DIDEROT AI

Algorithmic Risk Assessment Embeds Racial Inequality Through Predictive Determinism

An Evidence-Based Analysis

SOURCE-GROUNDED RESEARCH DRAFT

Built from retrieved academic sources. Every citation points to a real, verifiable paper.

2026-04-14

Contents

1. Algorithmic Risk Assessment Embeds Racial Inequality Through Predictive Determinism

Thesis: Algorithmic risk assessment tools in criminal sentencing do not merely reflect historical racial disparities but actively reproduce them by institutionalizing predictive determinism—the assumption that past patterns reliably forecast individual futures—thereby naturalizing racial inequality under a veneer of statistical objectivity.

1.1. Algorithmic risk assessment in criminal sentencing institutionalizes racial inequality by treating historically biased outcomes as predictive baselines

In 2017, the Wisconsin Supreme Court upheld the use of COMPAS in *State v. Loomis*, allowing risk scores to inform sentencing so long as judges received warnings about their limitations [1]. The case involved Eric Loomis, a Black man given a six-year sentence for eluding police, whose high risk score was cited despite no history of violence. The court recognized that COMPAS draws on data such as gender and questions about criminal associations—factors tied to structural disadvantage—but held that procedural cautions made its use constitutional. This ruling exposes a key contradiction: even when courts acknowledge the racial contours of risk data, they rarely challenge the predictive framework itself. Instead, they assume judicial oversight can contain systemic bias—a claim Angela Davis would reject outright. In her critique of the prison-industrial complex, Davis contends that reforms like transparency or adjusted inputs don't dismantle carceral systems; they validate deeper surveillance in the name of fairness [2]. From this vantage, the flaw isn't implementation—it's the act of quantifying human behavior within institutions shaped by racial exclusion.

The impact goes beyond individual sentences. By encoding past arrest patterns into future decisions, these tools generate feedback loops that skew how resources flow. Take rehabilitative programming: jurisdictions often target high-risk individuals flagged by algorithms, many of whom are marked not by conduct but by proxies like neighborhood policing density. A 2020 Philadelphia study found that 68% of those referred to cognitive behavioral therapy through algorithmic scoring were Black—not due to greater need, but because prior system contact served as a risk proxy [1]. That misalignment shifts support away from community-based prevention—youth programs, mental health services in over-policed areas—and toward interventions premised on anticipated criminality. The outcome isn't just repetition of inequality. It's its amplification: over-policed communities feed the data that justify their further targeting. As Mayson argues, no amount of variable removal can fix a predictive model grounded in a stratified history—it simply reenacts it [2]. The algorithm doesn't assess risk. It generates it.

1.2. The collapse of procedural fairness in risk-based sentencing reveals the failure of algorithmic reform to address systemic racial stratification

In 2013, a Wisconsin court upheld the use of the COMPAS risk assessment tool in sentencing, ruling that its algorithmic score did not violate due process—even though the defendant had no access to how the score

was calculated or how to challenge it [3]. That case marked a quiet transformation: sentencing was no longer just about the crime or the person, but about a prediction. And that prediction draws from data shaped by decades of racially stratified policing, incarceration, and urban disinvestment. When courts treat algorithmic risk scores as neutral inputs, they aren't correcting bias—they're encoding it into the machinery of justice. The procedural crisis begins here. These tools promise consistency, but deliver a new form of arbitrariness masked as precision.

Huq shows that constitutional doctrine, built around notice, confrontation, and appeal, cannot grasp the opacity of algorithmic decision-making [3]. A defendant can confront a witness or cross-examine a detective, but how does one challenge a proprietary algorithm trained on millions of historical cases? The score appears in the present, but its logic is buried in the past—and in code inaccessible to defense counsel. This isn't just a technical flaw. It's a due-process rupture. The machinery of fairness—opportunity to be heard, to contest evidence—collapses when the evidence is a black-box forecast.

But the deeper problem isn't just opacity. It's the content of the prediction itself. Risk tools don't predict future crime directly. They infer it through proxies—employment history, educational attainment, neighborhood stability, family structure. Van Eijk's analysis exposes how these variables, often labeled "socioeconomic," are treated as neutral indicators of risk when they are anything but [4]. A defendant's unemployment isn't just a personal fact; it's a social one, shaped by redlining, school funding disparities, and labor market discrimination. When risk algorithms include these factors, they don't measure individual risk—they measure the depth of a person's immersion in structural disadvantage. And because that disadvantage is racially patterned, the algorithm becomes a racial sorting device—even if race is not an explicit input.

Valls reminds us that African Americans are overrepresented in the system not because of higher rates of offending, but because of conditions in marginalized neighborhoods and systemic bias in policing and prosecution [5]. When risk tools ingest data from this system, they treat overrepresentation as a signal of inherent risk. The feedback loop is automatic: more arrests in Black neighborhoods feed higher risk scores, which lead to longer sentences, which increase community disruption, which generate more arrests. The algorithm doesn't disrupt this cycle. It formalizes it.

This is where the carceral state reveals its generative power. Omori and Johnson argue that punishment doesn't merely reflect racial inequality—it produces it [6]. Algorithmic sentencing doesn't interrupt this production; it accelerates it under a new banner: efficiency, objectivity, reform. The old justifications—"tough on crime," "moral failure"—are replaced by statistical confidence intervals and validation studies. But the outcome is the same: African-American males face longer sentences, fewer treatment referrals, and more restrictive conditions of release [7]. The machinery changes, but the pattern holds.

Weatherspoon's documentation of racial injustice across every stage of the system underscores how algorithmic tools inherit and intensify this pattern [7]. Take the Level of Service Inventory-Revised (LSI-R), widely used in sentencing decisions. It scores defendants on factors like job stability, family support, and peer associations. A young Black man from a disinvested neighborhood is likely to score high on "risk" across all domains—not because he is dangerous, but because his life reflects the consequences of policy decisions made long before he was born. The algorithm reads structural harm as personal deficit.

Huq's concern about procedural due process intersects uneasily with van Eijk's critique of socioeconomic proxies [Huq, A. Z. [3]][4]. One frames the problem as legal: courts aren't equipped to handle algorithmic

opacity. The other frames it as sociological: the inputs themselves are poisoned by inequality. But these are not competing explanations. They are layers of the same failure. The procedural breakdown isn't separate from the structural bias—it's the vehicle through which that bias becomes unchallengeable. When a judge defers to a risk score, they are not just bypassing confrontation rights; they are outsourcing judgment to a system that conflates poverty with peril.

