



Maturation, Bias and the Training Set

Developmental Fairness in AI-Supported Athlete Development

About this paper

This paper asks what fairness should mean when AI tools score, rank or profile young athletes on a development pathway. It is written for heads of talent, academy and performance directors, governing body staff who approve or buy such tools, and the coaches and parents who live with their outputs. Its concern is practical. A pathway that adopts a scoring tool inherits whatever the tool's training data did or did not contain, and in youth sport the people the data leaves out are often the people the pathway most needs to keep.

The paper is the fourth in the Athleet.AI series on AI in coaching and athlete development, and it carries the series argument that AI in coaching and athlete development must respect the non-linear, human and developmental nature of athletic progress. Earlier papers dealt with prediction and with the coach's place in the loop. This one joins two bodies of evidence that have mostly been read apart, the general literature on algorithmic bias and sport science's own record on relative age, maturation and the under-representation of women, and asks what follows when a model is trained on the second and judged by the standards of the first.

Editorial balance. Roughly half of what follows reports published evidence, about a third is analysis that connects that evidence to pathway decisions, and the remainder is my own framework of four fairness dimensions and a pre-deployment test, which is proposed for use and debate and has not been validated.

Legal and regulatory notice. Parts of this paper describe legislation, regulation, regulator activity or court decisions. 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 I draw a practical implication that a source does not state, the text says so.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim carries a tag stating what kind of evidence supports it. Two tags are combined where a claim rests on both.

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

REVIEW A systematic review, meta-analysis or narrative review.

CONSENSUS A governing body position, consensus statement or official guidance.

PREPRINT A working paper or preprint not yet peer reviewed.

INDUSTRY A company announcement, press report or industry source.

SCHOLARLY A scholarly book, edited volume or law journal article.

PRACTICE Practitioner opinion, including the author's own.

EVIDENCE GAP The available evidence does not answer the question.

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

CONTENTS

Contents

FRONT MATTER

A note from the author	5
Executive summary	6

PART ONE · THE INHERITED PROBLEM

1 Two Athletes, One Score	9
2 What General AI Has Already Learned	11

PART TWO · SPORT'S OWN DISTORTIONS

3 Sport's Existing Data Gaps	14
4 Maturation and Relative Age as Structural Bias	16

PART THREE · DESIGNING FOR DEVELOPMENTAL FAIRNESS

5 Four Dimensions of Fairness for Pathway AI	20
6 Testing Before Deployment	22

END MATTER

References	26
About the author	27

ILLUSTRATIVE MATERIAL

Every organisation, person and scenario described as illustrative in this paper is invented. Published research and official documents are cited and numbered, and invented material is never presented as evidence.

Research closed 22 May 2023. Sources and events are as they stood at that date.

CITATION

Forrest, P. (2023) *Maturation, Bias and the Training Set: Developmental Fairness in AI-Supported Athlete Development*. Athleet.AI Working Papers, AI in Coaching and Athlete Development, Paper 4. Athleet.AI, 24 May 2023. Available at: athleet.ai.

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

PAPERS IN THIS STRAND

- Paper one. *Predicting the Unpredictable*. Why Algorithmic Talent Identification Struggles With Non-Linear Athlete Development. 25 January 2023.
- Paper two. *The Coach in the Loop*. Tacit Knowledge, Professional Judgement and the Risk of Deskilled Coaching. 27 February 2023.
- Paper three. *The Measured Child*. Datafication, Consent and Autonomy Across the Athlete Pathway. 19 April 2023.
- **Paper four. *Maturation, Bias and the Training Set: Developmental Fairness in AI-Supported Athlete Development*. 24 May 2023.**
- Paper five. *Load, Readiness and False Precision*. The Limits of Algorithmic Training Prescription in Long-Term Endurance Development. 21 June 2023.
- Paper six. *Democratised or Stratified?* AI, Competitive Fairness and the

Grassroots Coaching Gap. 26 July 2023.

A NOTE FROM THE AUTHOR

The late developer I nearly overlooked

Most of what I believe about talent I learned by getting it wrong in small ways. Early in my coaching I ranked young athletes by what they could do that month, and I did it with a stopwatch and a clipboard, which felt objective. Some of the athletes I rated lowest were simply younger in their year or further from their growth spurt, and a few of them were outperforming my early favourites within a couple of seasons. Nothing in my numbers was false. They answered a narrower question than the one I thought I was asking.

That memory is why the current enthusiasm for scoring tools in youth sport makes me uneasy and hopeful at once. A well-built model could hold the context that a tired coach forgets on a wet Tuesday, including birth month, estimated maturity, training age and injury history, and could stop a selection panel from confusing size with promise. A badly built one would do the opposite.

In my advisory work I increasingly meet sports organisations that are asked to approve tools whose training data they have never seen. The questions they are encouraged to ask are about accuracy and security, which are fair questions, but I rarely hear anyone ask who was in the sample, how old they were relative to their cut-off, or whether any of them were girls. Those are the questions this paper tries to make routine.

I have written it for the people who sit on selection panels and procurement groups, and for the parent who is told that a number has closed a door. The recommendations at the end are mine. They are deliberately modest, because I would sooner see four tests run properly than a long charter that nobody applies.

