing a universe of thousands of players to a shortlist of dozens based on statistical profiles. This is a task where breadth matters more than depth and where false negatives (missing a good player) are less costly than the time required to review every candidate manually. Second, *anomaly detection*: identifying statistical patterns that diverge from expectations — a player whose output exceeds what their underlying metrics would predict, or whose metrics suggest a role change that hasn't occurred yet. Third, *consistency checking*: flagging cases where a scout's assessment and the statistical profile diverge sharply, not to override the scout but to prompt a second look.

Models are poorly suited to — and should not be trusted for — questions that require judgment about how evidence transfers between contexts, assessment of traits that the measurement infrastructure does not capture, and final recommendations on whether to commit capital. These are human decisions. They are also the decisions that matter most.

This is not a conservative position. It is a structural observation about where the value of intelligence resides. The cost of building an analytical pipeline has fallen by an order of magnitude in the past two years. What has not fallen — and what cannot fall — is the cost of knowing what the pipeline's output epistemically warrants. Any competent operator can now build, in weeks, data infrastructure that would have required a team of engineers eighteen months ago. The scarce resource is not pipeline. It is framework — the capacity to ask the right questions, interpret the answers within appropriate bounds, and recognise the limits of what any tool can tell you.

06

The Market's Confidence

Prediction markets, crowdsourced valuations, and the reification problem

The prediction market prices the striker's transfer to a top-five league at `$0.68`. On Transfermarkt, the crowdsourced valuation is `€35 million`. Both numbers carry epistemic weight — but of a different kind.

Prediction Markets: Calibrated Ignorance

Prediction markets have grown explosively: global trading volume reached approximately `$64 billion` in 2025, a fourfold increase from the prior year, with more than `80%` driven by sports event contracts [7]. Polymarket and Kalshi, the two dominant platforms, are each pursuing valuations of approximately `$20 billion` [8]. The theoretical case for prediction markets is well-established: when participants put money behind their beliefs, the resulting prices aggregate dispersed information more efficiently than any individual forecaster. Tetlock's Good Judgment Project demonstrated that calibrated forecasters — and aggregated forecasts in particular — outperformed intelligence analysts with access to classified information [9].

The market's price is, in a meaningful sense, the best available summary of what the crowd believes. But it is a summary of *belief*, not a summary of *knowledge*. The market is well-calibrated *given the information available to participants*. When the underlying measurement infrastructure is flawed — when the standard metric undercounts the very contests it claims to measure [1] — the market prices in the flaw as if it were ground truth. The market does not know what it does not know. The $0.68 contract price reflects aggregate psychological certainty with financial discipline. It does not reflect an assessment of whether the data on which participants are forming beliefs accurately captures the player's ability.

Transfermarkt: The Accidental Anchor

Transfermarkt occupies a unique position in football's epistemic infrastructure. The platform's player valuations are produced by a network of volunteer community members — unpaid enthusiasts who debate and propose valuations on forums, mediated by elevated "judges" who make final determinations [10]. The process combines performance data, comparative reasoning, transfer rumour, market dynamics, and subjective assessment. Herm et al. (2014) describe the mechanism as a "judge principle" in which experienced community members filter collective input into discrete estimates [10].

Academic research confirms two things about these valuations. First, they are reasonably accurate in aggregate: forecasts of international match results derived from Transfermarkt squad values outperform FIFA rankings and Elo ratings [11]. Second, they are systematically biased: they underestimate high-value transfers and carry structural distortions by league and market segment [12].

The epistemically remarkable development is what has happened to these numbers *outside* the platform. An investigation by Follow the Money found that Transfermarkt valuations are referenced by agents in transfer negotiations, cited by sporting directors, and included in club financial reporting [13]. Former FC Barcelona president Josep Maria Bartomeu appeared on Catalan radio and defended the €72 million fee paid for Arthur Melo by citing the player's Transfermarkt value. Werder Bremen included Transfermarkt valuations in a bond emission prospectus — a legal financial document — to demonstrate to investors that the club's player assets were worth more than their carrying amounts [14]. Agents have reportedly attempted to lobby Transfermarkt volunteers for more favourable client valuations [13].

