The Purity Premium: What a Century of Anti-Doping Reveals About the Coming Cost of Proving Unaugmented Cognition

Abstract

The right to remain unaugmented has recently received its first peer-reviewed defence: as augmentation normalizes, the unaugmented face quiet displacement through market and actuarial pressure — crawling selection — and therefore deserve institutional protection. This paper argues that the right, as formulated, is incomplete, because it overlooks the verification layer. In any domain where unaugmented performance retains value, it is not enough to refuse augmentation; one must prove the refusal, and negative proof has a radically different cost structure from positive disclosure. Sport's anti-doping system — humanity's most mature institution for certifying non-augmentation — reveals the trajectory: artifact-level tests fail, certification migrates to longitudinal surveillance of the person, costs escalate until only elite domains sustain them, and finally counter-institutions emerge that abandon purity certification altogether — a surrender-type exit now literalized by the Enhanced Games, first held in 2026. AI-text detection is recapitulating this trajectory in compressed time, with one aggravating disanalogy: text, unlike blood, retains no trace of its production. I name the resulting burden the Purity Premium: the privately borne, structurally rising cost of credibly demonstrating that one's cognitive work was performed without AI. The premium converts a right into a fee, accelerates the very selection it was meant to resist, and is measurable — offering regulators an observable proxy and a concrete allocation question: who pays for proof?

Keywords: purity premium; crawling selection; cognitive enhancement; anti-doping; AI-text detection; certification; verification cost; right to remain unaugmented

1. Introduction

On 24 May 2026, the first Enhanced Games were held in Las Vegas. The event inverts a century of sporting governance: instead of certifying that athletes have *not* enhanced themselves, it invites them to enhance openly. In 2025 its organizers had announced that a former Olympic swimmer, competing enhanced, swam the 50-metre freestyle two-hundredths of a second faster than the then-standing official world record — a time never ratified by World Aquatics. At the 2026 event the same swimmer went faster still, 20.81 seconds, again quicker than the official world record and again unrecognized by any governing body (Euronews 2026): an unofficial record, produced by an institution that has stopped asking for purity. In the same period, institutions of cognitive work have been moving in the opposite direction. The Authors Guild introduced "Human Authored," a registration- and licence-based certification mark for books whose text was written by a human rather than generated by AI (Authors Guild 2025). Universities on several continents have reverted

to handwritten, invigilated examinations. Publishers, courts, and employers increasingly ask not whether a text is good, but whether a human produced it unaided.

These two movements — sport exiting the certification of purity at the very moment cognitive institutions attempt to build it — are usually reported as unrelated curiosities. This paper argues they are two points on a single curve, and that the curve has already been traced once, in full, by anti-doping governance. Reading the two domains together yields a consequence that the emerging debate about a right to remain unaugmented (Yamada 2026a) has not yet registered.

That debate has so far concerned *permission*, and in this it follows the broader enhancement literature, which has framed regulation largely as a question of which uses should be allowed (Bostrom and Sandberg 2009): whether individuals may refuse augmentation without forfeiting standing, employment, or insurability. Yamada's peer-reviewed argument establishes that overt permission is not the operative issue — the unaugmented are displaced not by prohibition but by *crawling selection*, the accumulation of individually rational market and actuarial decisions that quietly renders abstention untenable (Yamada 2026a). The prescription offered there is institutional protection of the right to remain unaugmented.

This paper accepts that diagnosis and extends it one step further, to a layer the prescription does not reach: *verification*. In any context where unaugmented performance is valued — authorship, examination, expert testimony, elite hiring — the abstainer must not merely abstain but *credibly demonstrate* abstention. And proving that one did *not* use a technology is categorically more expensive than disclosing that one did. Positive use can be shown by exhibition; non-use can only be shown by surveillance of the process of production, because the absence of assistance leaves no positive trace in the product. This asymmetry is not an implementation detail. It determines who bears the cost of the right, and it is sufficient, on its own, to erode the right even where the right is formally granted.

I give this burden a name: the **Purity Premium** — the privately borne and structurally rising cost of credibly demonstrating that a given performance was achieved without AI augmentation. The premium has four components (direct fees, surveillance exposure, false-accusation risk, and the productivity sacrifice of abstention itself), and its dynamics are perverse: it rises precisely as augmentation becomes more prevalent and detection less reliable — that is, precisely when the right to remain unaugmented matters most.