Paul Forrest
Founder, Athleet.AI
24 May 2023

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

AI tools that score young athletes can be accurate on average and still unfair to the relatively young, the late maturing, girls and para athletes, because in a developing population the same score means different things for different children; pathways should test tools against those groups before use, and give families a real route to challenge what the tools say. General AI research has shown how aggregate accuracy conceals subgroup failure and how convenient proxies carry historic inequality into automated decisions. Sport science has separately documented the relative age effect, wide variation in biological maturity among children of the same age, and a persistent shortage of female participants in research. The two literatures are rarely read together. That matters, because pathway AI is trained on exactly the data those sport science findings describe, and so inherits the biases the second literature records while escaping the scrutiny the first would apply.

1

Aggregate accuracy can hide large failures in particular groups.

An audit of three commercial gender classification systems found error rates of up to 34.7% for darker-skinned women against a maximum of 0.8% for lighter-skinned men.¹ **PEER-REVIEWED** A youth sport model reported only on its overall accuracy could conceal gaps of the same kind by sex, birth quartile or maturity status. **EVIDENCE GAP**

3

Sport's own data under-represent women.

A study of 5,261 publications in six leading journals found that 6% studied female participants only, and that women made up 34% of all participants.³ **PEER-REVIEWED** Models built on sport science evidence start from that imbalance.

5

Developmental fairness is the dimension general AI ethics does not cover.

General fairness practice asks how a model performs for different groups and reports the gaps.⁶ **PEER-REVIEWED** It does not ask whether a prediction made at 12 prematurely narrows the options of someone who has not finished growing, and that is the question a pathway must answer. **ATHLEET.AI**

2

Proxies inherit the inequality built into the thing they stand in for.

A widely used health algorithm predicted costs as a stand-in for illness, and correcting the resulting bias would have raised the share of Black patients flagged for extra help from 17.7% to 46.5%.² **PEER-REVIEWED** Current performance used as a stand-in for potential carries the same risk in youth sport.

4

Relative age and maturation are structural biases, and models trained on selection data will learn them.

A meta-analysis of 253 samples found relative age effects to be consistent though small, and strongest in representative-level adolescent male sport.⁴ **REVIEW** Children of the same age can differ widely in biological maturity,⁵ **SCHOLARLY** so selection history is a record of maturity as much as of talent.

6

Contestability is on the policy agenda, and pathways can build it now.

The UK white paper on AI regulation lists fairness and contestability and redress among five cross-sector principles.⁷ **CONSENSUS** A pathway can give a family a clear route to question a tool-supported decision without waiting for legislation. **ATHLEET.AI**

What follows for a club or governing body

Before any scoring or profiling tool is used on young athletes, the organisation should ask for evaluation results broken down by sex, relative age quartile and maturity status, with the sample size behind each figure, and should decline to use the tool for selection where those figures are missing or rest on too few athletes to mean anything. It should treat maturity-adjusted outputs as the default for anyone in or near the growth spurt, record where a tool influenced a decision, and publish to families a short, plain route to challenge that decision with a named person who can overturn it.

All of it can be done with existing technology, provided the procurement conversation includes the athletes the pathway is most likely to lose and the organisation is willing to keep a tool out of selection until it has shown that it treats them fairly.



PART ONE

The inherited problem

Why fairness in pathway AI begins with biases that general AI research and sport science have each documented on their own.

CHAPTER ONE

Two Athletes, One Score

A model can give two young athletes the same number and still treat them unequally, because in a growing population the same number carries different meanings.

Talent pathways have always ranked children, and they have always done so with incomplete information. What changes when a model enters the process is the scale and the apparent authority of the ranking. A coach's hunch can be questioned over a cup of tea; a score on a dashboard arrives looking finished. The purpose of this chapter is to show, with an invented but ordinary case, why the standard questions asked of AI tools (is it accurate, and is the data secure) leave out the question that matters most on a development pathway.

ILLUSTRATIVE SCENARIO**Same score, different futures**

In an invented regional girls' pathway, two 13-year-olds sit on adjacent lines of a selection dashboard. Both carry a composite "potential" score of 71, produced by a purchased tool that blends sprint times, a jump test, coach ratings and match statistics. The first was born in the opening weeks of the selection year and began her growth spurt early; she is taller and heavier than most of her squad and has been among the quickest players at every age group. The second was born ten months later, has not yet started her spurt, and reaches the same score through first touch and decision-making that her coach rates highly. The tool was originally developed on data from a boys' programme and extended to girls with a smaller sample. The panel has fewer places than candidates, and with nothing to separate the two it chooses the athlete who looks further ahead. Her club responds to the decision by moving the second girl to a lower training group. No figure on the dashboard is wrong, and nobody on the panel acted in bad faith.

1.1 What the score cannot see

The two scores in the scenario are equal, and yet the second athlete has produced hers with fewer of the physical advantages that growth confers. If the score is a measure of present performance, the panel read it correctly. If it is meant as a measure of potential, which is how such tools tend to be sold and almost always how panels use them, then the second athlete's 71 is arguably the stronger signal, because it was earned against a physical handicap that will probably shrink. The tool, as described, has no way to say so.

That gap follows from how supervised models learn, and it would appear in most products built this way. A model trained to predict who was selected, retained or rated highly in past cohorts will learn whatever drove those outcomes, and the sport science reviewed in Part two suggests that birth month and maturity status drove a meaningful part of them.⁴ **REVIEW** The model then reports those learned associations

back as if they were properties of the athlete. My own reading is that this is the most likely route by which pathway AI would harden existing selection bias, since it needs nothing more than faithful learning from a biased history.