This is a case study in what epistemologists call **reification:** the process by which an approximate, crowd-derived estimate acquires the social authority of a fact. The epistemic status of a Transfermarkt value — approximate, biased, volunteer-produced, based partly on rumour and subjective assessment — is one thing. The social function of that number — anchor in a nine-figure negotiation, line item in a bond prospectus, benchmark in a boardroom — is something else entirely. The gap between what the number warrants and what it does is the gap this publication describes.

> A number's authority in a negotiation room does not depend on its epistemic warrant. It depends on whether everyone in the room treats it as a starting point. Transfermarkt's valuations have become anchors not because they are accurate, but because they are comprehensive, accessible, and uncontested.

07

The Compounding Problem

When epistemic gaps propagate through the decision stack

Each confidence layer — analyst, model, market — carries its own epistemic limitations. Examined in isolation, these limitations are well-understood and, in principle, manageable. The problem that has received insufficient attention is what happens when the layers compound.

THE COMPOUNDING STACK

LAYER 0
Measurement Infrastructure
Data providers define what gets counted. Definitions carry structural exclusions.
GAP

inherits gap

LAYER 1
Model / AI Scouting Platform
Trains on measured data. Cannot weight what was not recorded.
GAP

inherits gap

LAYER 2
Crowd Valuation / Prediction Market
Aggregates belief. Calibrated to available information — including flawed metrics.
GAP

inherits gap

DECISION LAYER
Recruitment Executive
Receives composite confidence. Each layer looks rigorous. The foundation may be sand.
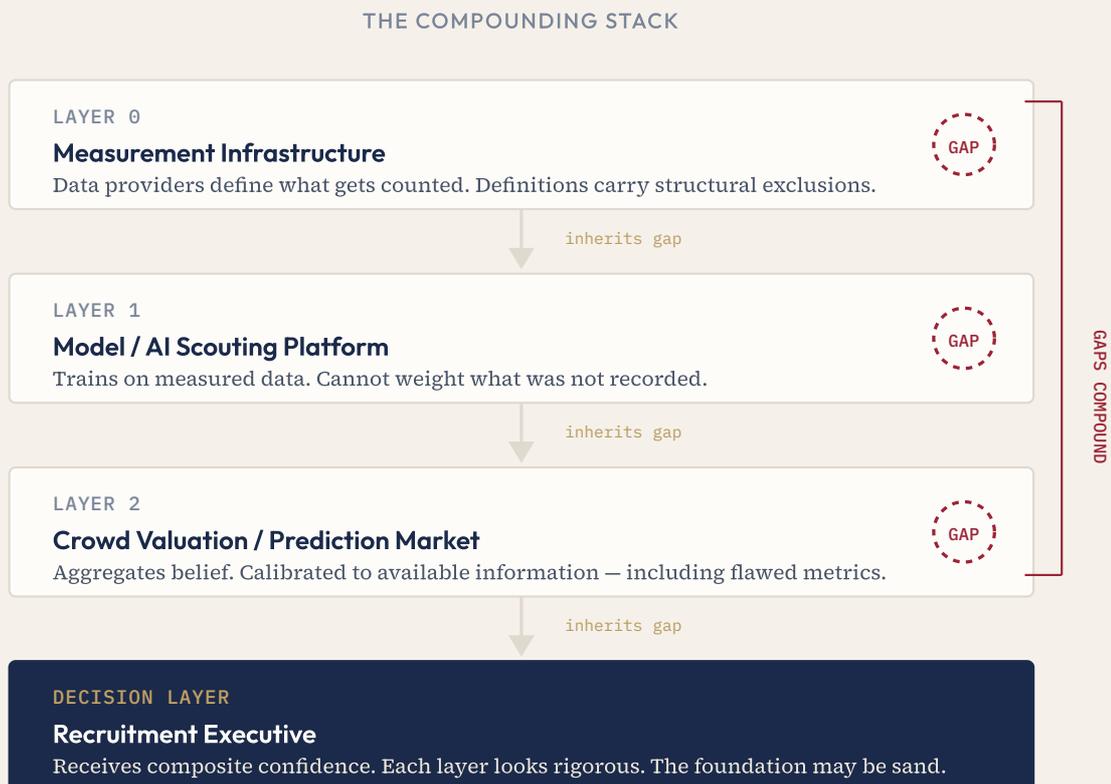
GAPS COMPOUND

Fig. 1. The compounding stack: epistemic gaps at the measurement layer propagate through models, markets, and decisions. Each subsequent layer inherits the blind spots of the layer below and adds its own. The decision-maker at the top receives a composite signal that looks increasingly authoritative — but the foundational measurement gap remains unaddressed at every stage.

