The Social Credit of Text: When Algorithms Judge Authorship

Weekly Analysis — https://ainews.social

Somewhere in the architecture of a system you will never see, a number is being assigned to a sentence you wrote. The number estimates the probability that a machine, rather than you, produced it. You did not consent to the calculation; you may never learn its result. But if the number crosses a threshold, things happen — an essay is flagged, an application is sorted into a lower pile, a parole assessment acquires a note, a protest leaflet is correlated against a watchlist. This is the quiet premise of a technology that has spread far past the classroom where most people first heard of it: that authorship is now a thing algorithms can adjudicate, and that their verdicts can be trusted enough to act upon. The premise deserves more suspicion than it is getting.

What I want to call the social credit of text is the slow extension of authorship-detection from a cheating-detection gadget into an infrastructure of judgment — one that increasingly governs hiring, policing, migration, and public speech. The detection industry sells confidence: a percentage, a verdict, a clean binary between human and synthetic. But the confidence is manufactured, and the cost of believing it falls unevenly. When a Palo Alto family filed suit after their son was accused of using AI to cheat, the dispute was framed as a school-discipline matter [1], but the underlying mechanism — a probabilistic guess about who authored a text, treated as fact — is the same mechanism now being piloted in contexts where the stakes are liberty and livelihood rather than a grade. The right question is not whether these tools work. It is who built them, whose writing they treat as suspicious, and who answers when they are wrong.

[1] A Palo Alto high schooler was accused of AI cheating. His family filed ...

This essay is about power: about who shapes the discourse around algorithmic judgment of authorship, whose voices are amplified in it, and whose are structurally absent. The false-positive problem — the wrongful flagging of genuinely human work — is usually discussed as a technical bug to be patched. I want to argue that it is something else. It is a preview of a society in which deviation from a statistically average prose style becomes grounds for suspicion, and in which the people whose writing deviates most are precisely those least able to contest the verdict.

The Verdict Machine and Its Makers

Begin with the makers, because the discourse rarely does. The companies building authorship-detection and behavioral-surveillance systems are not neutral instrument-makers; they are commercial and state actors with a stake in your believing their output. Amnesty International has documented how technology built by Palantir and Babel Street creates surveillance threats aimed squarely at pro-Palestine student protesters and migrants, scraping and correlating online speech to build profiles that travel into immigration and policing decisions [19]. The same machinery, in France, has been deployed to target migrants and foreign students [10]. These are not detection tools in the narrow plagiarism sense, but they share its logic: text and behavior are run through a model, a risk score emerges, and the score becomes the basis for consequence.

[19] USA/Global: Tech made by Palantir and Babel Street pose surveillance ...
[10] L'IA utilis00e9e pour cibler les migrants et les 00e9tudiants 00e9trangers aux ...

Police adoption of these systems is outrunning the rules meant to govern it. Law enforcement agencies are folding AI into their operations faster than legislatures can write standards for accuracy, auditing, or redress, so that the tools arrive already wired into decisions before anyone has decided whether they should [14]. This sequencing — deployment first, governance later, if ever — is itself an exercise of power. It means the people who build and sell the systems set the terms of the debate, because by the time critics arrive, the infrastructure is already load-bearing. Kate Crawford names the asymmetry precisely in [4]: tech companies rarely suffer serious financial penalties when their systems violate the law, and fewer still when they merely violate the ethical principles they themselves authored. The verdict machine, in other words, is built by parties who face almost no downside when it errs.

[14] Police use of artificial intelligence grows as rules lag behind

[4] The Atlas of AI

Notice what the standard framing of "AI detection" hides. It presents the technology as a referee — a disinterested third party that simply reports what is true about a text. But classification, as Crawford insists, is never disinterested. "Classification is an act of power," she writes in [4], and embedded in working infrastructure it becomes invisible without losing any of that power. To label a passage of writing "likely AI-generated" is to perform an act of categorization that carries the full weight of whatever consequence is attached to the category — and to do so while disclaiming responsibility, because the model "just" calculated a probability. The mystification is the product. What is being sold is not detection but the right to act on a guess while calling it a finding.

