



Load, Readiness and False Precision

The Limits of Algorithmic Training Prescription in Long-Term Endurance Development

About this paper

This paper examines what happens when contested training load science is packaged into software and shown to developing endurance athletes, their parents and the volunteer coaches who plan their weeks. It is written for club coaches, performance and education leads in governing bodies, and anyone commissioning or buying a tool that turns training data into a readiness score, a risk colour or a recommended load. It does not review any product, and it makes no claim about what any particular tool does.

Within the Athleet.AI strand, the paper carries the same single argument as its companions, namely that AI in coaching and athlete development has to respect the non-linear, human and developmental nature of athletic progress. The first paper in the strand made that case for talent identification. This one makes it for the weekly business of deciding how much a young runner, rower, swimmer or cyclist should train, where the pull towards a tidy number is strongest and the science beneath the number is thinnest.

Editorial balance. Roughly half of the paper reports and weighs published evidence on load monitoring, injury prediction, periodisation and individual response. About a third is my own analysis of how interface design converts that uncertain evidence into apparent precision, and the remainder sets out a proposed framework of design principles, marked throughout as the author's own and offered for testing.

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 · WHAT THE SCIENCE CAN BEAR

1	A red flag before the county race	9
2	What load monitoring can defend	11
3	Where the confident metrics fail	13

PART TWO · INDIVIDUALS, TIME AND THE INTERFACE

4	Plans, people and long horizons	16
5	How software manufactures certainty	18

PART THREE · DESIGNING FOR HONEST UNCERTAINTY

6	Principles for honest load tools	22
---	----------------------------------	----

END MATTER

References	25
About the author	26

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 19 June 2023. Sources and events are as they stood at that date.

CITATION

Forrest, P. (2023) *Load, Readiness and False Precision: The Limits of Algorithmic Training Prescription in Long-Term Endurance Development*. Athleet.AI Working Papers, AI in Coaching and Athlete Development, Paper 5. Athleet.AI, 21 June 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 week I stopped trusting the red square

For most of my coaching life I have kept training diaries by hand, and for the last few years I have also kept them in software. The software is better at arithmetic than I am. It is worse at knowing that a teenager slept badly because of exams, or that her easy run was slow because she ran it with a friend who was coming back from illness, and it has no way of knowing that a flat fortnight in March is often what precedes a good summer.

What pushed me to write this paper was a small moment. A colour on a screen changed, and the athlete concerned read it before I did. The number behind it was a ratio whose standing in the research literature had shifted considerably since it was first popularised, yet nothing on the screen said so. The colour carried the confidence of a medical test. I spent the next conversation undoing a worry that the data did not justify, and I came away thinking that the fault lay chiefly with a habit in our field of treating a descriptive figure as though it were a forecast.

In my advisory work outside sport I see the same habit whenever a dashboard is built faster than the evidence behind it. Sport has an additional reason for care, because many of the people reading these screens are young, many of the coaches are unpaid, and the timescale that matters is years, while the software thinks in days.

None of this makes me hostile to measurement. I would not want to coach without a record of what athletes have done and how it felt. The case in these pages is narrower. Tools should describe load honestly, show how uncertain their inferences are, and leave the decision about next week with a person who knows the athlete. Where I go beyond the evidence I say so, and the framework in the final chapter is mine, offered for challenge.

Paul Forrest
Founder, Athleet.AI
21 June 2023

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

Training load software can make a weak metric look authoritative, and in long-term endurance development that false precision is a greater risk than the absence of data. The answer is to build and buy tools that describe load and its trend, label any classification as descriptive and non-clinical, show their uncertainty, and leave prescription with the coach.

Load monitoring rests on some defensible foundations, including the distinction between internal and external load and a well-described pattern of intensity distribution in trained endurance athletes. Around those foundations has grown a set of confident derived metrics whose predictive value has been heavily criticised. When software presents those metrics as single numbers, colours and recommended loads, it strips out the uncertainty that the research literature itself reports, and it does so in front of readers who are often young and often unsupported.

1

The distinction between internal and external load is the most defensible part of load monitoring.

Internal load, the athlete's physiological and psychological response, is what drives adaptation, and external load, the work done, only makes sense alongside it.¹ **PEER-REVIEWED** Simple self-report measures appear to track training responses at least as well as many objective markers,² **REVIEW** so tools that record both kinds of load, and that treat self-report as data, stand on firm ground.

3

Injury prediction in sport remains weak, and machine learning has not yet changed that.

A critical review argued that screening tests are unlikely ever to predict injury well,⁶ **REVIEW** and a systematic review found that most published injury prediction models carry a high risk of bias.⁷ **REVIEW** Machine learning studies are growing in number, but a systematic review of them found only 11 eligible studies, with predictive performance ranging from poor to strong and methodological quality that could still be improved.⁸ **REVIEW**

5

Periodisation offers a plan under uncertainty, and software tends to present it as a mechanism.

Critics argue that much periodisation practice is tradition-led and rests on assumptions of predictable adaptation that biology does not honour.^{11,12} **REVIEW** An algorithm that prescribes next week's load to a precise figure imports those assumptions and hides them.

2

The acute:chronic workload ratio should not be presented as an injury predictor.

The ratio was popularised as a way to identify a safe zone of load progression.³ **PEER-REVIEWED** Later analyses argued that its conventional calculation produces spurious correlation and that the ratio has fundamental conceptual and statistical flaws, with no evidence supporting its use in managing injury risk.^{4,5} **PEER-REVIEWED**

4

Individual responses to the same training vary widely, and population models misfit individuals.