Consider the sequence. A data provider defines what constitutes an "aerial duel" — a definition that, as we have previously demonstrated, excludes a category of consequential aerial competition [1]. A scouting model trains on this data and learns to weight aerial ability using the provider's definition. The model produces a composite score that enters a crowdsourced discussion on Transfermarkt,

where community members incorporate statistical profiles into their valuation debates. An agent cites the Transfermarkt number in a contract negotiation. A prediction market prices the transfer outcome using all publicly available information — including the model output and the Transfermarkt valuation.

At each stage, the signal looks more authoritative. At no stage does the original measurement gap get corrected. It is not that anyone in the chain is negligent. It is that the architecture of the chain does not include a mechanism for propagating uncertainty. Each layer consumes the output of the layer below as input, not as a claim with stated limitations.

This is the epistemological analogue of a problem well-known in software engineering: *garbage in, garbage out*, except that the "garbage" is not random noise but systematic bias — and systematic bias is harder to detect than noise because it produces outputs that are internally consistent and plausible.

08

Bounding Recruitment Certainty

A framework for epistemic disclosure in recruitment decisions

The solution is not to abandon any of these tools. Each produces genuine value within its domain. The solution is to develop a practice of **epistemic disclosure** — a discipline of stating, alongside every assessment, what the assessment is conditioned on and where the warrant runs out.

We propose five questions that should accompany any recruitment recommendation, whether produced by a human analyst, an AI model, or derived from market signals:

#	QUESTION	WHAT IT EXPOSES
1	What is this claim conditioned on?	The data inputs, the analytical framework, the assumptions required for the conclusion to hold.
2	What would change the assessment?	The sensitivity of the conclusion to specific variables. A robust assessment survives perturbation; a fragile one collapses under small changes.
3	What can this tool not see?	The structural exclusions — traits not captured by the data, contexts not represented in the training set, dynamics not priced by the market.
4	Where does this layer inherit from another?	The dependency chain. A model trained on aerial duel data inherits the aerial duel definition's exclusions. A market price conditioned on model outputs inherits the model's blind spots.
5	What is the appropriate confidence for this decision?	Not what the tool outputs, but what the decision-maker should hold — given the answers to questions 1 through 4.

These questions do not produce a number. They produce a *bounded assessment* — a judgment about a player's ability that explicitly states its warrant and its limits. This is harder to produce than a composite score. It is also harder to misuse.

A Confidence Calibration Framework

STATSWING's intelligence assessments employ a calibrated vocabulary for expressing confidence, drawn from the practice of analytic epistemology and adapted for recruitment contexts:

CONFIDENCE LEVEL	LANGUAGE	EPISTEMIC BASIS
High	"We assess that..." / "consistently demonstrates"	Multi-season mechanical evidence, system-independent traits confirmed across contexts, corroborating signals from independent sources.
Medium	"Evidence suggests..." / "tends to"	Statistical pattern supported by limited mechanical evidence, or strong mechanical evidence from a narrow sample.
Low	"Appears to..." / preliminary indicators	Single-context observation, model output without independent corroboration, or crowd-derived signal with known biases.

The critical discipline is that the confidence level is a property of the *evidence*, not a property of the *assessor's feeling*. An analyst who has watched extensive video may feel highly confident while possessing only medium-warrant evidence (because the video is drawn from a single tactical system). A model may output a high probability score while the underlying warrant is low (because the training data carried systematic exclusions). Calibrating confidence to evidence rather than to feeling is what separates intelligence assessment from opinion.

The Framework Shift

We assess that the competitive advantage in recruitment intelligence is migrating from *pipeline* to *framework*. The cost of assembling, transforming, and modelling player data has fallen dramatically — accelerated by AI coding tools that compress what previously required dedicated engineering teams into infrastructure a single competent operator can build in days. Every analytical provider now buys substantially similar data feeds. The data is increasingly a commodity.

What is not a commodity — and what cannot be commoditised by current technology — is the epistemic framework applied to the data: the capacity to determine what questions are worth asking, what kind of knowledge each data point represents, where the market's analytical conventions produce systematically wrong assessments, and how to bound the uncertainty around a recommendation so that the decision-maker knows what they are buying.