[4] The Atlas of AI

What "Statistically Normal" Erases

Every detection model encodes a picture of what ordinary human writing looks like, and that picture is built from data. The data are not neutral. They overrepresent some kinds of writers and underrepresent others, and the model inherits those proportions as a definition of the normal. The consequence is not abstract. When researchers examined how an automated system scored student essays, Asian American students lost more points than their peers for writing that diverged from the model's expectations — a pattern of penalty applied not to the quality of the argument but to the shape of the prose [15]. The system was not built to discriminate. It discriminated because it was built to reward statistical typicality, and typicality is a demographic fact before it is a linguistic one.

[15] PROOF POINTS: Asian American students lose more points in an AI essay ...

This is the heart of the false-positive problem, and it is why "false positive" is too clean a phrase. A false positive in an authorship detector is not a random error scattered evenly across the population. It clusters. It lands on writers whose first language is not English, whose rhetorical traditions differ from the training corpus, whose syntax carries the fingerprint of a different educational or cultural formation. Researchers have repeatedly warned that AI tools deployed in classrooms carry exactly this potential for racial bias, precisely because the "normal" they enforce is the normal of the dominant group [17]. The peer-reviewed literature on algorithmic bias in education has reached the same conclusion from multiple angles: systems trained to recognize patterns reproduce the inequities baked into the patterns they were shown [8].

[17] Researchers Warn of Potential for Racial Bias in AI Apps in the Classroom

[8] Algorithmic Bias in Education | International Journal of Artificial ...

Ruha Benjamin gave this dynamic its sharpest name. In [4], she describes how systems that appear neutral, even benevolent, can deepen discrimination precisely because their bias is encoded as technical fact rather than human prejudice — what she calls the New Jim Code. An authorship detector that flags non-standard English as suspicious does not announce itself as a tool of exclusion. It announces itself as a measurement. And the measurement's authority depends on its users forgetting that "standard" was a choice. The detector that penalizes the Asian American student is not malfunctioning; it is functioning exactly as designed, enforcing a norm and treating distance from that norm as evidence of fraud.

[4] Race After Technology

The discrimination compounds when the judgment is about not just the text but the person who admits to using AI at all. A striking study found that women who use AI tools at work are perceived as less competent, while men using the same tools are read as pragmatic and efficient [20]. Layer this onto a detection regime and the asymme-

[20] Women Who Use AI Seen As Incompetent; Men Who Use AI Seen As Pragmatic

try doubles: the same disclosed AI assistance is excused in one writer and condemned in another, and a flag that says ”machine-influenced” means different things depending on who is wearing it. The number does not interpret itself. It is interpreted through the social hierarchies already in the room.

Liberty, Livelihood, Reputation

The reason any of this matters beyond the seminar is that the consequences attached to these judgments are escalating. Consider hiring, where authorship and aptitude judgments by algorithm have already moved from pilot to practice. The largest study to date of AI hiring algorithms found clear racial disparities in how candidates were scored — not at the margins but as a systematic feature of how the systems sorted applicants [12]. The cost of a false judgment here is a livelihood: a job not offered, an application never seen by human eyes, a career quietly redirected by a model that no applicant can interrogate. This is the same failure mode that Janelle Shane catalogs in [4], where an Amazon résumé-screening tool taught itself that the strongest predictors of a good candidate were being named Jared and having played lacrosse — a parody of merit that the machine treated as wisdom until engineers caught it.

[12] Largest study of AI hiring algorithms to date finds ’clear racial ...