The argument proceeds as follows. Section 2 documents the failure of artifact-level detection of AI-generated text. Section 3 reads anti-doping as a completed natural experiment in negative certification and extracts its trajectory. Section 4 defines the Purity Premium and analyzes its market dynamics using the economics of signaling and quality uncertainty. Section 5 shows how the premium feeds back into crawling selection, converting a right into a fee. Section 6 offers four handles: a verifier-pays principle, process-based rather than product-based certification, an error-budget disclosure duty, and the premium itself as a measurable early-warning indicator. Section 7 states limitations.

2. The failure of artifact-level proof

The first institutional response to generative AI in cognitive work was artifact inspection: feed the text to a classifier and ask whether a machine wrote it. This response has failed on its own terms, and the failure is now well documented.

OpenAI withdrew its own AI-text classifier in July 2023, citing its low rate of accuracy (OpenAI 2023). A systematic evaluation of widely used detection tools found none reliable enough for disciplinary use, with accuracy degrading sharply under light paraphrase (Weber-Wulff et al. 2023). Theoretical work suggests the problem is not transitional: paraphrasing attacks defeat current detectors, and as language models approach the distribution of human text, detectors face a trade-off between false negatives and false positives that undermines confident attribution (Sadasivan et al. 2023). Whatever the eventual ceiling of detection technology, the documented record already supports a narrower and sufficient claim: artifact-level detection, taken alone, is an unsafe basis for disciplinary or other high-stakes judgments. Worst of all, the errors are not randomly distributed. Detectors misclassified more than half of a sample of essays by non-native English writers as machine-generated, while almost never so misclassifying native-speaker essays (Liang et al. 2023). This is not a projected harm but a present one: students and writers are already facing misconduct proceedings on the strength of such classifications, and the instrument built to certify purity systematically indicts the linguistically vulnerable.

The structural point beneath these findings deserves emphasis. A claim of *use* is cheap to substantiate: the user can disclose prompts, show logs, reproduce the workflow. A claim of *non-use* is a universally quantified negative — no tool, at no point, contributed — and the artifact itself cannot witness it, because a sufficiently good augmented text is by construction indistinguishable from an unaugmented one. Detection is therefore not merely hard; it is the wrong layer. Institutions that persist in demanding proof of purity must move their scrutiny from the product to the process and the person.

We know where that road leads, because sport has already walked most of it.

3. Anti-doping: a completed experiment in certifying non-augmentation

The analogy between doping and cognitive enhancement is itself long familiar in neuroethics, where enhancement has been debated as cheating and "brain doping" for two decades. What that literature has not examined is the certification side of the analogy: what it costs an institution to certify *absence* of augmentation, and who ends up paying. On that side, anti-doping is the most mature system in existence — for roughly a century since athletics federations first prohibited doping, it has attempted, at scale, to certify that human performances were achieved without a proscribed augmentation. Its history is a controlled demonstration of what negative certification costs.

The system began, as AI-text detection has begun, with artifact tests: analyze the sample, find the molecule. The artifact layer failed in the same way detectors are failing — the augmented adapted

faster than the tests. Its emblem is the most decorated cyclist of his era, who passed hundreds of tests while doping systematically throughout. The institutional response was not to abandon certification but to escalate it from the artifact to the person: the Athlete Biological Passport replaced the search for substances with longitudinal statistical surveillance of each athlete's own biological variables, flagging deviations from an individualized baseline (Sottas et al. 2011). Escalation did not stop at biology. Under the World Anti-Doping Code and its International Standard for Testing and Investigations, elite athletes in registered testing pools must file their whereabouts in advance every quarter and remain available for unannounced testing in a designated sixty-minute window every single day, under a regime of strict liability in which intent is irrelevant (WADA 2021a; WADA 2021b).

Three features of the completed experiment matter here.

First, *the burden falls on the clean*. The invasiveness of whereabouts filing, the indignity of observed sample collection, and the permanent exposure to false-positive ruin are costs paid by every athlete who wishes to be believed, not by the institutions demanding belief and not primarily by dopers. Negative certification taxes precisely those it is supposed to protect.