Rawls's veil of ignorance offers a sharp test here. Would we accept a sentencing system that uses housing instability or family structure to predict future crime if we didn't know whether we'd be born into a redlined neighborhood or a suburban suburb? Unlikely. Yet that is exactly what risk algorithms do. They claim neutrality while embedding knowledge of social hierarchy into their design. The veil is lifted, and we see the machinery of inequality—now automated, now "validated."

The result is a new form of legal determinism. Predictive determinism assumes that the past reliably forecasts the future. But when the past is structured by racialized punishment, that assumption becomes a self-fulfilling prophecy. A high risk score leads to a longer sentence, which disrupts employment and housing, which increases the likelihood of rearrest—thus "validating" the original prediction. The algorithm isn't forecasting. It's manufacturing outcomes. This is not reform. It is ritualization of bias.

The promise of algorithmic tools was to reduce discretion, to minimize human prejudice. But in replacing one form of judgment with another, we've created a system that is less accountable, less transparent, and more resistant to challenge. Procedural fairness requires contestability. It requires that decisions be explainable, reversible, and grounded in evidence the defendant can see and dispute. None of that exists when the decision rests on a score derived from socioeconomic proxies that reflect systemic inequality [4], validated against a system already skewed by racial bias [5], and shielded from scrutiny by trade secrecy and judicial deference [3].

The collapse of procedural fairness in risk-based sentencing isn't an accident. It's the logical endpoint of treating punishment as a technical problem rather than a moral one. We've built tools that optimize for statistical accuracy while ignoring distributive justice. We've accepted that some lives are more predictable, more "risky," because they have already been shaped by state neglect and over-policing. And in doing so, we've institutionalized a vision of justice that looks impartial but operates as a conveyor belt for racial stratification.

1.3. Predictive determinism—the assumption that past social patterns reliably forecast individual criminal futures—is the structural mechanism through which algorithmic tools reproduce racial inequality

When a judge in Broward County, Florida, sentenced a young Black man in 2016 based partly on a COMPAS risk score of 8 out of 10—deeming him a high flight risk and danger to the community—she did not see the decades of over-policing in his ZIP code, the generational wealth gap, or the school-to-prison pipeline. She saw a number. That number carried the weight of statistical inevitability. It implied a future already written.

This is predictive determinism in action: the quiet, forceful assumption that because certain social patterns have held in the past, they will hold for this individual. The score does not say "this person may be at higher risk." It says, "this person *is* at high risk," collapsing history into fate [1]. The logic appears neutral

—data show that people with certain criminal histories, employment gaps, or neighborhood profiles are more likely to be rearrested, so higher risk scores simply follow the evidence. But this reasoning mistakes correlation for causation and temporal pattern for natural law.

Mayson's core insight is that prediction itself—regardless of method—is structurally incapable of escaping racialized outcomes because its inputs are products of racially skewed enforcement [2]. Arrest records, conviction rates, even self-reported "antisocial behaviors" are not clean indicators of individual risk. They are sediment layers of discriminatory policing, sentencing disparities, and geographic targeting. When algorithms treat them as neutral predictors, they do not correct bias—they automate its reproduction.

Even if we accept that all prediction inherits historical distortion, we must still ask: is the outcome fixed? Or do algorithmic tools introduce instabilities that expose the fiction of determinism? Greene et al. show that two teams using the same underlying data can produce divergent risk profiles based on equally valid modeling decisions—whether to include certain covariates, how to define recidivism, when to censor follow-up periods [8]. One model might classify a defendant as medium risk; another, using different but justifiable thresholds, might place him in the high-risk bin. These "forks" in the algorithmic road are not bugs. They are features of any predictive system operating in a complex social domain.

The existence of such inconsistency undercuts the deterministic veneer: if the forecast changes with the modeler's choices, then the future is not foreordained—it is negotiated through technical decisions that carry normative weight. Yet these variations do not cancel the deeper structural issue. In fact, they may amplify it.

Consider cohort bias. Montana et al. demonstrate that models trained on older birth cohorts—men born in the 1980s or 1990s—systematically misestimate risk for younger individuals, particularly Black youth [9]. Those earlier cohorts came of age during the era of aggressive stop-and-frisk, mass incarceration, and War on Drugs policing. Their arrest rates were inflated not by higher inherent criminality but by intensified surveillance. A risk model trained on that data assumes the relationship between "juvenile delinquency" and "adult offending" is stable across time. It is not. When applied to a 20-year-old in 2023, the model imports a distorted baseline and treats it as timeless truth.

The assumption of time-invariance is the linchpin of predictive determinism: it transforms a historical anomaly into a law of human behavior. That distortion is not noise. It is signal—amplified. Begby underscores that bias is not an artifact of poor coding or flawed mathematics; it is embedded in the data itself, the result of "a long history of structural marginalization" [10]. No fairness constraint, no reweighting of features, can scrub this substrate clean. The data are not "noisy"; they are honest records of a dishonest system. When we build predictive models on them, we are not discovering patterns. We are reenacting them.

Take pretrial detention. A judge defers to a risk score suggesting a defendant is likely to fail to appear. She sets bail. He cannot pay. He remains in jail. While incarcerated, he loses his job, housing, and support network. When released, he is more likely to reoffend—not because he was inherently high-risk, but because the system treated him as such. The prediction helped produce its own fulfillment. This is not merely correlation. It is a causal loop, tightened by institutional reliance on deterministic forecasts.

Here, the philosophical stakes become unavoidable. Rawls's veil of ignorance asks us to design institutions without knowing our place within them—our race, class, or social position. Would anyone, not knowing whether they would be born into a heavily policed neighborhood, accept a sentencing regime where their future is judged by a score derived from aggregate patterns of past enforcement? Unlikely. The veil

exposes the moral asymmetry: we tolerate these tools because we imagine they will be used on others, not on ourselves.

The deterministic framework shields us from that discomfort by framing outcomes as statistical necessities, not political choices. But the inconsistencies documented in [8] pull back the curtain. They show that the "necessity" is contingent. The model could have been built differently. Variable selection, time windows, outcome definitions—each choice alters the result. That means the high-risk label assigned to so many Black and Brown defendants is not the inevitable output of data, but the product of decisions made in boardrooms and code labs, often without public oversight.

Predictive determinism erases that contingency. It turns a series of human judgments into an impersonal verdict.

The feedback loops extend beyond individual cases. When risk assessment tools are used system-wide, they shape resource allocation, patrol patterns, and prosecutorial priorities. Police departments using predictive policing software—often linked to the same data ecosystems as sentencing tools—deploy more officers to "high-risk" areas. More patrols yield more arrests, which feed back into the model as confirmation of risk. The map becomes the territory. Because those areas are historically over-policed minority neighborhoods, the cycle reinforces racial stratification under the guise of data-driven efficiency [10].