Standardised training produces widely different fitness gains in different people.⁹ **REVIEW** Separating true individual response from measurement error and day-to-day variation requires designs that most monitoring data cannot supply.¹⁰ **PEER-REVIEWED**

6

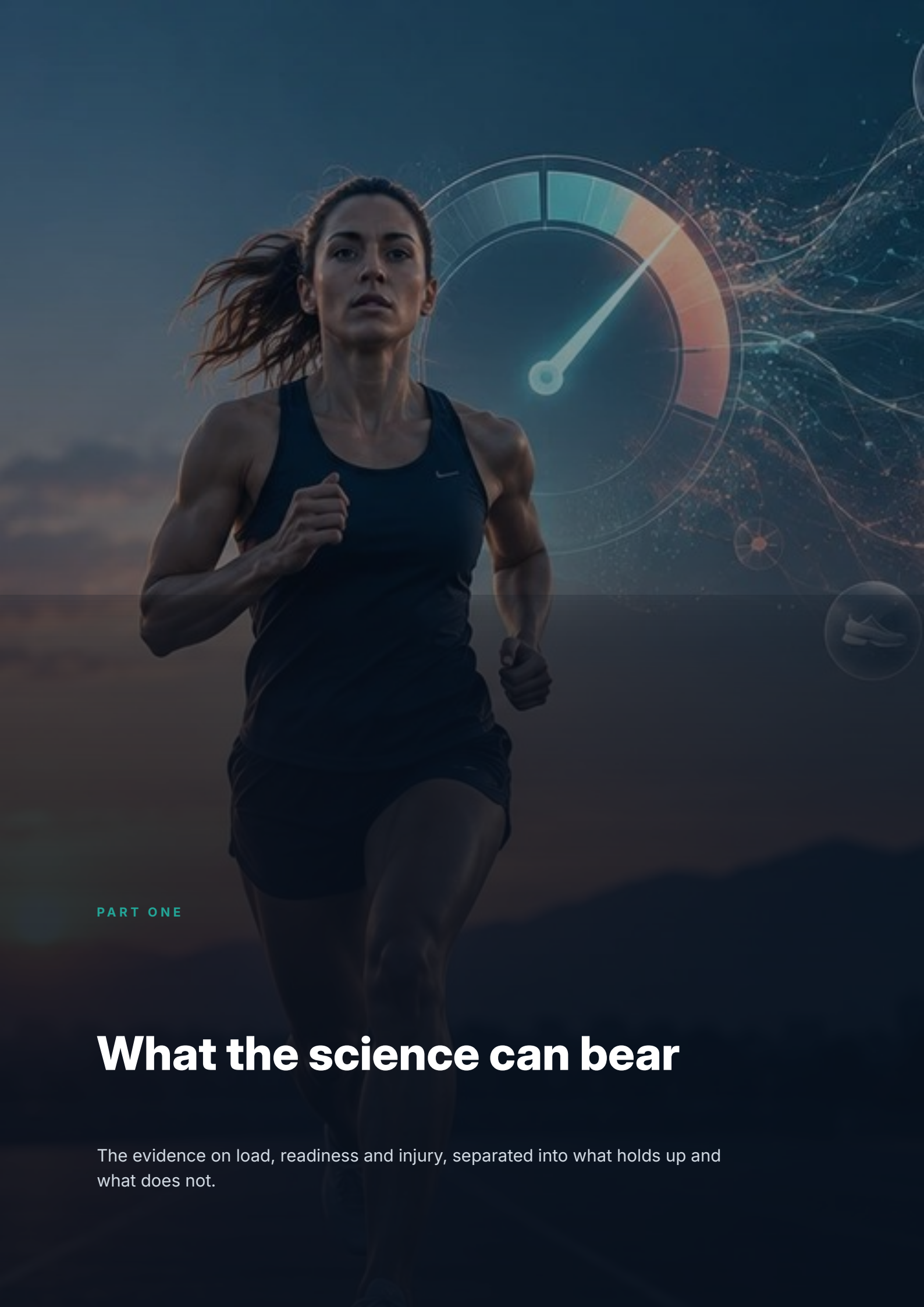
Interface choices, more than algorithms, are where false precision is manufactured.

Single-number scores, traffic-light colours and confident verbs imply a certainty the underlying science does not have. **ATHLEET.AI** Young athletes and volunteer coaches are the readers least equipped to discount that signal, and consumer platforms increasingly place it in front of them. **PRACTICE**

What follows for a club or governing body

Clubs and governing bodies do not need to reject load tools, but they should set conditions for their use with developing athletes. Those conditions include no injury-risk labels without validated evidence for the population concerned, visible uncertainty on every derived figure, descriptive language for any classification, and a plain statement that the coach, working with the athlete, owns the training decision.

Governing bodies are also well placed to publish guidance that volunteer coaches can use when a parent or athlete arrives worried about a colour on a screen. My own view is that such guidance would do more good, sooner, than any improvement in the underlying models, because the models will remain uncertain for years while the screens are already in use.

A woman with long brown hair is running towards the camera. She is wearing a black Nike tank top and black shorts. The background is a dark, hazy landscape. Overlaid on the image is a large, glowing speedometer with a needle pointing towards the right. To the right of the speedometer is a complex, glowing network of lines and nodes, resembling a data visualization or a molecular structure. In the bottom right corner, there is a small, glowing circular icon of a running shoe.

PART ONE

What the science can bear

The evidence on load, readiness and injury, separated into what holds up and what does not.