[4] You Look Like a Thing and I Love You

Move to the carceral context and the cost becomes liberty itself. When detection and surveillance systems feed into policing and parole, a false correlation is not a sorting error; it is a person held, watched, or denied release on the strength of a number. School surveillance offers a preview of how this plays out at scale. Systems like Gaggle, which monitor students’ words for signs of risk, have generated false alarms that escalated into police contact and arrests — children swept into the criminal-legal system because an algorithm misread a text [18]. The architecture that flags a student’s journal entry as a threat is continuous with the architecture that would flag a parolee’s written statement as deceptive. The same probabilistic guess, the same disclaimed responsibility, the same human being left to absorb the consequence.

[18] School AI surveillance like Gaggle can lead to false alarms, arrests ...

Reputation is the third casualty, and it is the one the detection industry most readily produces. A growing body of litigation has formed around students wrongly accused of AI cheating, and the documented cases reveal how flimsy the underlying evidence often is — a detector’s percentage presented as proof, a teenager forced to prove a negative [3]. The legal record is, tellingly, beginning to break in the accused’s favor on procedural grounds: courts have started to hold that an al-

[3] AI Detection Lawsuits: Every Student Case, Outcome, and What the Data ...

gorithmic flag is not, by itself, due process [2]. That is a meaningful check — but notice that it arrives only after harm, only for those with the resources to litigate, and only in a domain where the accused has standing to sue. The migrant profiled by Babel Street, the applicant screened out by a hiring model, the parolee assessed by a risk tool — these people have no comparable forum. The reputational damage in their cases is administered quietly and contested almost never.

[2] AI Cheating Lawsuit Myths 2014 Students Win on Due Process

It is worth naming the institutional response honestly. The dominant reaction across schools and agencies has been to adopt the tools first and write the rules afterward, if at all. AI use is growing across institutions while the policies meant to govern it lag conspicuously behind [7]. Advocacy groups documenting surveillance and detection in schools have had to assert something that should be obvious — that the people subjected to these systems have rights that the systems routinely ignore [5]. The pattern generalizes: the convenience accrues to the institution, the risk accrues to the judged, and the gap between adoption and accountability is where the harm lives.

[7] AI Use in Schools Growing, but District Policies Haven't Caught Up

[5] AI in Schools: Surveillance, Detection, and Student Rights - CPAI

The Discourse of Documented Harm

Here is the strange thing about the conversation surrounding all of this. We are not short of harm documentation. A remarkable share of the public discourse on AI and society — by my reading of this stretch of coverage, the largest single share — consists of cataloging ethical failures: bias found, discrimination measured, surveillance exposed, false alarm tallied. The peer-reviewed warnings exist [8]. The investigative journalism exists [18]. The human-rights reporting exists [19]. What is missing is not the diagnosis. What is missing is any commensurate movement toward building the systems that would prevent the harm, or toward giving the harmed the power to refuse it.

[8] Algorithmic Bias in Education | International Journal of Artificial ...

[18] School AI surveillance like Gaggle can lead to false alarms, arrests ...

[19] USA/Global: Tech made by Palantir and Babel Street pose surveillance ...

This imbalance is itself a fact about power. The dominance of ethical-failure documentation tells us who is allowed to participate in the discourse and in what role. Critics, researchers, and journalists are permitted to describe the damage at length; they are rarely positioned to set the standards that would constrain it. The standard-setting happens elsewhere — inside the vendors who write their own ethical principles, inside the agencies that adopt before they regulate. Crawford's observation in [4] that companies face few consequences even when they violate their own stated principles explains why the documentation can pile up indefinitely without producing change. Documentation that carries no enforcement mechanism is not a brake. It is a record kept by people who do not hold the controls.

[4] The Atlas of AI

Meredith Broussard's account of the technologists' worldview in [4] sharpens the point. After the 2016 election, she recalls, the people building these systems were not surprised that misinformation flooded their platforms; they were surprised that anyone took it seriously — "since when did people start believing everything on the Internet is true?" The reflex is to treat the harms as the users' problem, a failure of credulity on the part of the people downstream rather than a failure of design upstream. Transposed to authorship detection, the same reflex says: if a false positive ruined your week, the fault lies in your unusual prose, or in your having used the tool at all, not in the detector that misjudged you. The structural silence in the discourse is the silence of those downstream voices — the writers flagged, the applicants screened, the protesters surveilled — who appear in the coverage as data points but almost never as authorities on what should be done.