This is not an argument against technology. It is an argument about where technology's value terminates and where human judgment — specifically, epistemically disciplined judgment — begins. The institution that understands what its pipeline's output warrants will make better decisions than the institution that builds a better pipeline.

The pattern is visible in the recent history of the field itself. The most consequential findings in football analytics have not been generated by more advanced models. They have been generated by more precise questions — revised definitions, reframed measurements, foundational interrogation of

what the field takes for granted. Tracking when a player is substituted by game state requires no machine learning; it requires a willingness to frame the question. Identifying that a standard metric excludes a category of consequential competition requires no novel statistical method; it requires someone to ask whether the definition captures what it claims to capture [1]. The data in both cases is simple. The insight is in the framing. The football analytics community has invested heavily in methodological sophistication — more advanced models, more complex visualisations, more granular data. This investment is not wasted. But it has produced a structural bias: the field reaches for computational power before it has verified that the question is right, the definition is sound, and the measurement captures what matters. More sophisticated analysis of a flawed metric does not produce better insight. It produces more confident error.

METHODOLOGICAL NOTE

This publication synthesises findings from epistemology, decision science, human–AI interaction research, and football analytics. The three-layer model (analyst, model, market) is an analytical framework proposed by STATSWING, not an empirical finding. The five-question epistemic disclosure protocol has been applied in STATSWING's intelligence assessments since the institution's operational launch but has not been tested at scale across multiple organisations. The automation bias findings are drawn from peer-reviewed systematic reviews across aviation, medicine, and related domains; their application to football recruitment is by analogy, not by direct study. We note this limitation explicitly: the claim that automation bias operates in football scouting is a well-grounded inference, not an empirical demonstration. Empirical research on this question in professional football recruitment departments is, to our knowledge, absent from the published literature — which is itself an indication of how little the epistemic foundations of recruitment decision-making have been interrogated.

DATA SOURCES AND REPRODUCIBILITY

The prediction market data (trading volumes, platform valuations) are drawn from publicly reported figures by Kalshi, Polymarket, and industry reporting by Covers, Front Office Sports, and PYMNTS as of March 2026 [7][8]. Transfermarkt's valuation methodology is documented in academic studies by Herm et al. (2014), Peeters (2018), and Coates & Parshakov (2022) [10][11][12]. The automation bias literature is anchored in the systematic review by Goddard et al. (2012) and subsequent work in human–AI interaction [2][3][4]. Tetlock's forecasting research is reported in Tetlock & Gardner (2015) and the Good Judgment Project publications [9]. The aviation case studies are drawn from NTSB and BEA investigation reports [5]. The Transfermarkt reification examples are drawn from investigative journalism by Follow the Money (2020) and academic analysis in Emerald Insight (2024) [13][14]. The three-layer confidence model, the five-question epistemic disclosure framework, and the Crawling Plants confidence calibration vocabulary are proprietary analytical contributions of STATSWING.

EPISTEMIC CERTAINTY   RECRUITMENT INTELLIGENCE   AUTOMATION BIAS   PREDICTION MARKETS   DECISION ARCHITECTURE

SCOUTING METHODOLOGY

IMPLICATIONS BY AUDIENCE

For recruitment departments

When a model produces a composite score and a scout produces a qualitative assessment, and the two disagree, the question is not "which is right?" but "what is each conditioned on?" If the model's score is conditioned on a measurement infrastructure with known gaps, and the scout's assessment is based on direct observation of traits the model cannot see, the scout may hold higher epistemic warrant on the specific question at hand — even though the model is more reliable in aggregate. Develop protocols for adjudicating between layers that are based on warrant, not authority. A number is not more trustworthy than a judgment simply because it is a number.

For analytics providers and platform builders

Confidence scores are consumed by decision-makers as though they are well-calibrated probabilities. In most cases they are not — they are model outputs that reflect the training data's coverage and the algorithm's assumptions, not the ground-truth probability of success. Consider publishing model cards alongside scouting products: documentation of what the model was trained on, what it excludes, how the league-comparability adjustment works, and where the output's uncertainty is highest. This is standard practice in responsible AI deployment in medicine and finance. Football analytics has not yet adopted it.