1.2 Why the harm is a closed door

General discussions of algorithmic harm often concern a wrong answer, such as a misclassified face or a mispriced loan. In a pathway the characteristic harm is a closed door. A deselected 13-year-old seldom gets a second assessment a month later, and she usually gets less coaching, weaker competition and, often, a story about herself that she did not have before. Hancock, Adler and Côté propose that parents, coaches and athletes themselves amplify early advantages, parents through Matthew effects, coaches through Pygmalion effects and athletes through Galatea effects, with coaches' beliefs shaping how they treat players and athletes' own beliefs shaping their effort.⁸ **PEER-REVIEWED** A score that carries institutional authority seems likely to strengthen each of those loops, and I would expect the effect to be largest where the score is least understood. **PRACTICE**

This is why I use the term developmental fairness for the concern at the centre of this paper. It asks whether a tool's output, at the moment it is used, unfairly narrows the future options of someone whose abilities are still forming. For adults the question seldom arises, so the general literature has had little reason to answer it. Pathways do not have that luxury, although (to be fair to them) they were not answering it especially well before any software arrived, and the relative age literature is in large part a record of that difficulty.

The timing matters. In May 2023 the UK Competition and Markets Authority opened an initial review of competition and consumer protection questions raised by AI foundation models,⁹ **CONSENSUS** and in the same month OpenAI's chief executive told a US Senate subcommittee that regulatory intervention by governments would be critical to managing the risks of increasingly powerful models.¹⁰ **INDUSTRY** Neither development is aimed at sport. Both signal that questions about how AI systems treat the people they affect are moving from the seminar room to the regulator's desk, and sports bodies that buy these tools would do well to have their own answers ready.

TRAINING SET

The training set is the historical data a model learns from. Whatever patterns it holds, including who was selected and why, become the model's idea of what a good athlete looks like.

CHAPTER TWO

What General AI Has Already Learned

Research on algorithmic bias outside sport has already shown the two mechanisms that matter most for pathways, namely hidden subgroup failure and biased proxies.

The literature on algorithmic fairness is large, and most of it concerns adults in settings far from sport. Three findings from it transfer almost directly to youth development, and each changes what a pathway should ask a vendor.

2.1 Accuracy that hides its failures

Buolamwini and Gebru examined two widely used facial analysis benchmarks and found that their images were overwhelmingly of lighter-skinned people, at 79.6% in one and 86.2% in the other,¹ [PEER-REVIEWED](#) so they built a new benchmark balanced by gender and skin type and used it to evaluate three commercial gender classification systems. Error rates for darker-skinned women reached 34.7%, while the highest error rate for lighter-skinned men was 0.8%.¹ [PEER-REVIEWED](#) Each system could have reported a respectable overall accuracy on an unbalanced test set, because the group it served worst was the group least present in the data.

The lesson for sport is structural. A talent tool validated on a pathway's historical cohort will be validated mostly on the athletes who were plentiful in that cohort, which in many sports means relatively older boys who matured early. Its headline accuracy then describes how well it works for them. Whether it works for a late-maturing girl born in the last quarter of the year is a separate question, and the headline figure is silent on it. I am not aware of any published audit of a commercial youth sport tool broken down in this way. [EVIDENCE GAP](#)

Average accuracy is not a fairness test. A single accuracy figure describes the athletes who dominate the test data. Ask for results by sex, relative age quartile and maturity status, and ask how many athletes sit behind each result, before accepting that a tool works for the whole pathway.

2.2 The proxy problem

Obermeyer and colleagues studied a commercial algorithm that health systems used to identify patients with complex needs for extra support.² [PEER-REVIEWED](#) At any given risk score, Black patients were considerably sicker than White patients. The cause was the target the algorithm predicted, which was future healthcare cost, used as a stand-in for illness; because less money was spent on the care of Black patients, cost understated their need. The authors estimated that correcting the disparity would

raise the share of Black patients receiving additional help from 17.7% to 46.5%, and they argued that convenient, apparently effective proxies can be an important source of bias in many settings.² **PEER-REVIEWED**

Youth sport is full of proxies. Past selection stands in for talent, current sprint speed for athletic potential, minutes played for coach confidence, and a coach rating for a dozen things the coach could not measure. Each proxy carries its own history. If selection panels have favoured early maturers, then a model that predicts selection will learn to favour them too, and it will do so while reporting good predictive performance, exactly as the health algorithm did. The target variable needs as much scrutiny as the algorithm, and in my experience it receives far less. **PRACTICE**