CHAPTER ONE

A red flag before the county race

A readiness score is a claim about the future dressed as a description of the past, and in youth endurance sport the reader of that claim is often the least able to question it.

Endurance coaching has always involved judgements about load. How far, how fast, how often and with how much rest are questions every coach answers weekly, and for most of the history of the sport the answers were written in notebooks and argued over at the track. Software has changed the medium of those judgements more than their substance. What it has also changed, and what this paper is about, is the confidence with which the answers are presented.

ILLUSTRATIVE SCENARIO**A colour on a Thursday**

Maya is 15 and runs the 1500m for a club with two volunteer endurance coaches. Her mother pays for a consumer training app that links to her watch. On the Thursday before the county championships, after a week in which a school cross-country fixture was added to her usual sessions, the app shows her load ratio in red and a message that her injury risk is high. It suggests she reduce training by a set percentage for the coming week. Maya reads the message on the bus, tells her coach she is worried about getting injured and asks whether she should race at all. Her coach, who has known her for four years, thinks she is tired from exams but otherwise well, and knows that her weekly volume is modest by any standard. The coach has no way to see how the app reached its figure, and no way to tell whether the red square reflects anything more than one extra race in a quiet week.

The scenario is invented, but its elements are ordinary. Consumer fitness platforms increasingly surface readiness and load scores to amateur and youth athletes, often with colour coding and suggested adjustments. **PRACTICE** Much of this is sold within a wider enthusiasm for AI, and consultancy estimates of the economic value of generative AI alone run to trillions of dollars a year,¹³ **INDUSTRY** which gives some sense of the commercial pressure to add intelligent features to every product that holds data.

1.1 Why false precision is the main risk

My argument is that, in long-term endurance development, the most serious risk from algorithmic load tools lies in apparent precision, and that data scarcity is the smaller problem. A number on a screen implies that it was measured. A colour implies a threshold, and a threshold implies that someone validated it. A recommended load for next week implies that the relationship between this week's training and next week's adaptation is known well enough to be calculated. Each implication may be false, and the reader has no means of telling.

The harm runs in two directions. An athlete may be frightened away from a sensible session or a race by a signal that carries no predictive weight, as Maya nearly was. Equally, a green screen may reassure an athlete who is unwell, under-fuelled or carrying an early overuse problem that no load metric would detect. Both errors are more likely when the reader is young, when the coach is a volunteer without sports science training, and when the tool gives no hint of its own uncertainty. **ATHLEET.AI**

A third harm is slower and harder to see. When athletes learn to consult a score before they consult their own sense of how training is going, they may lose some of the self-knowledge that endurance sport depends on in the long run, since no device accompanies an athlete into the final lap of a race. I cannot point to research that measures this effect in young endurance athletes, and I raise it as a concern drawn from practice. **PRACTICE** **EVIDENCE GAP** An earlier paper in this strand, *The Coach in the Loop*, made a parallel point about coaches, whose judgement may weaken if it is exercised less often.

1.2 Scope and approach

The paper concentrates on endurance sports, where training volume is high, where the athletes who concern me are often between 12 and 18, and where development unfolds over many years. Its evidence is drawn mostly from sport science. Some of the literature predates the current wave of AI products, and I use it deliberately, because the metrics now being automated were developed and criticised in that earlier period. Where the evidence does not answer a question, the paper says so, and where I draw an inference, it is marked as mine.

CHAPTER TWO

What load monitoring can defend

Before criticising what software does with load data, it helps to be precise about which parts of load monitoring have earned their place.

The case against false precision could be misread as a case against measurement. It is nothing of the kind, and several parts of the load monitoring literature are well argued, widely used and useful to coaches at every level. The problem arises further along, when defensible descriptions are converted into predictions.

2.1 Internal and external load

The distinction between external load, the work an athlete performs, and internal load, the physiological and psychological response to that work, has become the organising idea of the field. Impellizzeri, Marcora and Coutts, reviewing its development over 15 years, argue that internal load is the stimulus that drives training outcomes, while external load matters chiefly for understanding and planning the work that produces it.¹ **PEER-REVIEWED** Two athletes running identical sessions may therefore experience internal loads that differ widely. A record of distance or pace alone says little about what the session cost either of them.

The International Olympic Committee consensus statement on load and injury, published in 2016, takes a similar line, treating load as a broad concept that includes both training and competition, alongside psychological and life stressors, and recommending that both internal and external measures be monitored.¹⁴ **CONSENSUS** The same statement drew attention to rapid changes in load as a risk factor, and it discussed the acute:chronic ratio with some approval, a position that later critiques challenged and that I return to in the next chapter.

2.2 Intensity distribution in endurance training

For endurance coaches, one of the better-described patterns concerns how training intensity is distributed. Seiler's review of best practice reports that successful endurance athletes across several sports perform about 80 per cent of their training sessions at low intensity, below the first ventilatory threshold, with most of the remainder at high intensity.¹⁵ **REVIEW** The review is descriptive of what well-trained adults do, and its author is careful about how far it supports prescription. It nonetheless gives coaches a vocabulary for reviewing a training week that software can support usefully, simply by showing the proportion of time spent in each zone.

That usefulness depends on the zones being set sensibly for each athlete, which for a growing adolescent is a moving target. A heart rate zone set from a test six months ago may no longer describe the same physiological boundary, and a tool that reports time in zone to the minute without saying when the zones were last checked is already slightly overstating what it knows. **PRACTICE**

INTERNAL AND EXTERNAL

External load is what the athlete did, such as distance, pace or power, and internal load is what it cost them, such as heart rate response or perceived effort. Most of the useful information sits in the relationship between the two.