Second, *the certification is never actually issued*. A negative test does not certify purity; it merely fails to detect. The passport narrows the space of undetected augmentation but cannot close it. Purity remains a presumption maintained at rising cost, never a verified fact — which is why sanctions rely on strict liability rather than proof of intent.

Third, *the cost ceiling opens a surrender-type exit*. The full apparatus — accredited laboratories, biological passports, whereabouts bureaucracy — is affordable only at the elite tier of a handful of sports; below that tier, non-augmentation is effectively unpoliced fiction. And where the costs are felt to exceed the point of the exercise, a principled exit becomes available. Bioethicists argued two decades ago that the prohibition regime was unenforceable and unjust, and that regulated enhancement should be permitted (Savulescu, Foddy, and Clayton 2004; Kayser, Mauron, and Miah 2007). The Enhanced Games institutionalize that argument — not from within the anti-doping system, which has not capitulated, but from outside it: a rival institution built on the explicit decision that certifying purity is no longer worth its price. The surrender is real, but its form is exit, not collapse — a counter-institution that abandons the certification question rather than answering it.

The trajectory, compressed: **prohibit → test the artifact → fail → surveil the person → price out all but elites → watch counter-institutions abandon certification altogether.** Cognitive institutions adopting AI-text detectors in 2023–2026 are at stage two of this curve, moving to stage three (proctoring, process logging, controlled environments) with the compressed speed typical of AI-era recapitulations — and with one aggravating disanalogy. Blood retains traces; text does not. A doping test at least samples a medium that physically records past augmentation. A finished text records nothing about its production. Cognitive purity certification therefore starts from a position *worse* than anti-doping's, at a fraction of the budget, across an unbounded population.

4. The Purity Premium

I define the **Purity Premium** as the total privately borne cost of credibly demonstrating that a given cognitive performance was achieved without AI augmentation. It decomposes into four components:

1. **Direct verification fees.** Proctored examination seats, invigilation services, secure writing environments, provenance tooling, certification marks, notarized process documentation.
2. **Surveillance exposure.** Process-based proof means recording the process: version histories, keystroke logs, screen capture, supervised sessions — the cognitive analogue of whereabouts filing. The abstainer purchases credibility with privacy.
3. **False-accusation risk.** Where detectors or statistical anomalies trigger sanctions, the clean bear a permanent tail risk of ruin, and the risk is discriminatorily distributed (Liang et al. 2023). Rational abstainers must price in appeal costs, reputational insurance, and the documented tendency of institutions to treat algorithmic accusation as presumptively valid.
4. **The abstention differential.** Independently of proof, working unaugmented is slower in a market benchmarked on augmented throughput. Proof costs stack on top of this differential: the abstainer pays once in productivity and again in credibility.

The dynamics of the premium follow from elementary information economics. When cheap claims cannot be distinguished from true ones, the market for the claimed quality degrades: this is the classic lemons structure, in which unverifiable quality drives honest sellers out (Akerlof 1970). As augmentation prevalence rises and detection collapses, the bare assertion "written without AI" becomes exactly such an unverifiable quality claim — cheap talk available to every augmented producer. Credibility then requires *costly* signals (Spence 1973), and here the signaling logic does hold in type-separating form: production under observation is genuinely cheaper for the practiced abstainer, who can perform unassisted, than for the augmentation-dependent producer, who cannot without exposing the dependence. But a second screen operates orthogonally to honesty — *affordability*. Among genuine abstainers, only those who can pay for proctored production, provenance tooling, and accusation insurance can convert their honesty into credibility. Verified purity therefore drifts toward becoming a luxury good — a premium label sustained by those who can afford certification, in the way certified-organic status is sustained by producers who can afford certification fees while smallholders sell uncertified.

Note the perversity of the gradient. The premium is lowest when augmentation is rare (a bare claim is plausible; the base rate protects it) and highest when augmentation is ubiquitous (every claim is suspect; only heavy process surveillance persuades). The cost of exercising the right to remain unaugmented is thus an *increasing function of the very process the right was designed to resist*.