2.3 Fairness definitions that cannot all hold at once

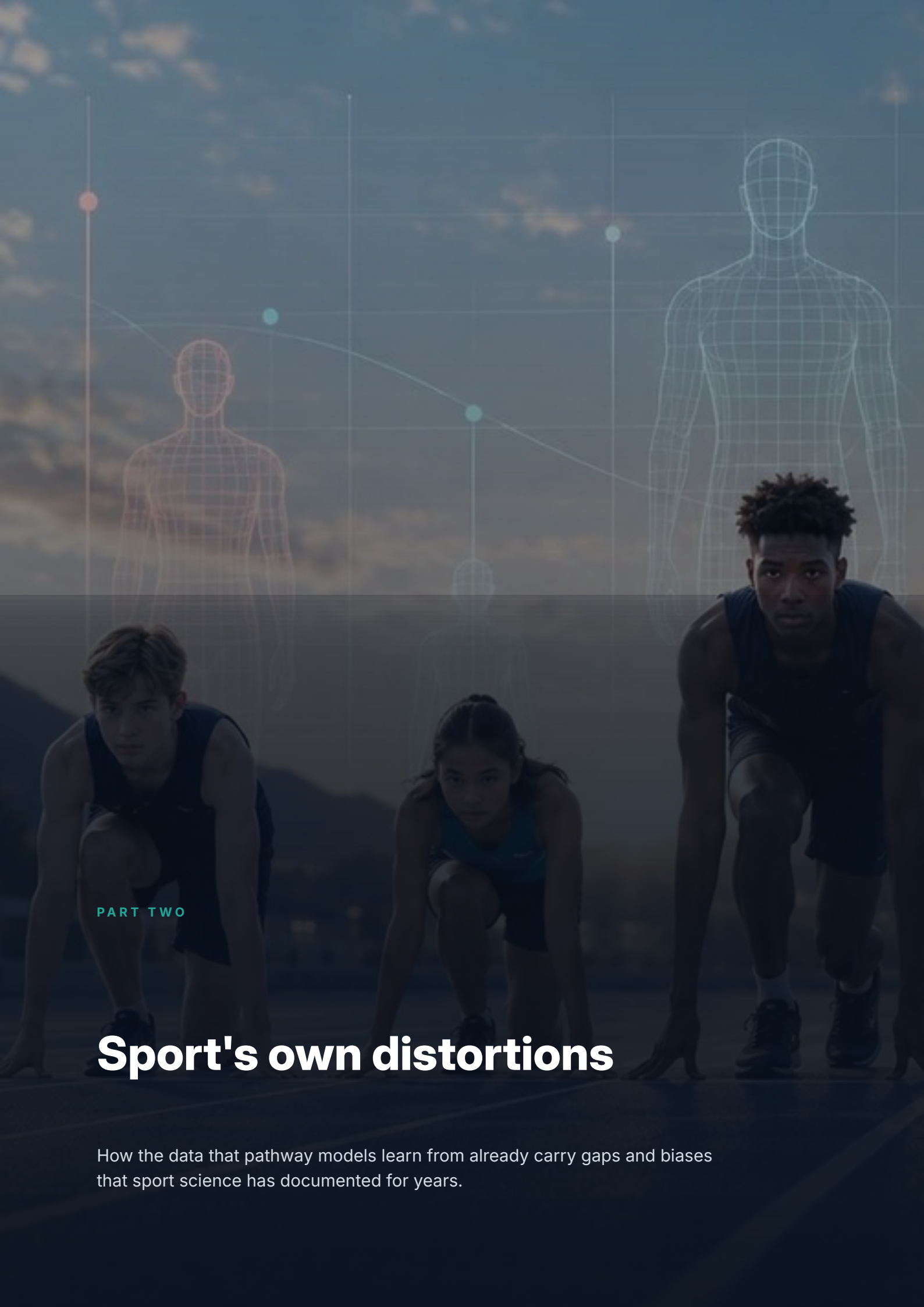
A third finding complicates any hope that fairness can be ticked off with one statistic. Kleinberg, Mullainathan and Raghavan set out three reasonable conditions for a fair risk score, one about calibration within groups and two about balance in how groups are scored, and showed that outside special cases, such as perfect prediction or identical base rates across groups, no score can satisfy all three together.¹¹

PEER-REVIEWED The special cases rarely hold in youth sport, because the base rate of later success is itself shaped by relative age and maturity.

For a governing body the practical meaning is that a vendor who claims its tool "is fair" has made a choice among incompatible definitions, whether or not it knows it. The pathway should ask which definition was chosen and why, and should decide for itself which kind of error it most wants to avoid. My own position, argued in Part three, is that on a development pathway the costlier error is usually the wrongful exclusion of a late developer, since a wrongful inclusion can be corrected at the next review while an exclusion tends to remove the athlete from the evidence altogether.

PROXY

A proxy is a measurable stand-in for something harder to observe. Its weakness is that it inherits every bias in how it was produced.

The image features three athletes in starting blocks on a track at dusk. Overlaid on them are wireframe models of human figures in various poses, connected by a line with colored dots, suggesting a data-driven analysis of movement. The background is a sunset sky with a grid overlay.

PART TWO

Sport's own distortions

How the data that pathway models learn from already carry gaps and biases that sport science has documented for years.

CHAPTER THREE

Sport's Existing Data Gaps

Before any model is trained, the evidence base it draws on is already uneven, most visibly for women and girls, para athletes and small elite groups.

Pathway tools learn from two kinds of data. Some learn directly from an organisation's own records of tests, selections and results. Others encode assumptions drawn from published sport science, such as normative test values, injury thresholds and growth curves. Both sources have known gaps.

3.1 The sex data gap

Cowley and colleagues examined human-participant research published between 2014 and 2020 in six leading sport and exercise science journals.³ [PEER-REVIEWED](#) Of 5,261 publications, 63% included both sexes, 31% included males only and 6% included females only; across the more than 12 million participants counted, 66% were male and 34% female.³ [PEER-REVIEWED](#) The authors present the result as a continuing sex data gap. The phrase in their title, "invisible sportswomen", names the problem plainly.