2.3 Self-report and perceived exertion

Perhaps the most robust tool available to a volunteer coach costs nothing. Foster and colleagues described a session rating of perceived exertion (RPE) method in which the athlete rates the whole session's effort and the rating is multiplied by its duration, giving a simple index of internal load.¹⁶ PEER-REVIEWED A later systematic review concluded that subjective self-reported measures, such as mood, perceived stress and recovery, responded to changes in training load with greater sensitivity and consistency than commonly used objective measures such as blood markers and heart rate indices.² REVIEW

This matters for the design argument that follows. If simple self-report stands up so well, then a tool that asks an athlete how she feels, records the answer faithfully and shows it beside the training record is doing something defensible. The difficulty begins when the same tool combines that answer with other inputs through an undisclosed weighting and returns a single score with two decimal places.

CHAPTER THREE

Where the confident metrics fail

The derived metrics most attractive to software designers are those whose predictive claims have been most heavily criticised.

Between the defensible descriptions of the previous chapter and the scores that appear on consumer screens lies a layer of derived metrics, thresholds and prediction models. The research literature on that layer is less reassuring, and a reader who met these metrics only through an app would have little reason to know it.

3.1 The rise and fall of a ratio

The acute:chronic workload ratio compares an athlete's recent load, typically over one week, with her longer-term load, typically over four weeks. Gabbett's influential 2016 paper argued that athletes are better protected by building high chronic loads gradually than by avoiding hard training, and it described a range of the ratio, roughly 0.8 to 1.3, associated with lower injury risk, with values of about 1.5 and above associated with higher risk.³ **PEER-REVIEWED** The idea was intuitive, easy to compute and easy to display, and it spread quickly through team and individual sport.

Subsequent work questioned it on two fronts. Lolli and colleagues showed that when the acute week is included within the chronic period, as in the conventional calculation, the two terms are mathematically coupled, which produces a spurious correlation between them.⁴ **PEER-REVIEWED** Impellizzeri and colleagues went further, arguing that the ratio lacks a sound conceptual basis, that it does not normalise load in the way its users assume, and that the evidence for it is weak enough that it should be dismissed as a framework for managing injury risk.⁵ **PEER-REVIEWED** The debate has continued in the journals, and some practitioners still find the underlying idea about avoiding sudden spikes in load a useful rule of thumb. **PRACTICE** On the specific question of whether a ratio value should be shown to an athlete as a measure of her injury risk, however, my own reading is that the critiques leave little room.

A metric can be contested in the literature and uncontested on the screen.

Nothing in a coloured ratio tells its reader that the calculation behind it has been challenged in the same journals that popularised it. The asymmetry between what researchers argue and what users see is the core of the false precision problem.

3.2 Screening, prediction and the problem of bias

The weakness of the ratio is part of a wider pattern. Bahr's critical review argued that screening tests are unlikely to predict sports injury, and probably never will, because a risk factor that is strongly associated with injury across a group can still discriminate poorly between the people who will and will not be injured.⁶ **REVIEW** The distinc-

tion between association and prediction is simple to state and easy to lose in software, where any variable correlated with injury in a published study may be turned into a personal risk flag.

Bullock and colleagues assessed the conduct of published musculoskeletal injury prediction models in sport and found that most were at high risk of bias, with common problems including small samples, inadequate assessment of model performance and poor reporting, and none of the included studies externally validated a model.⁷

REVIEW A model that has not been tested on athletes outside the group that produced it gives little assurance for a different population, and developing endurance athletes are rarely the population on which models were built. **EVIDENCE GAP**