[4] Artificial Unintelligence

There is a deeper silence still, which the global pattern of who builds these systems makes visible. The judgments that get rendered on text in the wealthy world depend on labor performed cheaply elsewhere. Hundreds of thousands of workers in poorer countries perform the data labeling and content moderation that train and tune these models, doing the invisible classification work that makes the verdict machine appear automatic [13]. The dependency runs deeper than wages. Analysts of digital colonialism describe how the entire stack — data, infrastructure, standards — is arranged so that the Global South supplies raw material and absorbs risk while the power to define what counts as legitimate output stays in the metropole [9]. Even the more optimistic policy literature on AI in the Global South concedes that inclusive governance remains aspirational, that the people most affected by these systems have the least say in how they are built and ruled [6]. The "statistically normal" writer the detector defends is, in the end, a writer from a particular place, validated by labor extracted from people whose own writing the same systems are most likely to flag.

[13] Los cientos de miles de trabajadores en pa00edses pobres que hacen ... - BBC

[9] El colonialismo digital en la era de la IA: siete dimensiones ... - Medium

[6] AI in the Global South: Opportunities and challenges ... - Brookings

Who Pays, Who Profits, Who Escapes

Follow the accountability and you find it flowing in exactly the wrong direction. When an authorship detector errs, the burden of proof lands on the accused. The student must demonstrate that they wrote their own essay; the applicant must somehow show that a screening model misjudged them; the migrant must contest a profile assembled from scraped speech. The institutions that deploy these tools, and the vendors that sell them, are insulated at every step — by the opacity of

the models, by the disclaimer that a probability is not an accusation, by the simple fact that the cost of error is externalized onto people without the standing to bill it back. This is the structural signature of the New Jim Code that Benjamin warns about in [4]: a system that distributes suspicion downward and immunity upward, all while presenting itself as the impartial arithmetic of merit.

[4] Race After Technology

The regulatory exceptions prove the rule by their rarity. When Spain's data-protection authority sanctioned the processing of biometric data by AI, it was a genuine instance of a state actor imposing a real consequence on a deployer [11]. But such interventions are scattered and reactive, arriving after deployment and covering only fragments of the field. They demonstrate that accountability is possible while underscoring how exceptional it remains. For every sanctioned biometric system there are countless detection and surveillance tools running unaudited, their error rates unpublished, their training data undisclosed, their false positives absorbed silently by the judged.

[11] La AEPD sanciona el tratamiento de datos biom00e9tricos con IA en la ...

The integrity of the discourse is itself now under threat from the technology it describes, which closes the loop in an unsettling way. Researchers studying online behavior have had to develop methods to detect "LLM pollution" — the contamination of supposedly human research data by AI-generated responses [16]. The irony is exact: the same generative systems that make authorship detection seem necessary are degrading the data on which detection's own claims of accuracy must rest. We are building verdict machines to police a boundary that the machines themselves are erasing, and we are doing it with growing confidence even as the ground shifts under the confidence. The people profiting from selling certainty have no incentive to disclose how unstable that certainty has become.

[16] Recognising and mitigating LLM Pollution in online behavioural research

So who escapes? The vendors, who write their own principles and face few penalties for breaking them. The agencies, which adopt before they regulate and treat the resulting harms as the cost of doing business. The technologists, who locate the failure in the credulous downstream user rather than the upstream design. And who pays? The writer whose prose is too unusual, the worker whose disclosed AI use is read through a hierarchy that condemns some and excuses others, the protester whose speech is correlated against a watchlist, the labeler whose invisible work makes the whole apparatus run. The gap between who is harmed and who speaks is not a flaw in the discourse. It is the discourse's organizing principle, and naming it is the first condition of changing it.