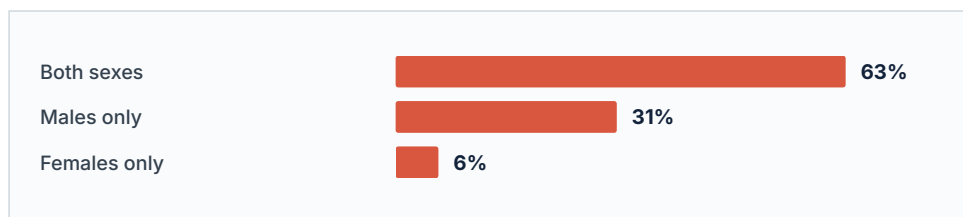


FIGURE 1 Share of 5,261 sport and exercise science publications by participant sex, 2014 to 2020. Source: Cowley et al. (2021).

Studies that include both sexes do not necessarily analyse them separately, so the true share of evidence that speaks specifically to female athletes is probably smaller than the 6% figure suggests. For a pathway tool the consequence is direct. Thresholds, growth assumptions and training-load rules that were derived mainly from male samples may be applied to girls without anyone checking that they hold, and girls' maturation follows a different timetable from boys', which makes the transfer especially doubtful around adolescence.⁵ [SCHOLARLY](#) A tool developed on a boys' academy and extended to girls, as in the opening scenario, is an easy thing to buy and a hard thing to validate.

3.2 Para sport and the small-sample problem

Para sport raises the same problem in a more acute form. Athletes are fewer, impairments vary widely, and classification systems already group athletes for competition. A model trained on a handful of athletes in each class will produce estimates with wide uncertainty, and a model trained across classes risks treating impairment-re-

lated differences as differences in talent. I have found little published work on how machine learning tools for talent or performance should be validated in para pathways, and I suspect this reflects the size of the populations as much as a lack of interest. **EVIDENCE GAP**

Classification itself adds a complication that I do not try to resolve here. Where a pathway's structure already sorts athletes by an assessment of impairment, a scoring tool layered on top may interact with that sorting in ways that neither the classifier nor the vendor anticipated, and nobody may be responsible for noticing. **PRACTICE**

3.3 Elite samples are small by construction

The final gap concerns elite sport generally. The athletes a talent pathway most wants to predict, those who reach senior international level, are by definition few. A model trying to learn what distinguishes them has a small number of positive cases, a long wait between the adolescent measurement and the adult outcome, and a population whose composition has been shaped by earlier selection. Baker, Cobley and Schorer's edited collection on talent identification brings together international perspectives on these difficulties, and its contributors return repeatedly to how hard it is to predict adult performance from measurements taken in adolescence.¹² **SCHOLARLY**

Small samples still permit modelling, although they leave subgroup claims fragile. If a tool's validation set holds 40 athletes who reached a senior squad, perhaps five of them were born in the last quarter of their selection year, and any statement about how well the tool serves that group rests on five people. The numbers in that sentence are illustrative, but the arithmetic is general, and it is the reason the recommendations in Part three insist on reporting the count behind every subgroup figure.

CHAPTER FOUR

Maturation and Relative Age as Structural Bias

Relative age and biological maturity already skew who is selected, so any model trained on selection history will learn the skew unless it is designed not to.

Two related phenomena distort youth sport selection. The first, relative age, concerns when in the selection year a child was born. The second, biological maturation, concerns how far through puberty a child has progressed, which varies widely among children of the same age and is easily mistaken for the first.

4.1 Relative age, a small effect that compounds

Cobley and colleagues conducted the first meta-analysis of relative age effects, re-examining 38 studies published between 1984 and 2007 that together contained 253 independent samples across 14 sports and 16 countries.⁴ **REVIEW** They found relative age effects to be consistently present but generally small, with effect size increasing as the relative age difference grew. The effect was moderated by age category, skill level and sport context, and adolescent males aged 15 to 18 competing at regional and national level in popular sports appeared most exposed.⁴ **REVIEW** The authors also noted that the evidence on female and more culturally diverse settings was thin. They called for policy and practitioner action to reduce the inequality.

A small effect at one selection point can matter a great deal over a pathway, because selection is repeated. Each round tilts the next cohort slightly towards the relatively older, and each tilted cohort becomes the data from which future decisions, and now future models, are made. Hancock, Adler and Côté's model explains part of the compounding through social agents, proposing that parents contribute through Matthew effects, coaches through Pygmalion effects and athletes through Galatea effects.⁸ **PEER-REVIEWED** A scoring tool is, in effect, a new social agent in that model, one that never tires and is rarely questioned. That reading is mine; the authors wrote before such tools were common in pathways.

4.2 Maturation and how it is estimated

Biological maturity describes progress towards the adult state, and it varies considerably among children of the same chronological age, most visibly around the adolescent growth spurt.⁵ **SCHOLARLY** An early-maturing 13-year-old may be bigger, stronger and faster than a late-maturing peer for reasons that have little to do with long-term potential, and the advantage tends to narrow once the later maturer catches up.

Because direct assessment of skeletal age requires imaging, many sports use anthropometric estimates. Mirwald and colleagues developed sex-specific equations predicting years from peak height velocity, known as maturity offset, from age, height, body mass, sitting height and leg length, using a sample of 152 Canadian children (79 boys

and 73 girls) followed through adolescence.¹³ **PEER-REVIEWED** The models explained much of the variance, with R^2 values of 0.92 for boys and 0.91 for girls and standard errors of about half a year, and they were cross-validated on further Canadian and Flemish samples. Even so, the authors advised caution in implementation and recommended treating maturity offset as a categorical measure.¹³ **PEER-REVIEWED**

Two points follow for pathway AI. The maturity estimate is itself a model, trained on a particular sample in particular decades, so a tool that adjusts for maturity inherits that model's error and its population limits. And an error of half a year is large in the context of a growth spurt, which means a tool that treats offset as a precise continuous input may be more confident than its inputs allow. Whether the original equations hold equally well for today's athletes from more diverse backgrounds remains, as far as I can find, an open question. **EVIDENCE GAP**

MATURITY OFFSET

Maturity offset is an estimate of how many years a child is from peak height velocity, the point of fastest growth. It carries an error of roughly half a year.