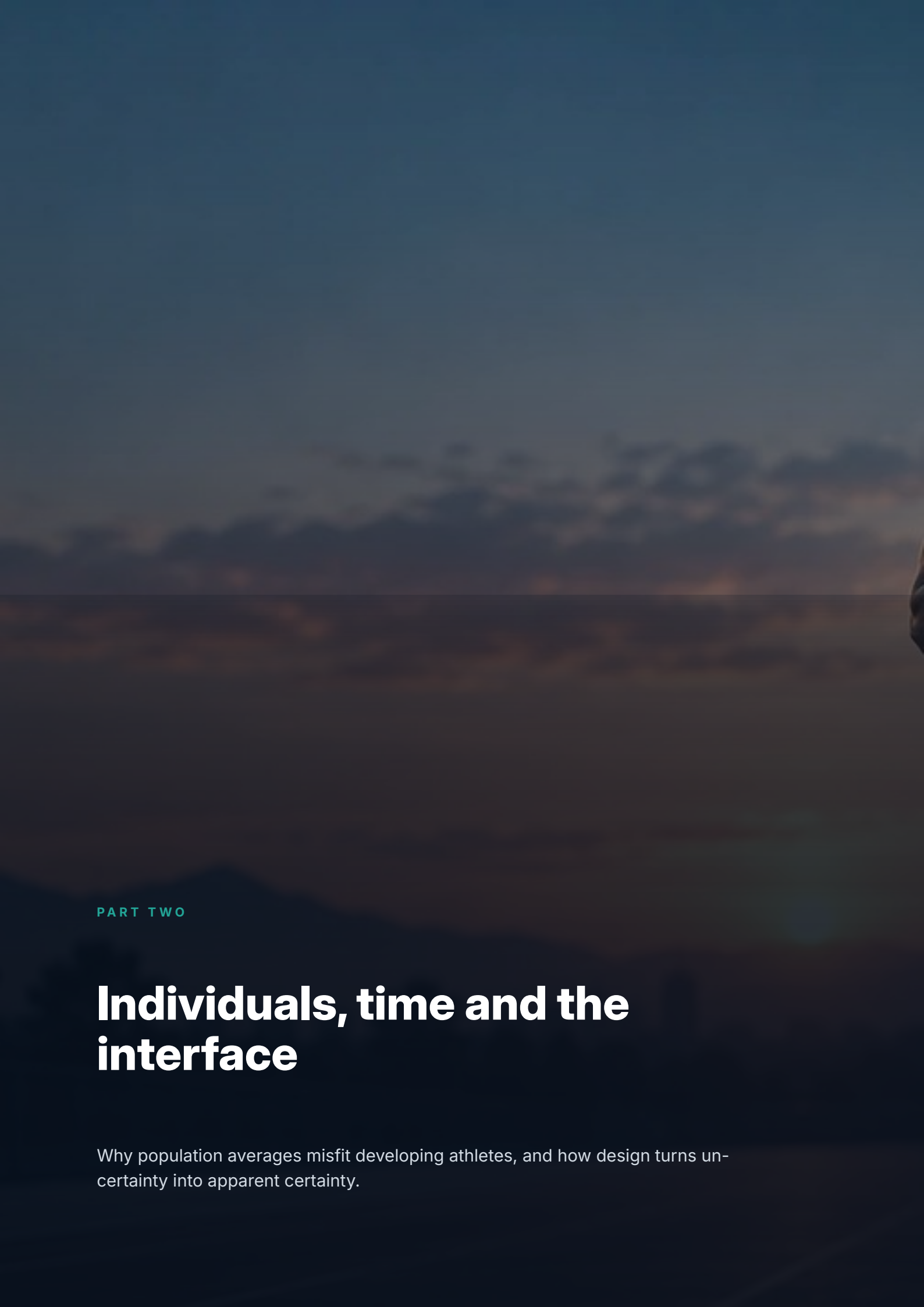
3.3 Machine learning inherits the problem

Machine learning is sometimes presented as the way past these limitations, since it can combine many inputs and find non-linear relationships. Rossi and colleagues, working with GPS data from a single professional football club across one season, reported that a machine learning approach forecast injuries better than several baselines, including ratio-based methods.¹⁷ **PEER-REVIEWED** The study is careful, but its setting is a single squad of adult professionals. Its design therefore cannot show whether the approach would transfer elsewhere.

A systematic review of machine learning in sport injury prediction and prevention found only 11 eligible studies, reported predictive performance ranging from poor to strong, and judged methodological quality generally sufficient but open to improvement.⁸ **REVIEW** Nothing in that literature, as I read it, supports showing a personalised injury probability to a teenage distance runner, and, on my reading, most of the included work concerned adult team sports whose injury patterns differ considerably from those of young endurance athletes. More flexible models can also fit noise more readily, which makes external validation more important. Yet external validation is the step most often missing.

Metric or claim	What the evidence supports	Suitable for automated display to developing athletes
External load (distance, time, pace)	Accurate description of work done	Yes, as a record
Session RPE and self-report	Useful index of internal load and response	Yes, shown as the athlete reported it
Time in intensity zones	Descriptive, depends on zone accuracy	Yes, with the date zones were set
Acute:chronic workload ratio	Contested concept and calculation	Trend only, never as a risk label
Injury risk score or probability	Most models at high risk of bias	No, absent validation in that population
Recommended next-week load	Rests on assumed predictable adaptation	As an option for the coach, never as an instruction

TABLE 1 Load metrics against the evidence reviewed in this paper and the author's proposed display rule. Sources are cited in the text; the final column is the author's framework.



PART TWO

Individuals, time and the interface

Why population averages misfit developing athletes, and how design turns uncertainty into apparent certainty.

CHAPTER FOUR

Plans, people and long horizons

Training theory describes average tendencies over months, while a young athlete lives through one particular body over years.

Even if the derived metrics were sound, two further problems would limit what software could safely say to a developing athlete about her training. The first is that planned training theory has weaker foundations than its textbooks imply, and that its most careful critics have been saying so for more than a decade. The second is individual variation, which is large and hard to measure.

4.1 Periodisation as a plan under uncertainty

Kiely asked whether periodisation practice is led by evidence or by tradition. He concluded that much of it rests on inherited planning assumptions that are no longer scientifically justifiable.¹¹ **REVIEW** In a later paper he argued that periodisation theory assumes that the body responds to a planned stimulus in predictable ways, when adaptation is shaped by many interacting stressors and is inherently uncertain, so that plans should be held loosely and revised in light of how athletes actually respond.¹²

REVIEW Mujika and colleagues, writing from applied practice, described an integrated approach in which periodisation spans nutrition, recovery, psychological preparation and skill, each adjusted to the athlete and the context.¹⁸ **REVIEW**

Read together, these papers point towards a view of the training plan as a working hypothesis revised through conversation. Software, by its nature, favours a fixed plan that the athlete is expected to follow, with deviations recorded as failures of compliance. A model that outputs next week's load to a precise figure assumes that the mapping from stimulus to response is stable enough to calculate. The critics dispute exactly that assumption. **ATHLEET.AI** (I should add that many experienced coaches already work this way and would not describe it in these terms; they simply adjust the plan when the athlete arrives looking flat.)