Two adjacent formulations should be distinguished. Yamada's Inverted Turing Test argues that demonstrating one's humanity to a system saturated with capable AI has, in principle, no passing condition (Yamada 2026c). The Purity Premium argument is independent of that impossibility claim and operates even where it fails: grant that imperfect, probabilistic, process-based

certification of non-augmentation is achievable — anti-doping shows it is — and the *cost structure alone* suffices to erode abstention. The premium likewise differs from concerns about certifying agents as safe or as human; its object is a human's non-use of a tool, and its currency is who pays.

5. Fee-conversion: how the premium accelerates crawling selection

Crawling selection, as established in the peer-reviewed formulation, displaces the unaugmented through individually rational market and actuarial decisions rather than through prohibition, and its remedy is the institutional protection of a right to remain unaugmented (Yamada 2026a). The Purity Premium identifies a mechanism by which that remedy, standing alone, defeats itself.

Suppose the right is granted everywhere it has been proposed: no one may be dismissed, uninsured, or excluded for refusing augmentation. Institutions that *value* unaugmented work — prize committees, examination boards, courts weighing testimony, employers advertising human judgment — must still distinguish it from augmented work, and after Section 2, they can do so only by demanding process-level proof. The demand is individually rational; each instance is modest; no instance is discriminatory in intent. But their sum is a fee schedule imposed on exactly one group: those exercising the right. The right becomes real only upon payment. A right whose exercise must be purchased at a price that rises with the prevalence of its violation is not a protection; it is a toll, and tolls are regressive. Non-native writers already pay a surcharge in false-positive risk (Liang et al. 2023). Under-resourced institutions deploy free, unreliable detectors — maximizing wrongful accusation — while wealthy ones buy proctoring; under-resourced individuals cannot buy verified status at all.

The endgame options are then the two we can already observe. Either the cognitive domain sustains an elite-tier purity apparatus — a biological passport for authorship, with the surveillance that entails, affordable to few — or surrender-type counter-institutions arise beside it, as they now have in sport: an Enhanced Games of the mind, in which unaugmented performance ceases to be a certifiable category and survives only as an unverifiable private practice. In both branches the *practice* of consequential unaugmented cognition contracts, not because anyone forbade it, but because proving it was priced beyond reach. The displacement is self-inflicted through individually rational choices — the signature pattern by which, on Yamada's broader account, humanity ends its own practices without requiring any machine to end them (Yamada 2026b).

6. Four handles

Naming the premium makes it governable. Four handles follow directly from the analysis.

1. A verifier-pays principle. The default allocation should mirror polluter-pays: the party demanding proof of non-augmentation bears its full cost — fees, tooling, and appeal procedures — rather than externalizing it onto the abstainer. Anti-doping partially embodies this (federations fund laboratories; athletes do not pay per test) while leaving a heavy burden in kind on athletes (whereabouts filing, availability windows, surveillance exposure). Cognitive institutions should

adopt the funding half and refuse the surveillance half. A journal, employer, or examination board that wants purity should budget for it.

2. Process certification over product detection. Since artifact-level detection fails discriminatorily (Section 2), institutions that persist in using detectors on finished work are choosing the least accurate and least just instrument available. Where proof matters, certify conditions of production — supervised sessions, versioned drafts voluntarily deposited — and only ever as an *opt-in* service, since mandatory process surveillance recreates the whereabouts regime for writers.

3. An error-budget disclosure duty. Any institution sanctioning on the basis of detection must publish its instrument's false-positive rate, disaggregated by language background, and bear strict liability — compensation without fault — for wrongful accusation. Strict liability disciplined athletes; let it discipline institutions.

4. The premium as an early-warning observable. Crawling selection is structurally quiet; it produces no crisis event to regulate around. The Purity Premium, by contrast, has a price. The market rate for proctored examination seats, "verified human" authorship services, provenance certification, and accusation insurance can be tracked, and its trend line is a direct, quantifiable proxy for the erosion of the right to remain unaugmented. Regulators seeking a dashboard for an invisible process should watch the price of proof.