4.3 What bio-banding teaches model design

Bio-banding groups young athletes by maturity status, often estimated as a percentage of predicted adult stature, for some competition, training or assessment purposes. Cumming and colleagues describe its uses in competition, talent identification and strength and conditioning, and present it as a way to separate the effects of maturity from those of skill and ability, as a complement to age-group competition.¹⁴ **REVIEW** In a study of 66 Premier League academy players from four clubs, grouped at 85 to 90% of adult stature, early-maturing players reported that bio-banded games were more physically demanding and required more emphasis on technique and tactics, while later-maturing players valued the greater chance to use and show their technical, physical and psychological abilities.¹⁵ **PEER-REVIEWED** The sample was small and the evidence is self-reported, but the direction is informative.

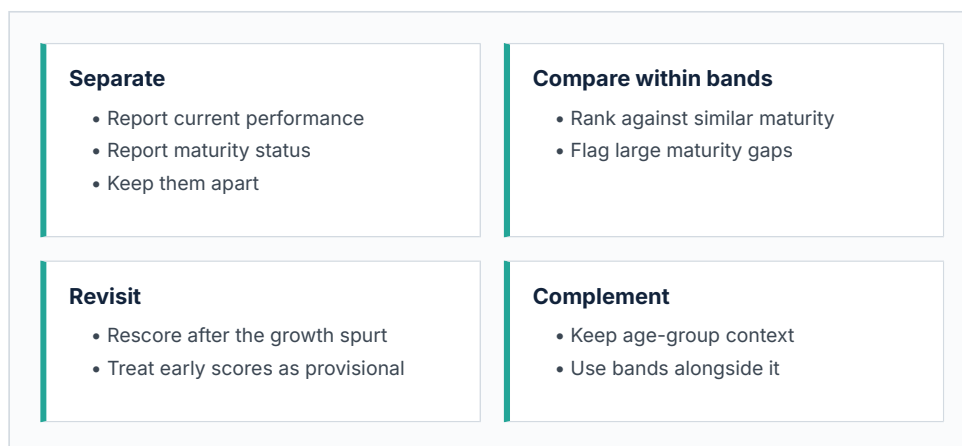


FIGURE 2 Three lessons from bio-banding translated into model design. Source: author's framework, drawing on Cumming et al. (2017; 2018).

The translation into model design is fairly direct. A tool should report current performance and estimated maturity as separate quantities, so that a panel can see when a strong score is being driven mostly by physical advancement. Where it produces comparisons, it should be able to compare athletes against peers of similar maturity as well as similar age. And it should treat any output produced during the growth

spurt as provisional, with a scheduled rescore once growth has settled. All three are technically simple, but each requires the vendor to have collected maturity data during development, and the buyer to ask whether it did. **ATHLEET.AI**



PART THREE

Designing for developmental fairness

A four-part account of fairness for pathway AI, and the tests and routes to challenge that should come before any tool is switched on.

CHAPTER FIVE

Four Dimensions of Fairness for Pathway AI

Fairness on a pathway has at least four dimensions, and the developmental one is the dimension that general AI ethics leaves undressed.

The general literature offers many fairness definitions, and, as Chapter 2 showed, they cannot all hold together. For a pathway I find it more useful to ask four questions about any tool, each tied to a different point where a young athlete can be treated unfairly. The four dimensions are my own framing, drawing on the sources above.

ATHLEET.AI

5.1 Representational and allocative fairness

Representational fairness asks who is in the data. A tool is representationally unfair to a group when that group is missing or thin in the training and validation data, so that the tool's behaviour for them is unknown or poor. Gender Shades is the canonical example outside sport,¹ [PEER-REVIEWED](#) and the sex data gap is its counterpart within it.³ [PEER-REVIEWED](#) The practical test is whether the vendor can show, for each group the pathway cares about, how many athletes were in the data and how the tool performed for them.

Allocative fairness asks who receives resources because of an output. Places, coaching time, sports science support, funding and competition exposure are all allocated on pathways, and a tool that shifts those allocations systematically towards one group is allocatively unfair even if its predictions are accurate by some measure. The health algorithm case shows how an allocation can go wrong through the choice of target,² [PEER-REVIEWED](#) and the relative age literature shows which groups a sporting allocation tends to favour.⁴ [REVIEW](#) The test is to run the tool on a past cohort and compare the resulting allocation by birth quartile, sex and maturity status with the pathway's intended one.

5.2 Developmental fairness

Developmental fairness asks whether an output, at the time it is used, prematurely narrows the options of an athlete who is still developing. It differs from the other three because it concerns time. A score can be well calibrated, drawn from representative data, and fairly allocated at the moment it is produced, and still be developmentally unfair if it is used to end an athlete's access to the pathway before the factors it measures have settled. The late maturer in the opening scenario was harmed in this way.