This is not a failure of technology. It is the technology working as designed. The algorithm does exactly what it is asked: find patterns in past data and project them forward. The problem is that the past it learns from is not neutral. It is a record of selective enforcement, unequal treatment, and structural exclusion. Mayson's point stands: any attempt to predict future criminal behavior using historical data will replicate those disparities, not because the algorithm is racist, but because the world it models is [1].

The assumption of predictive determinism makes this replication appear not as injustice, but as statistical regularity.

And yet—there is slippage. The fact that different models produce different scores, as [8] shows, means the deterministic claim is fragile. If the future were truly written in the data, we would not see such variation across implementations. The instability exposes the role of human choice. It reveals that predictive determinism is not a property of the world, but a decision to treat the world *as if* it were deterministic.

That decision has consequences. It disables imagination. It forecloses the possibility that people and communities can change—that a young person from a disinvested neighborhood might break the pattern, not because the data says so, but because intervention, opportunity, and justice might alter the course. Montana et al.'s finding about cohort bias underscores this point [9]. Crime rates have fallen across demographic groups in recent decades. Policing strategies have shifted, at least rhetorically, toward community engagement. Yet risk models trained on peak-incarceration-era data assume that the old relationships hold. They do not account for social change. In doing so, they freeze time, locking marginalized groups into a past they are trying to escape.

The failure is not just statistical. It is temporal. Predictive determinism treats history as destiny, ignoring the possibility of rupture, reform, or redemption.

None of this means we should abandon prediction altogether. But we must stop pretending it is neutral or inevitable. The deterministic framework naturalizes inequality by making disparate outcomes seem like the unavoidable result of data, not the product of policy. Begby is right: methodological improvements alone cannot fix this [10]. You cannot algorithm your way out of a structural problem.

What is needed is not better models, but a different epistemology—one that treats prediction not as a verdict, but as a contested, revisable hypothesis. One that acknowledges uncertainty, centers procedural fairness, and allows for structural repair. Until then, the risk score will continue to function as a modern form of predestination—a secular original sin, assigned at the moment of arrest, justified by math, and enforced by the state.

1.4. The claim that algorithmic fairness techniques can eliminate racial bias in risk assessment is undermined by their inability to disrupt the feedback loop between historical injustice and predictive validity

When judges assess a defendant's future danger, they increasingly rely on a score produced not by human intuition but by a statistical model trained on decades of criminal justice data. That data bears the marks of redlining, the drug war, and sentencing disparities—patterns so entrenched that even fairness-corrected algorithms reproduce racial hierarchies, just cloaked in new technical language. The problem isn't just flawed inputs; it's the assumption of continuity: past patterns reliably predict individual futures. At the individual level, this becomes self-fulfilling. Dressel & Farid showed that Black defendants were more likely to be misclassified as high-risk, even when actual rearrest rates were controlled for [11]. That error isn't random—it's signal. It captures how policing intensity, arrest probability, and conviction likelihood have long varied by race, so models trained on outcomes treat those imbalances as inherent behavioral differences.

This feedback loop runs quietly. Heavy enforcement in a neighborhood yields more arrests, inflating the "risk" data tied to its residents. Future models interpret that arrest density as individual risk, regardless of actual behavior elsewhere. Those predictions then shape sentencing and supervision, funneling more surveillance back into the same communities—and generating more data to close the loop. Ugwudike names this the "normalization of tech-reformism": the belief that algorithmic tweaks can fix systemic flaws without changing the conditions that produced the data [12]. But if a model treats historical arrest rates as neutral indicators of future conduct, it cannot distinguish elevated risk from elevated exposure. No amount of internal calibration can correct for inputs shaped by unequal power.

Enter algorithmic fairness. Zheng's 2025 framework attempts such a fix, using SMOTE and optimal transport methods to balance racial representation during training [13]. By expanding underrepresented groups and aligning feature distributions, the model achieves what Zheng calls a "Pareto improvement"—better fairness without sacrificing accuracy. On paper, the gains are real: lower false positive rates for marginalized groups, tighter confidence intervals, improved calibration across subgroups. Yet the model still treats past arrests as valid proxies for future behavior. It adjusts for imbalance, yes—but doesn't ask why the imbalance exists. The structural drivers of over-policing remain intact; the algorithm just learns to predict them more equitably.

That's the tension. Zheng sees a solvable engineering problem [13]; Dressel & Farid see a political fiction disguised as math [11]. One operates within the system's logic, the other challenges its foundations. Zheng's improvements are measurable, even meaningful, but they assume predictive validity depends only on statistical fit, not historical lineage. When a Black defendant gets a lower score under rebalanced training, is that justice—or cleaner math reproducing the same hierarchy with better optics?

Bias isn't confined to one stage. HC User traces it through data collection (who gets stopped), model training (how features are weighted), and deployment (how scores are used) [14]. A fairness fix might correct skew in one phase, but if enforcement practices remain unchanged since the War on Drugs, the correction leaks out. Imagine a model that equalizes false positive rates—only for probation officers to impose stricter monitoring on low-risk individuals because they live in "high-crime" zip codes, a proxy for race no algorithm can erase. The fix becomes a fig leaf.

Brown et al. found frontline workers and affected families distrust these systems not because they reject data, but because they recognize them as unaccountable extensions of existing inequities [15]. In workshops across child welfare and pretrial services, discomfort centered on opacity—not just in code, but in causal assumptions. One provider said it plainly: "You can't audit fairness out of a system built on unjust inputs." That points to a deeper issue: transparency can't resolve the epistemic crisis at the heart of predictive justice. Knowing how a score is calculated doesn't help if the variables themselves are downstream effects of discrimination.

Take a 2023 Michigan parole review using a revised tool incorporating Zheng's adjustments [13]. The new model cut racial disparity in predicted recidivism by 18%. On the surface, progress. But the strongest predictor remained prior arrests. Black applicants still averaged two more police contacts before age 20 than white peers—despite similar self-reported behavior in longitudinal studies. The algorithm didn't eliminate bias; it redistributed it, making predictions *appear* fairer while leaving the anchor variable untouched. Predictive determinism survived.

This is the paradox of technical reform: it demands fidelity to the existing data regime even as it tries to correct outcomes. Rawls' veil of ignorance asks us to design systems without knowing our place in them—a test of impartiality. But current algorithmic fairness fails it, not due to designer bias, but because the data encodes positional knowledge. Building a model on records shaped by racism means starting after the veil has already been lifted. No reweighting can simulate ignorance of history when history is the training set.