4.2 Responders, non-responders and noise

Bouchard and Rankinen reviewed evidence, much of it from large family studies of standardised training, showing wide individual differences in how people's fitness responds to the same exercise programme, with some people gaining a great deal and others gaining little.⁹ **REVIEW** This finding is sometimes quoted as proof that a personalised algorithm is needed. It could equally be read as a warning, since a model trained on group averages will fit a great many people poorly.

Hecksteden and colleagues added an important statistical caution. Much of what looks like individual variation in response to training may reflect measurement error and within-subject variation in how well training works from one period to another, and separating true individual response from that noise requires repeated measurement and, in their view, repeated interventions.¹⁰ **PEER-REVIEWED** Routine monitoring

data from a club athlete rarely has that structure. A tool that claims to have learned how a particular athlete responds may therefore be describing noise with great confidence, and the athlete has no way to tell which. **EVIDENCE GAP**

4.3 Development takes years, and it is not linear

The first paper in this strand, *Predicting the Unpredictable*, argued that athletic development is non-linear and that early indicators are poor guides to later achievement. The same applies to load. The International Olympic Committee consensus on youth athletic development emphasises that growth, maturation and psychosocial development proceed at individual rates, and it warns against early specialisation and excessive training loads, with their associated risks of overuse injury and burnout.¹⁹

CONSENSUS During the adolescent growth spurt, a training load that suited an athlete a year earlier may be excessive, while a flat or even declining performance may be a normal feature of growth.

Growth also disturbs the measures themselves. Body mass, stride length and cardiovascular dimensions all change during adolescence, so a pace or heart rate that meant one thing in September may mean something different by the following summer, and baselines drawn from the earlier period can quietly go stale. **PRACTICE**

Software designed around weekly optimisation is poorly matched to that timescale. A tool that rewards steady week-on-week progression may nudge athletes and parents towards chasing short-term gains in exactly the years when patience matters most, and a readiness score that dips during a growth phase may be read as a problem to be solved when it is a stage to be lived through. **ATHLEET.AI** Whether any current tool accounts for maturation in its load models is something I have not been able to establish from published evidence. **EVIDENCE GAP**

NOISE AND SIGNAL

A change in a daily score may reflect a real change in the athlete, an error in the sensor, or ordinary biological variation. Without repeated measures and a sense of typical variation, a single day's change says little.

CHAPTER FIVE

How software manufactures certainty

The step from uncertain science to apparent precision is taken mostly in the interface, through choices about numbers, colours and words.

Much of the debate about AI in sport focuses on algorithms. For load and readiness tools, I would argue that the more consequential choices are made in presentation, in the decisions about what to compress, what to colour and what to say, and those decisions are often taken by product teams with little contact with the research literature. An honest model shown through a misleading interface still misleads.

5.1 What a readiness score contains

The term readiness has no agreed definition in the research literature reviewed here, and in consumer products it usually names a composite. Typical ingredients include overnight heart rate measures, sleep duration estimated from a wrist sensor, recent training load and, in some products, a short self-report. **PRACTICE** Each ingredient carries its own measurement error. The method of combining them is rarely published in a form that an outside reader can examine. The athlete sees the composite and almost never the parts.

That opacity matters for a specific reason. The component with the strongest evidence behind it, the athlete's own report of how she feels,² **REVIEW** may carry only a modest weight in the composite, while a wrist-derived sleep estimate of uncertain accuracy for adolescents may carry more. Whether that is so for any given product I cannot say, and that inability is itself the point. **EVIDENCE GAP** A coach who asked the athlete directly would at least know what the answer was based on.

5.2 Single numbers and traffic lights

Collapsing several uncertain inputs into one score hides which input moved. A readiness score of 62 invites comparison with yesterday's 68 as though the six-point difference meant something, even when the variation in the underlying measures might easily account for it. Displaying decimals compounds the problem, because precision in the display is read as precision in the measurement. **ATHLEET.AI**

Colour works in a similar way. Traffic-light schemes borrow the visual language of clinical and safety warnings, and red in particular signals danger to almost any reader. When a threshold between amber and red was drawn from a contested ratio, or chosen by a product team for visual balance, the colour still carries the full authority of a warning. **PRACTICE**

1	Contested metric	A ratio or model with weak or disputed predictive evidence.
2	Single score	Several noisy inputs combined into one number without error bars.
3	Threshold and colour	An arbitrary cut point drawn in the language of warning.
4	Confident wording	Verbs such as predicts or detects imply validated capability.
5	Prescription	A precise load for next week presented as an instruction.
6	Belief	Athlete or parent acts on the signal without the coach.

FIGURE 1 How uncertain evidence becomes apparent precision on a screen. Source: author's framework.

5.3 Confident language and the missing denominator

Words complete the effect. A message that says injury risk is high implies a known probability for this athlete, while a message that says this week's load is higher than your recent average describes a fact. The first requires a validated prediction model and the second requires arithmetic. Both fit on a phone screen, and the first sounds more useful, which is exactly the problem. **ATHLEET.AI**

Risk statements also tend to omit their denominator. Telling a young athlete that her risk is elevated says nothing about how common injury is in people like her, how much the risk has risen, or how often such warnings turn out to be wrong. The prediction literature reviewed earlier suggests that, for most published models, those figures are either unknown or unreliable,⁷ **REVIEW** yet the message on the screen gives no hint of that.