Three features make a tool more likely to be developmentally unfair. The first is using outputs from the growth spurt period to make lasting decisions about who stays on the pathway. The second is scoring athletes against age peers alone, without maturity

context.¹⁴ **REVIEW** The third, which is subtler, is a feedback loop in which the tool's outputs influence who gets the coaching that improves future scores, so that the tool appears to be confirmed by outcomes it helped to cause. That last mechanism mirrors the Pygmalion and Galatea effects Hancock and colleagues propose,⁸ **PEER-REVIEWED** and I suspect it would be almost impossible to detect from within a single pathway's data. **EVIDENCE GAP**

5.3 Procedural fairness and contestability

Procedural fairness asks whether athletes and parents can understand a decision in which a tool played a part, and whether they can challenge it. It is the dimension closest to current policy discussion. The UK government's white paper on AI regulation, published on 29 March 2023, proposes five cross-sector principles for regulators to apply, among them appropriate transparency and explainability, fairness, and contestability and redress, and it states that the government does not intend to put the principles on a statutory footing initially.⁷ **CONSENSUS** I read the white paper as an indication of the direction of policy; it does not, on its own terms, create new duties for sports bodies, and any organisation unsure of its legal position should take advice.

My own view, which goes beyond what the white paper says, is that sport should not wait for regulation to define contestability for it. A young athlete's route to challenge a selection decision already exists in most pathways in some form. The task is to make that route work when a tool was involved, which means families must know the tool was used, must be able to see the inputs that concerned their child, and must be able to put their case to a person with the authority to reverse the outcome. **ATHLEET.AI**

Dimension	Question it asks	Typical failure on a pathway	Test before use
Representational	Who is in the data?	Girls or para athletes thinly sampled	Subgroup counts and results from the vendor
Allocative	Who gets resources because of the output?	Places tilt to relatively older athletes	Replay a past cohort and compare allocations
Developmental	Does the output close options too early?	Growth-spurt scores used for deselection	Check maturity reporting and rescore rules
Procedural	Can families understand and challenge it?	No record that a tool was used	Walk a mock appeal through the process

TABLE 1 Four dimensions of fairness for pathway AI, with a pre-deployment test for each. Source: author's framework.

CHAPTER SIX

Testing Before Deployment

Pathways should establish whether a tool is fair to the athletes they are most likely to lose before it influences a single selection, and should give families a working route to challenge it.

The four dimensions matter only if they change what happens before a tool goes live. This chapter sets out what a pathway should require, then turns to contestability in practice, and ends with my recommendations and the questions I cannot yet answer.

6.1 Subgroup evaluation before use

Mitchell and colleagues proposed model cards, short documents that accompany a trained model and report its performance across relevant groups, including intersections of groups, along with its intended uses and known limits.⁶ **PEER-REVIEWED** The idea transfers well to sport. A pathway tool should come with a card that reports performance separately by sex, by relative age quartile and by maturity band, and for intersections where the numbers allow, such as late-maturing girls in the youngest quartile.

Two reporting rules make such a card useful. Each subgroup figure should show the number of athletes behind it and carry an uncertainty range, because a point estimate from a small group can look reassuring by chance. Groups below a minimum size should be marked as untested and kept out of the average. I have not found published guidance on what that minimum should be for youth sport models, and it is one of the open questions below. **EVIDENCE GAP** In its absence a pathway should set its own threshold, write it down, and refuse to use a tool for selection in any group that falls below it. **ATHLEET.AI**

1	Ask for the card	Obtain results by sex, relative age quartile and maturity band, with counts.
2	Check the target	Establish what the tool predicts and what bias that target carries.
3	Replay a cohort	Run the tool on a past intake and compare allocations by group.
4	Set the floor	Mark groups below the minimum sample as untested for selection.
5	Rehearse a challenge	Walk a mock family appeal through the process before go-live.
6	Schedule a rescore	Fix when growth-spurt outputs will be reviewed and replaced.

FIGURE 3 A pre-deployment fairness test for pathway AI tools. Source: author's framework.

The target check matters particularly, given the health algorithm case. A tool that predicts "selected at 16" has learned from the pathway's own selection history, with all the relative age and maturity bias that history contains. A tool that predicts a later, less selection-dependent outcome is harder to build and slower to validate, but it is less likely to launder past bias into a new score. The buyer should know which of the two it is buying.

6.2 Contestability in practice

Contestability is easy to promise and hard to deliver, mainly because the people with the least power in a pathway are the ones most likely to need it. A family challenging a deselection is usually doing so against a club or governing body on which the child's sporting future depends.

ILLUSTRATIVE SCENARIO

A parent asks why

In an invented county pathway, a father learns that his 12-year-old son has not been retained, and at the review meeting he is told that the decision "took account of the performance model". He asks what the model said and is shown a single ranking, with his son in the bottom third. Nobody present can tell him which inputs drove the ranking, whether the model adjusted for his son's August birthday, or whether the ranking was checked against the coach's view. He leaves with no route to take the matter further. In a pathway that had prepared for this conversation, the same father would have been told beforehand that a tool was used, shown the separate current performance and maturity figures for his son, told that outputs from this age are treated as provisional, and given the name of a person who could order a rescore or a fresh assessment.