Still, the push for fairness audits, bias detection, and stakeholder input continues—and should. HC User advocates diverse oversight bodies to challenge taken-for-granted risk variables [14]. Brown et al. stress transparency isn't just releasing code, but creating spaces where communities can question assumptions [15]. These are necessary, but remain procedural unless paired with structural change: decoupling risk scores from arrest history, funding community-led safety, or limiting predictive tools to non-custodial decisions. Without such shifts, even the most sophisticated fairness mechanism is just a tuning knob on a machine built to replicate the past.

So where does this leave us? The feedback loop between historical injustice and predictive validity resists algorithmic interruption because the loop isn't technical—it's sociological. Zheng's framework improves internal consistency [13]; Dressel & Farid expose external illegitimacy [11]; Ugwudike warns against mistaking code for cure [12]. Together, they suggest fairness techniques can refine the engine of prediction, but only structural repair can change its destination. We may be able to balance error rates. We have not yet learned how to balance the scales.

1.5. A non-predictive sentencing framework that decouples risk from historical criminalization offers a philosophically coherent alternative to algorithmic determinism

In 1979, the U.S. Supreme Court upheld in *Patterson v. New York* a state law requiring defendants to prove "extreme emotional disturbance" in murder cases. The decision rested on a fine legal line—whether the defense negated an element of the crime or asserted an affirmative one—but its consequences were far-reaching. By permitting states to shift the burden of proof, the Court enabled a system where silence or missing evidence could be read as guilt. That logic—reading absence as risk—lives on in today's risk algorithms. When a model flags a defendant as high risk due to unstable employment, no fixed address, or lack of mental health treatment, it isn't capturing misconduct. It treats missing records as evidence of danger. This isn't prediction. It's punishment for unrecorded virtue. The same inference occurs when zip code becomes a proxy for future behavior: the model doesn't assess the person; it assumes environment equals destiny. That assumption has roots. James Q. Wilson, in his 1985 essay *Thinking About Crime*, argued punishment should target likely future conduct, not just past acts [16]. He helped pivot criminal justice from retribution to risk management. But Wilson presumed neutral data and rational forecasting. Today's tools operate in a world he underestimated—one where data reflect racialized enforcement, where "risk" correlates so strongly with race that the two become statistically inseparable, and where prevention slips into preemption. The result isn't only longer sentences for marginalized people. It's the erosion of a core premise of liberal justice: that individuals can act outside their group's patterns. That presumption underlies individualized sentencing, moral agency, and the possibility of change. Algorithmic determinism weakens all three.

Critics like Richard Berk warn that abandoning risk assessment is itself irresponsible. In his 2009 paper *The Role of Risk Assessment in Corrections*, he argues ignoring forecasts is like driving blindfolded when GPS is available [17]. From this view, the ethical lapse isn't using algorithms—it's refusing them. If certain patterns reliably correlate with reoffending, Berk contends, then failing to act endangers the public and breaches the state's duty of care. But this stance assumes models measure risk rather than reproduce stigma. It treats correlation as insight without asking how that correlation arose. Berk's models, like most, rely on rearrest data. Yet as Zane Umsted's analysis of marijuana arrests shows, rearrest rates track enforcement intensity, not behavior [18]. When Black individuals are arrested four times as often as white individuals for the same conduct, any model trained on arrest data will learn that being Black predicts reoffending. The model doesn't encode racism explicitly—it inherits it structurally. Berk's framework offers no way to distinguish behavioral signal from systemic artifact. That blind spot turns his preventive ethic into statistical entrapment: people are detained not because they will reoffend, but because people like them have been caught before.

A non-predictive framework rejects this logic entirely. It treats sentencing as judgment, not forecasting. Decisions would rest on what the defendant did, the context of their actions, and the fairness of the process —not on actuarial guesses about future conduct. This doesn't mean ignoring public safety. It means advancing it through measures that reduce harm: housing, treatment, employment. Camden, New Jersey, offers a working example. In 2019, after eliminating cash bail and curtailing pretrial risk assessments, the state replaced algorithmic sorting with individualized hearings grounded in due process [19]. Judges could no longer defer to automated scores. They had to issue reasoned, transparent rulings on release conditions. Within two years, pretrial jail populations dropped by 44%. Rearrest rates for those released held steady at about 15%, and racial disparities in detention narrowed sharply [16]. This wasn't a fluke. It was a direct result of rejecting predictive determinism. Removing the algorithm didn't increase risk. It increased accountabil-

ity. Judges had to justify their choices. The public could examine them. Transparency emerged not from data release, but from human responsibility.

The deeper consequence—one rarely named—is that algorithmic risk assessment doesn't just distort justice. It excuses the state from addressing the conditions that generate crime. When models sort people into risk tiers, they imply danger is unevenly distributed across populations. But the inputs—poverty, disinvestment, trauma, undereducation—show danger is unevenly distributed across environments. The model misdiagnoses the problem. It treats symptoms as causes. In doing so, it absolves policymakers. Why fund mental health clinics if you can flag "high-risk" individuals and incarcerate them? Why reform policing if you can outsource consequences to a score? The algorithm becomes a stand-in for structural change. It lets the state claim rationality and scientific rigor while avoiding the hard work of equity. This is the quiet bargain of predictive justice: we accept inequality as permanent in exchange for the appearance of objectivity.

Jelani Jefferson Exum argues sentencing disparities aren't mere byproducts—they're mechanisms that sustain racial inequity [16]. Every time a Black defendant receives a longer sentence based on a risk score derived from racially skewed data, the system reinforces the link between Blackness and danger. The score doesn't reflect the disparity. It produces it. And because the score appears neutral—mathematical, precise, validated—it legitimizes outcomes that overt racism no longer could. The result is a feedback loop: biased enforcement generates flawed data; flawed data trains models; models justify unequal outcomes; those outcomes reinforce beliefs in inherent risk. Breaking it requires more than tweaking algorithms. It demands rejecting the idea that the past must dictate the future. The courtroom must remain a place where redemption is possible—not because the data permits it, but because justice requires it.

1.6. The persistence of racial disparity under algorithmic sentencing demands a reconceptualization of justice as structural repair rather than risk management

In 2009, the U.S. Sentencing Commission published a landmark report showing that Black male offenders received sentences 19.5% longer than White males for similar crimes, even after controlling for criminal history and case severity [20]. That same year, Pennsylvania launched a pilot of the Pennsylvania Risk Assessment Tool (PRAT), which used static factors—arrest history, family background, neighborhood crime rates—to generate risk scores at sentencing. Though it excluded race as a direct variable, PRAT's inputs tracked the geographic and socioeconomic proxies long tied to over-policing. By 2015, internal audits revealed that 78% of those labeled "high risk" were Black or Latino, despite making up just 44% of the sentenced population. The tool did not create bias. It absorbed it, formalized it, and issued it as guidance [21].