Base rates add a further twist that is easily missed. Where injury in a group is fairly uncommon over a given week, even a reasonably discriminating model will issue many more warnings to athletes who stay healthy than to athletes who go on to be injured, simply because the healthy athletes are so much more numerous. This follows from Bahr's distinction between association and prediction,⁶ **REVIEW** and it means that a young athlete who sees a red warning is, on most plausible assumptions, more likely to be fine than not. Nothing on a typical screen says so. **ATHLEET.AI**

5.4 Who reads the screen

The readers of these interfaces are not the performance scientists for whom the underlying metrics were designed. In the club setting they include teenagers, parents who are paying for the product and volunteer coaches with limited time and limited

training in statistics. Such readers are the least likely to discount a confident signal and the most likely to act on it without discussion. **PRACTICE** A previous paper in this strand, *The Measured Child*, discussed the wider questions of consent and autonomy that arise when data about young athletes is collected. Here the narrower concern is what the data is made to say.

Regulation offers some context, though not an answer. On 14 June 2023 the European Parliament adopted its negotiating position on the proposed AI Act, with the final text still to be agreed with member states, and the Parliament describes the rules as aiming to protect health, safety and fundamental rights from harmful effects of AI.²⁰

CONSENSUS In the United Kingdom, the government announced on 7 June 2023 that the country would host a global summit on AI safety.²¹ **CONSENSUS** Neither development, as far as I can see, addresses consumer training apps specifically, and my own view, which goes beyond what either source says, is that clubs and governing bodies should not wait for regulation before setting their own expectations. Whether any given tool falls within a particular regulatory category is a question for qualified advisers.



PART THREE

Designing for honest uncertainty

A proposed framework for building, buying and using load tools with developing athletes.

CHAPTER SIX

Principles for honest load tools

The alternative to false precision is a tool that describes clearly, qualifies its inferences and hands the decision to a person.

The principles below are my own proposal. They draw on the evidence reviewed in earlier chapters, but they have not been tested as a set, and I offer them for challenge by coaches, scientists and designers. **ATHLEET.AI**

6.1 Five design principles

The first principle is to describe before predicting. A tool should show what the athlete did, how it felt and how both compare with her recent history, and it should treat that description as its main product. Trends in external load, session RPE and self-reported recovery rest on the most defensible evidence available,^{16,2} **PEER-REVIEWED REVIEW** and they are what a coach most needs to see.

The second is to label any classification as descriptive and non-clinical. Where a tool groups weeks as higher or lower than usual, the label should say exactly that, in plain words, and it should avoid terms such as risk, danger or warning. The third is to make uncertainty visible, showing ranges, the date that zones or baselines were last set, and the amount of missing data, so that a reader can judge how much weight a figure can bear.

The fourth principle is that the coach owns the decision. Any suggested load should appear as an option for the coach and athlete to consider together, and a tool used by developing athletes should not send prescriptive messages directly to them without that route. The fifth is that no injury-risk claim should appear without validated evidence for the population concerned, published in a form that an independent reviewer can assess. Given the state of the prediction literature,^{6,7} **REVIEW** that principle would at present exclude injury-risk labels for young endurance athletes altogether.

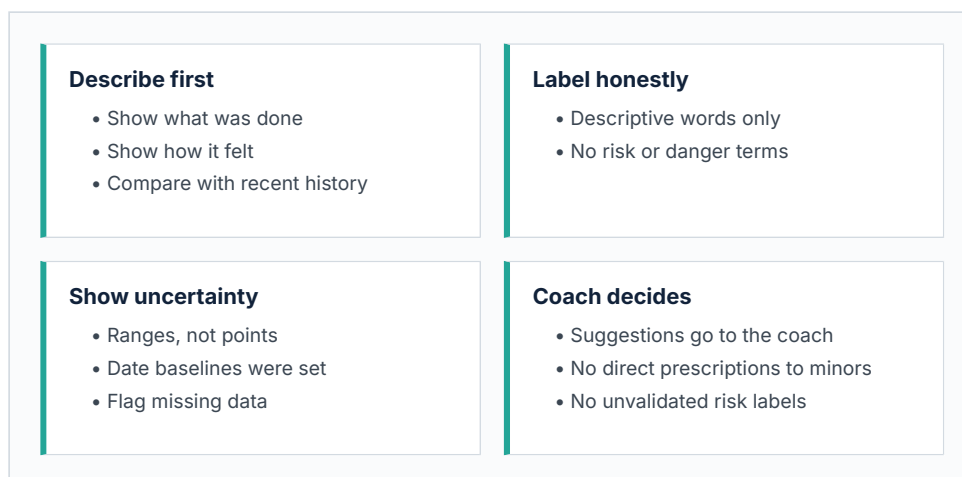
ATHLEET.AI

FIGURE 2 The five proposed principles, grouped for tools used with developing athletes. Source: author's framework.

6.2 What a club or governing body can ask

Most clubs will buy or recommend tools, and few will build them. The principles therefore translate into questions that a club committee or governing body can put to a vendor, or to a parent who has already bought a product. What does each score contain, and how are the inputs weighted? Has any risk classification been validated, on whom, and where is that validation published? How does the tool show uncertainty, and what happens to the display when data is missing? Does the tool send messages directly to athletes under 18, and can a coach see and moderate what they receive?

ATHLEET.AI

Governing bodies could go further by publishing short guidance for volunteer coaches on how to read and discuss load data with young athletes, including what to say when an athlete arrives worried about a colour on a screen. Such guidance would sit comfortably alongside existing consensus positions on youth development, which already stress individualised, developmentally appropriate training.¹⁹ **CONSENSUS** It would also give coaches a basis for telling a parent, calmly, that a ratio is a description of last week and nothing more.