The difference between those two versions of the meeting costs little. It needs a record that the tool was used, a plain explanation of what it measures, access for the family to their own child's inputs, and a named reviewer with authority to overturn the outcome. Where personal data about children is involved, organisations will also have data protection obligations that are outside the scope of this paper, and should take their own advice on them. **PRACTICE**

6.3 Recommendations

The recommendations below are my own and are proposed for pathways that use, or are considering, AI tools that score or profile young athletes. **ATHLEET.AI**

1. Require a subgroup evaluation, with counts and uncertainty ranges, by sex, relative age quartile and maturity band, before any tool is used in selection.
2. Establish what each tool predicts, and prefer targets that do not simply reproduce the pathway's past selection decisions.
3. Set and record a minimum subgroup size, and treat any group below it as untested for selection purposes.
4. Require current performance and estimated maturity to be reported separately, and treat outputs from the growth spurt as provisional with a fixed rescore date.
5. Replay each tool on at least one past cohort and compare the allocation it would have produced with the pathway's intended one.
6. Tell families when a tool contributed to a decision, give them their child's inputs, and name a person with authority to overturn the outcome.
7. Review each tool annually against the four dimensions, and suspend its use in selection if subgroup performance cannot be shown.

These are deliberately procedural. None asks a pathway to build a model, and each can be written into a procurement specification or a selection policy within a season.

6.4 Conclusion and open questions

The argument of this paper is that fairness on a development pathway has a temporal dimension that general AI ethics has not needed to consider, and that pathway tools trained on selection history will, unless designed otherwise, learn the relative age and maturation biases that sport science has documented for years. The general AI literature supplies the mechanisms, in hidden subgroup failure and biased proxies, and the sport literature supplies the populations at risk. Joining the two points to a modest set of tests that any pathway can require before use.

Several questions remain open, and I would welcome evidence on any of them.

- What minimum subgroup sample should a youth sport model meet before it is used in selection, and should the threshold differ by decision type?
- Should maturity status be a mandatory model input, or does collecting it at scale create privacy and misuse risks that outweigh the benefit?
- How should developmental fairness be assessed in para sport, where classification already shapes who competes with whom?

- Can a pathway detect the feedback loop in which a tool's outputs shape the coaching that later appears to confirm them, or does that require comparison across organisations?

The last of these is the one I find most troubling, partly because I suspect the answer is that a single pathway cannot see it at all.

END MATTER

References

1. Buolamwini, J. and 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.
2. Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464), 447-453.
3. Cowley, E.S., Olenick, A.A., McNulty, K.L. and 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.
4. Cobley, S., Baker, J., Wattie, N. and McKenna, J. (2009) Annual age-grouping and athlete development: a meta-analytical review of relative age effects in sport. *Sports Medicine* 39(3), 235-256.
5. Malina, R.M., Bouchard, C. and Bar-Or, O. (2004) *Growth, Maturation, and Physical Activity*. 2nd edn. Champaign, IL: Human Kinetics.
6. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019) Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.
7. Department for Science, Innovation and Technology (2023) *A pro-innovation approach to AI regulation*. Command Paper CP 815, 29 March 2023.
8. Hancock, D.J., Adler, A.L. and Côté, J. (2013) A proposed theoretical model to explain relative age effects in sport. *European Journal of Sport Science* 13(6), 630-637.
9. Competition and Markets Authority (2023) CMA launches initial review of artificial intelligence models. Press release, 4 May 2023.
10. US Senate Committee on the Judiciary, Subcommittee on Privacy, Technology and the Law (2023) Oversight of A.I.: Rules for Artificial Intelligence. Hearing, 16 May 2023.
11. Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017) Inherent trade-offs in the fair determination of risk scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, LIPIcs 67, 43:1-43:23.
12. Baker, J., Cobley, S. and Schorer, J. (eds) (2012) *Talent Identification and Development in Sport: International Perspectives*. London: Routledge.
13. Mirwald, R.L., Baxter-Jones, A.D.G., Bailey, D.A. and Beunen, G.P. (2002) An assessment of maturity from anthropometric measurements. *Medicine and Science in Sports and Exercise* 34(4), 689-694.
14. Cumming, S.P., Lloyd, R.S., Oliver, J.L., Eisenmann, J.C. and Malina, R.M. (2017) Bio-banding in sport: applications to competition, talent identification, and strength and conditioning of youth athletes. *Strength and Conditioning Journal* 39(2), 34-47.
15. Cumming, S.P., Brown, D.J., Mitchell, S., Bunce, J., Hunt, D., Hedges, C., Crane, G., Gross, A., Scott, S., Franklin, E., Breakspear, D., Dennison, L., White, P., Cain, A., Eisenmann, J.C. and Malina, R.M. (2018) Premier League academy soccer players' experiences of competing in a tournament bio-banded for biological maturation. *Journal of Sports Sciences* 36(7), 757-765.

END MATTER

About the author

Paul Forrest is the founder of Athleet.AI. He advises boards and leadership teams on the adoption and governance of artificial intelligence, drawing on a background in business process reengineering and on hands-on work building machine learning systems, including small, domain-specific language models. He has served as a special adviser to UK government on the Global Entrepreneur Programme and works as a board member and business adviser.

He is also a qualified athletics coach who coaches sprint, hurdle and endurance athletes as a volunteer, and much of his thinking about how people learn, decide and improve comes from the track.

This paper is part of the Athleet.AI Working Papers, AI in Coaching and Athlete Development. The views expressed are the author's own.

athleet.ai

Maturation, Bias and the Training Set

Athleet.AI Working Papers, AI in Coaching and Athlete Development
Paper four of six · 24 May 2023