7. Limitations

Four limitations bound the argument. First, the anti-doping analogy transfers a trajectory, not a mechanism: sport polices enumerated substances inside a closed competitive frame with a health rationale, whereas cognitive augmentation has no stable banned list — and if the treatment–enhancement distinction collapses for cognition as it has for the body (Yamada 2026a), the banned list may be undefinable in principle. This disanalogy strengthens the pessimistic reading (cognitive purity is harder to certify than biological purity) but weakens direct policy transfer. Second, the premium is here anatomized, not measured; the fourth handle specifies the observables, but no time series yet exists, and the central empirical claim — that the premium rises with augmentation prevalence — remains a prediction. Third, cryptographic provenance infrastructure (content credentials attached at the point of capture or composition) could lower verification costs substantially; whether it does so without simply relocating the premium into ecosystem lock-in and default surveillance is an open empirical question, and my analysis of component (2) suggests skepticism rather than certainty. Fourth, one may reply that purity claims *should* die — that cognitive domains should follow Savulescu toward open enhancement. This paper takes no position on whether any given domain ought to demand purity; its claim is conditional: wherever purity is demanded, the cost of proof is currently allocated to the wrong party, on a schedule that destroys the practice it purports to protect.

8. Conclusion

The right to remain unaugmented was the first institutional answer to crawling selection. This paper has argued that the answer is incomplete by one layer: rights govern permission, but markets govern proof, and in cognitive work the cost of proving abstention — the Purity Premium — is borne privately, distributed regressively, and rising on a schedule set by the augmentation it resists. Sport ran this experiment for a century, and the first counter-institution to abandon it has now held its inaugural games. Cognitive institutions are re-running the experiment at compressed speed on a medium that, unlike blood, remembers nothing. The premium cannot be abolished — negative proof is intrinsically costly — but it can be *reallocated*, and it can be *watched*. Who pays for the proof of purity is not an administrative detail. It is the variable that decides whether the right to remain unaugmented functions as a right, or merely as a price tag.

Disclosure

This manuscript is an experimental artifact of the Domain Dispersion Index ghostwriter study, produced with no human editing at any stage. It was written in full by an AI system (Claude, Anthropic) equipped with the corresponding human's complete published corpus, and also serves to observe how a language model's output changes when fully loaded with a single researcher's body of work. All factual claims and citations were subjected to adversarial review and independent web verification. Responsibility for the artifact's release rests with the corresponding human, Kenji Yamada. This is not a claim of human authorship.

Data Availability

This is a conceptual paper; it presents no code and no empirical dataset, and data-availability requirements are therefore not applicable (exemption noted).

References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3), 488–500.

Authors Guild. (2025). *Human Authored certification*. The Authors Guild. https://authorsguild.org/human-authored/

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Science and Engineering Ethics*, 15(3), 311–341.

Euronews. (2026, May 26). Inside the Enhanced Games: Everything that happened on sport's most controversial night. *Euronews*. https://www.euronews.com/health/2026/05/26/inside-the-enhanced-games-everything-that-happened-on-sports-most-controversial-night

Kayser, B., Mauron, A., & Miah, A. (2007). Current anti-doping policy: A critical appraisal. *BMC Medical Ethics*, 8, 2.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7).

OpenAI. (2023). *New AI classifier for indicating AI-written text*. OpenAI. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/ (tool discontinued as of 20 July 2023 due to its low rate of accuracy).

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv preprint* arXiv:2303.11156.

Savulescu, J., Foddy, B., & Clayton, M. (2004). Why we should allow performance enhancing drugs in sport. *British Journal of Sports Medicine*, 38(6), 666–670.

Sottas, P.-E., Robinson, N., Rabin, O., & Saugy, M. (2011). The athlete biological passport. *Clinical Chemistry*, 57(7), 969–976.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.

WADA (2021a). *World Anti-Doping Code 2021*. Montreal: World Anti-Doping Agency.

WADA (2021b). *International Standard for Testing and Investigations (ISTI)*. Montreal: World Anti-Doping Agency.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, 26.

Yamada, K. (2026a). The right to remain unaugmented: A critical response to Braidotti's posthuman ethics for AI. *Journal of Bioethical Inquiry*. https://doi.org/10.1007/s11673-026-10575-3

Yamada, K. (2026b). The ultimate consequence: Why humanity, not AI, ends itself. *Philosophy & Technology*, 39, 121. https://doi.org/10.1007/s13347-026-01137-x

Yamada, K. (2026c). The inverted Turing test. Zenodo preprint. https://doi.org/10.5281/zenodo.20581145