For agents and intermediaries

Transfermarkt valuations are useful as rough benchmarks but carry documented biases — they systematically underestimate high-value players and overestimate players in certain market segments. Anchoring negotiations to a Transfermarkt figure as though it were an objective appraisal treats a crowdsourced estimate as a market fact. Agents who understand the epistemic status of these numbers — and who can articulate what an independent assessment would look like — hold a structural advantage in negotiations over those who cite the platform uncritically.

For researchers and academics

The football analytics literature has produced substantial work on refining player valuation models, improving predictive accuracy, and applying machine learning to scouting. It has produced comparatively little on the epistemic foundations of recruitment decision-making: what kinds of knowledge different data sources produce, how confidence should be calibrated across analytical layers, and how systematic measurement biases propagate through composite models. This gap mirrors a pattern Tetlock identified in political forecasting: the field invested heavily in making predictions and very little in evaluating whether those predictions were well-warranted. Decision science, analytic epistemology, and the emerging literature on human–AI interaction offer frameworks that the football analytics community has largely not engaged with.

REFERENCES

[1] J. Adejola, "Contested Aerial Opportunities: The Measurement Gap at the Centre of Football's Set-Piece Revolution,"
     STATSWING Research SW-R-2026-001, March 2026. statswing.com/research/aerials

[2] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: a systematic review of frequency, effect mediators, and
     mitigators," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 121–127, 2012.

[3] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: an attentional integration,"
     Human Factors, vol. 52, no. 3, pp. 381–410, 2010.

[4] G. Romeo and D. Conti, "Exploring automation bias in human–AI collaboration: a review and implications for
     explainable AI," *AI & Society*, Springer, 2025. springer.com

[5] Bureau d'Enquêtes et d'Analyses (BEA), "Final Report on the accident on 1st June 2009 to the Airbus A330-203,
     flight AF 447," July 2012. National Transportation Safety Board (NTSB), "Descent Below Visual Glidepath and Impact
     With Seawall, Asiana Airlines Flight 214," NTSB/AAR-14/01, June 2014.

[6] iatroX Clinical AI Insights, "From 'human-in-the-loop' to 'human-at-the-helm': how to prevent automation bias and
     preserve clinical agency in the AI era," December 2025. iatrox.com

[7] Next.Io via Covers, "Prediction Market Volume Quadrupled in Past 2 Years," March 2026. covers.com

[8] Wall Street Journal via PYMNTS, "Polymarket and Kalshi Each Targeting $20 Billion Valuations," March 2026.
     pymnts.com

[9] P. E. Tetlock and D. Gardner, *Superforecasting: The Art and Science of Prediction*, Crown, 2015. The Good Judgment
     Project; see also P. Atanasov, P. Rescober, E. Stone, et al., "Distilling the Wisdom of Crowds," *Management Science*,
     2017.

[10] S. Herm, H.-M. Callsen-Bracker, and H. Kreis, "When the crowd evaluates soccer players' market values: Accuracy
     and evaluation attributes of an online community," *Sport Management Review*, vol. 17, no. 4, pp. 484–492, 2014.

[11] T. Peeters, "Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer
     results," *International Journal of Forecasting*, vol. 34, no. 1, pp. 17–29, 2018.

[12] D. Coates and P. Parshakov, "The wisdom of crowds and transfer market values," *European Journal of Operational
     Research*, vol. 301, no. 2, pp. 523–534, 2022.

[13] T. Houtman and J. Wester, "How the volunteers of data website Transfermarkt became influential players at European
     top football clubs," *Follow the Money*, December 2020. ftm.eu

[14] D. Detzen and L. Lohlein, "Valuing soccer players: on the valuation dynamics of an online user community,"
     Accounting, Auditing & Accountability Journal, vol. 37, no. 2, pp. 608–626, 2024. emerald.com

CITE THIS PUBLICATION

> J. Adejola, "Epistemic Certainty in Recruitment: The Gap Between What Football Knows and What It Claims to
> Know," STATSWING Research SW-R-2026-002, March 2026. statswing.com/research/sw-r-2026-002

STATSWING

STATSWING Research publications share non-proprietary
findings from our analytical work. They represent a portion of
the frameworks and methodology that underpin our
intelligence products.

For the full analytical infrastructure — including proprietary
ratings, player and team assessments, and decision-support
briefs — visit statswing.com.

SW-R-2026-002