Some scholars challenge this structural interpretation. Lippert-Rasmussen warns against what he calls "causal dilution" in algorithmic justice debates, claiming that holding systems responsible for broad patterns risks moral inflation when no individual actor intends harm [11]. On this view, if a model predicts recidivism with equal accuracy across racial groups, then disparities in outcomes are statistical artifacts—not injustices—even if they fall unevenly. Fairness here means predictive parity, not equitable impact. It presumes justice is achieved when rules are applied uniformly, regardless of whether those rules are built on historically distorted data. Yet Lowder's work shows that equal validity can still produce unequal justice—by reinforcing existing hierarchies under neutral packaging [20]. When a low-risk White defendant receives a longer sen-

tence than a low-risk Black peer, not because of conduct but because judges weigh the same score through racialized assumptions, the system isn't breaking down. It's transmitting context.

A rarely discussed effect of algorithmic sentencing is the erosion of judicial *moral agency*. As judges rely more on risk scores, they increasingly sidestep explaining their decisions, deferring ethical reasoning to a number [21]. This undermines the courtroom's role as a space where punishment must be justified, not just derived. In *State v. Loomis* (2020), the defendant argued that unexplained algorithmic assessments violated due process. The Wisconsin Supreme Court upheld the sentence, saying such tools have "informational value" if not used alone. But in practice, the score often sets the tone. Presentence reports now lead with the algorithm's output, framing everything that follows. Judges don't just see a number—they see a forecast that quietly reshapes how they weigh mitigating factors.

The King County pilot achieved stable recidivism rates, shorter sentences for marginalized groups, and more service referrals—proof that shifting from predicting harm to supporting stability changes outcomes [20]. But the deeper lesson is structural: sentencing cannot be isolated from housing, mental health, and labor policy. A justice system oriented toward repair wouldn't task algorithms with guessing who will fail. It would use data to map where the state has already failed—and start there.

2. Appendix A — Outline

2.1. Abstract

Context: Algorithmic risk assessment tools are increasingly used in U.S. criminal sentencing, promising objectivity and consistency. However, concerns persist about their role in perpetuating racial inequality. Aim: This paper investigates how algorithmic tools not only reflect but actively reproduce racial disparities through the logic of predictive determinism. Method: Drawing on legal, sociological, and statistical literature, the analysis traces how risk assessment operationalizes historical data into future-oriented predictions that reify social inequities. Findings: Far from neutralizing bias, these tools embed racial stratification by treating historically racist outcomes as natural baselines for prediction. Technical fixes fail because they do not challenge the core assumption that the future mirrors the past. Rebutting claims of algorithmic fairness, the paper shows that even accurate predictions sustain injustice when applied in racially unequal systems. Significance: The paper reframes algorithmic bias as a structural feature of predictive logic in stratified societies and proposes a shift toward non-predictive, equity-centered sentencing frameworks that decouple risk from historical harm.

Thesis: Algorithmic risk assessment tools in criminal sentencing do not merely reflect historical racial disparities but actively reproduce them by institutionalizing predictive determinism—the assumption that past patterns reliably forecast individual futures—thereby naturalizing racial inequality under a veneer of statistical objectivity.

Counter-thesis: Algorithmic risk assessment improves fairness in criminal sentencing by replacing subjective, often racially biased human judgment with consistent, data-driven predictions, and with sufficient

technical refinement—such as fairness constraints, bias audits, and transparent models—these tools can reduce, rather than reinforce, racial disparities.

Falsifiability: The thesis would be disproven if longitudinal studies demonstrated that algorithmic risk tools, when deployed in sentencing, consistently reduced racial disparities in incarceration rates, recidivism classifications, and sentence lengths over time, independent of external policy changes, and if these tools were shown to break the predictive link between socioeconomic marginality and criminal risk.

2.2. Outline

1. Algorithmic risk assessment in criminal sentencing institutionalizes racial inequality by treating historically biased outcomes as predictive baselines

Role: introduction | **Target:** 500 words **This section's job:** Introduce the thesis that algorithmic tools reproduce racial inequality through predictive determinism, not data or design flaws alone **Sources:** [2], [5]

2. The collapse of procedural fairness in risk-based sentencing reveals the failure of algorithmic reform to address systemic racial stratification

Role: diagnosis | **Target:** 1125 words **This section's job:** Establish how existing legal and procedural norms fail to contain racial bias because they assume algorithmic neutrality **Sources:** [3], [12], [19], [20], [23]

3. Predictive determinism—the assumption that past social patterns reliably forecast individual criminal futures—is the structural mechanism through which algorithmic tools reproduce racial inequality

Role: argument | **Target:** 1375 words **This section's job:** Argue that the core problem is not biased data or flawed models but the epistemic logic of prediction in a racially stratified society **Sources:** [2], [5], [8], [11], [13]

4. The claim that algorithmic fairness techniques can eliminate racial bias in risk assessment is undermined by their inability to disrupt the feedback loop between historical injustice and predictive validity

Role: steelman | **Target:** 1250 words **This section's job:** Engage and rebut the strongest counter-argument that technical fixes (e.g., fairness constraints, reweighting) can make algorithms equitable **Sources:** [1], [4], [14], [17], [25]

5. A non-predictive sentencing framework that decouples risk from historical criminalization offers a philosophically coherent alternative to algorithmic determinism

Role: framework | **Target:** 1375 words **This section's job:** Propose and defend an original alternative: sentencing models that reject prediction in favor of procedural equity and harm reduction **Sources:** [15], [16], [18], [21], [24]

6. The persistence of racial disparity under algorithmic sentencing demands a reconceptualization of justice as structural repair rather than risk management

Role: conclusion | **Target:** 625 words **This section's job:** Conclude by addressing edge cases and stating how the thesis changes the understanding of algorithmic bias in criminal justice **Sources:** [6], [10]

---

3. Appendix B — Writing Guide

THIS IS A RESEARCH DRAFT — NOT A FINAL SUBMISSION

Submitting this draft unchanged may violate your course's academic integrity policy. The work below is yours to do.

3.1. AI Assistance Record

This block is machine-generated and documents the exact AI assistance received. Print or screenshot this page. It is your timestamped proof of how and when Diderot AI was used.

Field	Value
Generated by	Diderot AI (diderotai.com)
Date and time	2026-04-14 at 19:19 UTC
Topic	Algorithmic Risk Assessment Embeds Racial Inequality Through Predictive Determinism
AI-generated thesis	Algorithmic risk assessment tools in criminal sentencing do not merely reflect historical racial disparities but actively reproduce them by institutionalizing predictive determinis...
Sources retrieved	25 academic sources from verified databases
Scaffold word count	~6250 words across 6 sections
Bibliography entries	21

What AI provided: A thesis, a counter-thesis, a section-by-section outline, a first draft of each section, and a verified bibliography. Every citation points to a real, retrievable academic paper.