The Standard We Have Not Demanded

What would it mean to take the false-positive problem seriously — not as a bug awaiting a patch, but as the warning it is? It would mean refusing the central move of the detection industry: the conversion of a probability into a verdict, performed while disclaiming responsibility for the verdict's consequences. It would mean insisting that any system permitted to judge human output meet evidence-based standards before deployment, not after litigation — published error rates, disaggregated by the very groups the systems are most likely to misjudge, audited by parties with no stake in the result. The early legal victories on due-process grounds point the way [2], but a society cannot litigate its way to fairness one wronged individual at a time, least of all when most of the wronged have no standing to sue.

[2] AI Cheating Lawsuit Myths 2014 Students Win on Due Process

The deeper demand is about who holds the controls. The discourse, as it stands, lets the harmed appear as evidence and the powerful appear as decision-makers, and it treats that arrangement as natural. It is not natural. The labelers in poorer countries who train these systems [13], the students whose unusual prose is flagged as fraud [15], the migrants and protesters whose speech is scraped into watchlists [19] — these are the people with the most precise knowledge of how the verdict machine fails, and they are the people the discourse most thoroughly excludes from the room where its rules are written. A society that placed their testimony at the center would build very different systems, or build far fewer of them.

[13] Los cientos de miles de trabajadores en pa00edses pobres que hacen ... - BBC

[15] PROOF POINTS: Asian American students lose more points in an AI essay ...

[19] USA/Global: Tech made by Palantir and Babel Street pose surveillance ...

The social credit of text is not yet fully built. That is the only good news, and it is real. The tools are spreading, the policies lag, the accountability flows upward — but the infrastructure is not yet so embedded that we have forgotten it was a choice. The MIT account of these dilemmas in [4] keeps returning to a question that the detection industry would prefer we not ask: how many decisions, and how much of those decisions, do we actually want to delegate to a machine — and who is responsible when something goes wrong? For authorship, the honest answer is that we have delegated far more than the evidence warrants, to systems whose makers face almost no consequence for being wrong, against people who can least afford the error. The false positive is the tell. It reveals that we have extended trust to these systems well ahead of any reason to, and that the trust, like the harm, is being distributed by power rather than by proof. The judgment of who wrote a sentence is becoming the judgment of who is suspect — and that is a verdict no society should let an unaudited algorithm render in the dark.

[4] AI Ethics

References

1. A Palo Alto high schooler was accused of AI cheating. His family filed ...
2. AI Cheating Lawsuit Myths 2014 Students Win on Due Process
3. AI Detection Lawsuits: Every Student Case, Outcome, and What the Data ...
4. AI Ethics
5. AI in Schools: Surveillance, Detection, and Student Rights - CPAI
6. AI in the Global South: Opportunities and challenges ... - Brookings
7. AI Use in Schools Growing, but District Policies Haven't Caught Up
8. Algorithmic Bias in Education | International Journal of Artificial ...
9. El colonialismo digital en la era de la IA: siete dimensiones ... - Medium
10. L'IA utilis00e9e pour cibler les migrants et les 00e9tudiants 00e9trangers aux ...
11. La AEPD sanciona el tratamiento de datos biom00e9tricos con IA en la ...
12. Largest study of AI hiring algorithms to date finds 'clear racial ...
13. Los cientos de miles de trabajadores en pa00edses pobres que hacen ... - BBC
14. Police use of artificial intelligence grows as rules lag behind
15. PROOF POINTS: Asian American students lose more points in an AI essay ...
16. Recognising and mitigating LLM Pollution in online behavioural research
17. Researchers Warn of Potential for Racial Bias in AI Apps in the Classroom
18. School AI surveillance like Gaggle can lead to false alarms, arrests ...

19. USA/Global: Tech made by Palantir and Babel Street pose surveillance ...

20. Women Who Use AI Seen As Incompetent; Men Who Use AI Seen As Pragmatic