For the individual coach, much of this can start without any policy at all. A coach can ask athletes to share the underlying record alongside any score, can set aside a few minutes each week to talk through how training felt, and can say plainly to athletes and parents that app messages about injury risk are not medical advice and should be brought to the coach before anyone acts on them. **PRACTICE** None of that requires a sports science qualification, and all of it restores the conversation that a coloured square tends to replace. Coaches who work in clubs with a physiotherapist or a medical lead can route genuine concerns about pain or illness there, which is where they belonged in the first place.

6.3 Conclusion and open questions

The argument of this paper is that load tools do most harm when they present contested science as though it were proven, and that in youth endurance sport the harm falls on the readers least able to question what they see. Internal and external load, intensity distribution and simple self-report give coaches a sound basis for describing training. Ratios, risk scores and precise prescriptions do not have comparable support, and individual variation, measurement noise and the long, uneven course of development make it unlikely that they will acquire it soon. **ATHLEET.AI**

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

- Should consumer platforms be permitted to show injury-risk labels to athletes under 18 at all, and who should decide?
- What would honest uncertainty look like on a coach's phone screen, and would coaches and athletes read it as intended?
- Which load metrics are robust enough to automate, and which should remain with the coach?
- How should load tools account for growth and maturation, given how little published evidence addresses that question?

The last of these is the one I find hardest, partly because the research that would answer it would need to follow young athletes for years, which is exactly the timescale that product development cycles tend to ignore.

END MATTER

References

1. Impellizzeri, F.M., Marcora, S.M. and Coutts, A.J. (2019) Internal and external training load: 15 years on. *International Journal of Sports Physiology and Performance* 14(2), 270-273.
2. Saw, A.E., Main, L.C. and Gatin, P.B. (2016) Monitoring the athlete training response: subjective self-reported measures trump commonly used objective measures: a systematic review. *British Journal of Sports Medicine* 50(5), 281-291.
3. Gabbett, T.J. (2016) The training-injury prevention paradox: should athletes be training smarter and harder? *British Journal of Sports Medicine* 50(5), 273-280.
4. Lolli, L., Batterham, A.M., Hawkins, R., Kelly, D.M., Strudwick, A.J., Thorpe, R.T., Gregson, W. and 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.
5. Impellizzeri, F.M., Tenan, M.S., Kempton, T., Novak, A. and Coutts, A.J. (2020) Acute:chronic workload ratio: conceptual issues and fundamental pitfalls. *International Journal of Sports Physiology and Performance* 15(6), 907-913.
6. 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.
7. Bullock, G.S., Mylott, J., Hughes, T., Nicholson, K.F., Riley, R.D. and Collins, G.S. (2022) Just how confident can we be in predicting sports injuries? A systematic review of the methodological conduct and performance of existing musculoskeletal injury prediction models in sport. *Sports Medicine* 52(10), 2469-2482.
8. Van Eetvelde, H., Mendonça, L.D., Ley, C., Seil, R. and Tischer, T. (2021) Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics* 8, 27.
9. Bouchard, C. and Rankinen, T. (2001) Individual differences in response to regular physical activity. *Medicine and Science in Sports and Exercise* 33(6 Suppl), S446-S451.
10. Hecksteden, A., Kraushaar, J., Scharhag-Rosenberger, F., Theisen, D., Senn, S. and Meyer, T. (2015) Individual response to exercise training: a statistical perspective. *Journal of Applied Physiology* 118(12), 1450-1459.
11. Kiely, J. (2012) Periodization paradigms in the 21st century: evidence-led or tradition-driven? *International Journal of Sports Physiology and Performance* 7(3), 242-250.
12. Kiely, J. (2018) Periodization theory: confronting an inconvenient truth. *Sports Medicine* 48(4), 753-764.
13. McKinsey and Company (2023) *The Economic Potential of Generative AI: The Next Productivity Frontier*. McKinsey Global Institute, 14 June 2023.
14. 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.
15. Seiler, S. (2010) What is best practice for training intensity and duration distribution in endurance athletes? *International Journal of Sports Physiology and Performance* 5(3), 276-291.
16. Foster, C., Florhaug, J.A., Franklin, J., Gottschall, L., Hrovatin, L.A., Parker, S., Doleshall, P. and Dodge, C. (2001) A new approach to monitoring exercise training. *Journal of Strength and Conditioning Research* 15(1), 109-115.
17. Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernández, J. and Medina, D. (2018) Effective injury forecasting in soccer with GPS training data and machine learning. *PLOS ONE* 13(7), e0201264.
18. Mujika, I., Halson, S., Burke, L.M., Balagué, G. and Farrow, D. (2018) An integrated, multifactorial approach to periodization for optimal performance in individual and team sports. *International Journal of Sports Physiology and Performance* 13(5), 538-561.
19. Bergeron, M.F., Mountjoy, M., Armstrong, N., Chia, M., Côté, J., Emery, C.A. et al. (2015) International Olympic Committee consensus statement on youth athletic development. *British Journal of Sports Medicine* 49(13), 843-851.
20. European Parliament (2023) *MEPs ready to negotiate first-ever rules for safe and transparent AI*. Press release, 14 June 2023.
21. Prime Minister's Office (2023) *UK to host first global summit on Artificial Intelligence*. Press release, GOV.UK, 7 June 2023.

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

Load, Readiness and False Precision

Athleet.AI Working Papers, AI in Coaching and Athlete Development
Paper five of six · 21 June 2023