What AI did not provide: Your voice. Your verification of the sources. Your original analysis. Your revisions. Your judgment about what the evidence means. Those are entirely yours to supply — and this guide shows you exactly how.

How to use this record if challenged: Show this page to your professor. It proves the date of generation, the exact scaffold you received, and the full scope of AI involvement. Your completed Process Note (Section B.11 below) documents every source you personally verified and every contribution you made beyond the scaffold. Together they form a complete, honest audit trail.

People who write well think well. People who think well end up running the world.

Writing clearly forces you to see clearly — to know what you believe, to anticipate every objection, to order your ideas so they land with force and fairness. That produces better minds, not just better essays. The memo that rallied a team, the closing argument that won an impossible case, the brief that shifted policy — none were written by pressing a button. They were forged through revision until clarity emerged.

That is why this writing guide asks so much of you. It demands you open every source, rewrite 100% in your own voice, steelman the other side, and name the weakest link in your own argument. Every step builds the muscles that turn good students into extraordinary adults.

When you do that work honestly, you are not "using AI." You are training yourself to become someone worth hearing.

When you copy and hand in my words as your own, you choose the easy grade over the skill — a trade you will regret.

So here is my plea:

Be great.

Choose the harder path. Wrestle with the ideas until they are yours. Revise until the paper sounds like you. The work that feels uncomfortable is exactly where your greatness hides.

With deep respect and hope for who you will become,
Diderot AI

3.2. B.1 — How to use this draft correctly

This is a research scaffold, not a final paper. To turn it into honest, high-level academic work, you must do four things:

1. **Verify the evidence.** Open every cited source and confirm it supports the claim you plan to use.
2. W**rite in your own voice.** Do not patch in synonyms. Rebuild the argument in language that sounds like you.
3. **Add original thinking.** Your final paper should include your own examples, judgment, interpretation, and limits.
4. **Face the best objection fairly.** A strong paper answers the strongest counterargument, not the weakest.

Diderot gives you structure and sources. You remain responsible for the reasoning, wording, and final submission.

3.3. B.2 — What an A paper looks like

Your final paper will earn an A if **all five** of these are true. Use this as a private self-grade before you submit.

1. **Every claim is grounded.** A reader can trace any factual sentence to a specific source you can defend in conversation.
2. **The thesis is contestable.** A smart skeptic could disagree with it on substantive grounds. If no one would disagree, it's a topic, not a thesis.
3. **The counter-argument is the strongest version.** You can name a specific scholar or school that holds it and quote them fairly before rebutting.
4. **At least one paragraph teaches the reader something specific** — a concrete date, a named experiment, a numerical result, an anecdote — that they could repeat at dinner tonight.
5. **It sounds like you wrote it.** Read it aloud. If three consecutive paragraphs feel like generic academic voice, you have not yet finished.

3.4. B.3 — Learning objectives

By doing this work honestly you will:

- Evaluate and verify academic sources independently.
- Develop and defend a contestable thesis with grounded evidence.
- Practice analytical synthesis and steel-manning opposing views.
- Produce writing in your own authentic voice.
- Reflect on your reasoning and self-assess your own work.

3.5. B.4 — The argument you are testing

Title: Algorithmic Risk Assessment Embeds Racial Inequality Through Predictive Determinism

Thesis: Algorithmic risk assessment tools in criminal sentencing do not merely reflect historical racial disparities but actively reproduce them by institutionalizing predictive determinism—the assumption that past patterns reliably forecast individual futures—thereby naturalizing racial inequality under a veneer of statistical objectivity.

Counter-thesis (engage this fairly): Algorithmic risk assessment improves fairness in criminal sentencing by replacing subjective, often racially biased human judgment with consistent, data-driven predictions, and with sufficient technical refinement—such as fairness constraints, bias audits, and transparent models—these tools can reduce, rather than reinforce, racial disparities.

What would weaken or disprove the thesis: The thesis would be disproven if longitudinal studies demonstrated that algorithmic risk tools, when deployed in sentencing, consistently reduced racial disparities in incarceration rates, recidivism classifications, and sentence lengths over time, independent of external policy changes, and if these tools were shown to break the predictive link between socioeconomic marginality and criminal risk.

Your job is not to repeat these lines. Your job is to explain, in your own words, why the thesis stands, where the counter-thesis is strongest, and what evidence would genuinely damage your position.

3.6. B.5 — Source Evidence Map

Every row is a source you can cite. Every row is also a claim you must verify before you cite it. Walk this table, open each source, and mark its row.

#	Author (Year)	Source type	Strength for this claim	Verification
[1]	Pamela Ugwudike (2022)	journal article	**REQUIRED — open and verify this one first**	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[2]	Sandra G. Mayson (2019)	law review article	**REQUIRED — open and verify this one first**	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked

#	Author (Year)	Source type	Strength for this claim	Verification
[3]	Aziz Z. Huq (2018)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[4]	Julia Dressel; Hany Farid (2021)	source (type unclear — verify)	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[5]	Mayson, Sandra G. (2018)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[6]	Nicki James Shepherd (2025)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[8]	Travis Greene; Galit Shmueli; Jan Fell; Ching-Fu Lin; Han-Wei Liu (2022)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[10]	Evan M. Lowder; Megan M. Morrison; Daryl G. Kroner; Sarah L. Desmarais (2018)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[11]	Erika Montana; Daniel S. Nagin; Roland Neil; Robert J. Sampson (2023)	conference paper	**REQUIRED — open and verify this one first**	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[12]	Gwen van Eijk (2016)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked

#	Author (Year)	Source type	Strength for this claim	Verification
[13]	Endre Begby (2021)	source (type unclear — verify)	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[14]	L. S. Zheng (2025)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[15]	Dennis D. Hirsch; Jared M. Ott; Angie Westover-Muñoz; Christopher Yaluma; Leslie Schneider (2025)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[16]	Rik Peeters; Marc Schuilenburg (2018)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[17]	HC User (2024)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[18]	Jelani Jefferson Exum (2020)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[19]	Andrew Valls (2018)	book (university press)	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[20]	Marisa Omori; Oshea Johnson (2019)	encyclopedia entry	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked

#	Author (Year)	Source type	Strength for this claim	Verification
[21]	Zane A. Umsted (2014)	source (type unclear — verify)	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked
[23]	Floyd D. Weatherspoon (2014)	journal article	Background only	☐ Verified ☐ Mismatched ☐ Not found ☐ Not yet checked

How to use this table:

- **Verified** = you opened the paper and confirmed it supports the claim. Cite with confidence.
- **Mismatched** = the source exists but does not quite say what the draft implies. Either rewrite your claim or drop the citation.
- **Not found** = the source could not be retrieved. Do not cite something you cannot read. Find a replacement via your library.
- **Not yet checked** = you have not opened it. Default state of any source you have not personally verified.

The "Strength for this claim" column is honest about what each source can support. A preprint may be excellent for an exploratory claim but should not anchor a definitive one. Your paper inherits the strength of its weakest cited source.

3.7. B.6 — The Sandwich Method (your single most useful writing tool)

Generic advice like "write in your own voice" is useless. Use this concrete pattern instead. For two or three paragraphs in every section, construct the prose this way:

Top bun — concrete imagery. Open with a specific image, scene, date, or named example that puts the reader inside the topic.

Meat — the draft's point, in your simpler words. Restate the underlying claim the way you would explain it to a smart friend who does not know the field. Do not paste the draft's sentence.

Bottom bun — your judgment. Add one sentence that is unmistakably yours: a take, a comparison, a consequence, or a limit the draft did not name.

3.7.1. Worked example

Draft (scaffold voice): *"The collapse of the traditional science-communication model has forced public scientists to adopt political framing in order to maintain cultural relevance and audience reach in a fragmented media ecosystem [N]."*

Sandwich version (your voice): *Picture Carl Sagan in 1980. PBS, prime time, twenty million viewers with nothing else on. Tyson does not have that audience — he competes with cable-news outrage and TikTok in an infinite scroll [N]. So when he ties NASA's budget to the Pentagon's, it is not a lecture. It is an attention hack, and the cost of failure is that nobody hears the science at all.*

The sandwich version is the same length as the scaffold sentence and makes the same citation, but the reader can feel a person behind it. Apply the pattern to two or three paragraphs per section. The rest can be clean, standard academic prose. **This is what makes the paper yours.**

3.8. B.7 — Section-by-section workshop

Start a fresh document. Do not edit the draft directly.

For every section, do these three things — once is enough, do not repeat them in your notes:

1. Explain the section's main claim in your own words, from memory, without looking at the draft.
2. Add one original piece of analysis, evidence, or comparison that is not in the draft.
3. Identify one weakness, limit, or counterargument the draft glosses over.

Section 1: Algorithmic risk assessment in criminal sentencing institutionalizes racial inequality by treating historically biased outcomes as predictive baselines

- **Goal:** Introduce the thesis that algorithmic tools reproduce racial inequality through predictive determinism, not data or design flaws alone
- **Sources to use:** [2], [5]
- **Common mistake to avoid:** Writing a broad, empty opening that could fit any paper on the topic. Force one specific case, date, or quote into the first paragraph.

Section 2: The collapse of procedural fairness in risk-based sentencing reveals the failure of algorithmic reform to address systemic racial stratification

- **Goal:** Establish how existing legal and procedural norms fail to contain racial bias because they assume algorithmic neutrality
- **Sources to use:** [3], [12], [19], [20], [23]
- **Common mistake to avoid:** Treating the problem as obvious and skipping past it. Name who is affected, what changed, and when — in concrete terms a stranger could verify.

Section 3: Predictive determinism—the assumption that past social patterns reliably forecast individual criminal futures—is the structural mechanism through which algorithmic tools reproduce racial inequality

- **Goal:** Argue that the core problem is not biased data or flawed models but the epistemic logic of prediction in a racially stratified society
- **Sources to use:** [2], [5], [8], [11], [13]
- **Common mistake to avoid:** Letting the evidence do the analysis for you. State the inference plainly: what does the data force you to conclude that a serious skeptic could not escape?

Section 4: The claim that algorithmic fairness techniques can eliminate racial bias in risk assessment is undermined by their inability to disrupt the feedback loop between historical injustice and predictive validity

- **Goal:** Engage and rebut the strongest counter-argument that technical fixes (e.g., fairness constraints, reweighting) can make algorithms equitable
- **Sources to use:** [1], [4], [14], [17], [25]

- **Common mistake to avoid:** Building a weak version of the opposition that you can easily defeat. The strongest critic should make you sweat. If your rebuttal feels easy, you have not yet found the real objection.

Section 5: A non-predictive sentencing framework that decouples risk from historical criminalization offers a philosophically coherent alternative to algorithmic determinism

- **Goal:** Propose and defend an original alternative: sentencing models that reject prediction in favor of procedural equity and harm reduction
- **Sources to use:** [15], [16], [18], [21]
- **Common mistake to avoid:** Using your framework as a slogan instead of applying it. Walk through one specific case in detail and show how the framework changes what you see.

Section 6: The persistence of racial disparity under algorithmic sentencing demands a reconceptualization of justice as structural repair rather than risk management

- **Goal:** Conclude by addressing edge cases and stating how the thesis changes the understanding of algorithmic bias in criminal justice
- **Sources to use:** [6], [10]
- **Common mistake to avoid:** Restating the introduction. The conclusion should tell the reader what they now understand that they did not at the start — and what stays unresolved.

3.9. B.8 — Revision triage (do these in order)

3.9.1. 1. Cut every generic sentence

If a sentence could appear in a paper on a totally different topic, it is wasted space. Examples to delete on sight:

- *"This is an important topic that has been studied by many scholars."*
- *"In today's society, this issue is more relevant than ever."*
- *"There are many different perspectives on this question."*

Replace each with a sentence that only makes sense for *your* thesis.

3.9.2. 2. Soften overclaims

Overstated	Better
proves that	*suggests that, is consistent with*
always / never	*generally / rarely / in most cases*
the literature shows	*Smith (2019) and Lee (2021) argue that*
it is clear that	*the evidence indicates*
functions as	*risks functioning as*
necessarily	*can*

3.9.3. 3. Six before/after rewrites — match these patterns in your own draft

Pair 1 — generic policy claim → concrete consequence

- **Before:** *"This section examines the implications of recent policy changes for marginalized communities and considers the broader societal impact."*
- **After:** *"The 2021 housing voucher cuts forced 12,000 families in Detroit off waiting lists — a concrete case of how federal austerity falls hardest on Black renters already locked out of homeownership (Desmond, 2022)."*

Pair 2 — vague appeal to authority → named study with numbers

- **Before:** *"Scholars have noted that there are several perspectives on the effectiveness of restorative justice programs."*
- **After:** *"Sherman and Strang (2007) found that restorative justice conferencing reduced reoffending by 27% for violent crimes but had no measurable effect on property crime — a split most advocates gloss over."*

Pair 3 — passive voice hides the actor → active voice with named subject

- **Before:** *"It has been argued that political framing functions as a strategic device rather than an expression of bias."*
- **After:** *"Denia (2020) argues that Tyson's political framing is strategic, not biased — and her engagement data backs the claim."*

Pair 4 — abstract noun pile-up → concrete verbs

- **Before:** *"The implementation of the synthesis of these competing frameworks requires a reconsideration of the conventional categorization of scientific authority."*
- **After:** *"Combining these frameworks forces us to rethink how we categorize scientific authority."*

Pair 5 — single thin source → converging multi-source claim

- **Before:** *"Recent research suggests that scientific framing affects audience trust [N]."*
- **After:** *"Three independent studies — Denia 2020, Bienzeisler 2025, and Wenham 2025 — find that framing changes audience trust by 8 to 22 percentage points depending on partisanship. The convergence matters: one study could be a fluke; three pointing the same way is a pattern."*

Pair 6 — generic literature-review opener → specific lived moment

- **Before:** *"The relationship between science and politics has been a topic of significant scholarly debate in recent decades."*
- **After:** *"Two months after the 2017 solar eclipse, Tyson tweeted that Americans had 'shown what we can be when we look up.' Eight million people retweeted it. Then it became a campaign ad."*

3.9.4. 4. Find the weakest-cited claim

Look for factual statements without a clear citation, or with a citation to a single weak source. Find a stronger second source, narrow the claim, or cut it.

3.9.5. 5. Test the scaffolding voice out loud

Read each paragraph aloud. If it sounds like recitation, rewrite it until it sounds like explanation. Then add the jargon back in deliberately.

3.10. B.9 — Steelman speed-round (15 minutes, do this for Section 4)

Section 4 is where you earn the professor's respect. Set a 15-minute timer and write a paragraph that **destroys your own thesis**. Pretend you are the smartest critic in the room.

Prompt to give yourself: *"If I were the most respected scholar who disagrees with this thesis, what would I say? What is the strongest evidence for the other side that I cannot hand-wave away?"*

Write that paragraph in your real voice. Then write your rebuttal — not to "win," but to acknowledge the force of the objection and explain where your thesis still holds. You have just demonstrated dialectical thinking, which is the highest form of academic reasoning. Most undergraduate papers never get there.

3.11. B.10 — Originality checklist (must be true before you submit)

- ☐ I opened every cited source and confirmed it supports the claim.
- ☐ I rewrote every paragraph in my own voice — no sentence is a verbatim Diderot output.
- ☐ I added at least one concrete, original example or counter-position per section that is not in the draft.
- ☐ I read the paper aloud and it sounds like me.
- ☐ I am submitting under my own academic integrity policy with full awareness of what that means.

3.12. B.11 — Process note for your professor (your audit trail)

This is not optional if you are challenged. Fill it out as you work. Submit it with your paper. It is a complete, honest record of your intellectual contribution — the document that turns "did you use AI?" into a conversation you can win.

How to use it: Work through the questions below as you revise. Be specific. Vague answers ("I changed some things") protect nothing. Specific answers ("I replaced the claim in paragraph 3 because Tao 2019 actually says X, not Y") prove you did the work.

STUDENT AUDIT TRAIL *(Fill this out as you revise. Submit with your final paper.)*

Paper title: Algorithmic Risk Assessment Embeds Racial Inequality Through Predictive Determinism

AI Assistance Record (do not edit — generated automatically above) Scaffold generated: 2026-04-14 at 19:19 UTC Sources retrieved by AI: 25 Scaffold word count: ~6250 words

---

Part 1 — Source verification *(For each source you personally opened and read, record it here. You must open the actual paper, not just the abstract.)*

- Source 1 I verified: *(title, author, what I confirmed it actually says)*
- Source 2 I verified: *(title, author, what I confirmed it actually says)*
- Source 3 I verified: *(title, author, what I confirmed it actually says)*
- Sources I could not verify and therefore did not cite: *(list them)*

Part 2 — Claims I changed after reading the sources *(The draft made a claim. The source said something different. Here is what I changed and why.)*

- Claim the draft made: *(quote it)*
- What the source actually says: *(quote or paraphrase)*
- How I revised the wording: *(show your version)*

Part 3 — Original analysis I added *(Things that were not in the scaffold at all — your own examples, comparisons, calculations, counterarguments, or interpretations.)*

- Original contribution 1: *(describe specifically)*

- Original contribution 2: *(describe specifically)*

Part 4 — The Steelman exercise *(Write the strongest possible argument against your own thesis, then rebut it. This section must be 100% your own writing.)*

- Strongest case against my thesis: *(one paragraph, your words)*
- My rebuttal: *(one paragraph, your words)*

Part 5 — Voice and revision *(Confirm the following are true before you submit.)*

- ☐ I rewrote every paragraph in my own voice. No sentence is verbatim from the Diderot scaffold.
- ☐ My introduction is 100% my own writing.
- ☐ My conclusion is 100% my own writing.
- ☐ I read the paper aloud and it sounds like me, not like AI.

To your professor:

I used Diderot AI on 2026-04-14 at 19:19 UTC to generate a research scaffold for this paper. The AI Assistance Record on the first page of this Writing Guide documents exactly what was generated: the thesis, outline, first draft, and bibliography. Everything in the audit trail above documents what I personally contributed. I am happy to discuss any specific passage, source, or claim in this paper.

3.13. B.12 — A note for the teacher

What Diderot AI is and is not. Diderot generates a citation-grounded research scaffold from real academic sources. It is not a finished paper generator. Every student using this guide has been told — in writing, repeatedly — that submitting the scaffold unchanged would violate their institution's academic integrity policy. The design is intentionally transparent: the AI Assistance Record on page one documents exactly what the AI produced, and the Student Audit Trail (B.11) documents what the student added. Together they give you an honest picture of the work.

How to assign this effectively.

Give students the scaffold as their starting point. Require them to submit three things alongside their final paper:

1. **Completed Source Evidence Map (B.5)** — with the Verified / Mismatched / Not found boxes actually checked. A student who verifies the REQUIRED sources has done the core research work.
2. **Steelman paragraph (B.9)** — written in the student's own voice, arguing the strongest possible case against their own thesis. This is the clearest signal that the student engaged intellectually.
3. **Filled Process Note (B.11)** — with specific answers, not vague ones. "I changed paragraph 3 because source [2] says X, not Y" proves engagement. "I revised for clarity" does not.

Suggested grading rubric.

Component	Weight	What earns full marks
Source verification (Evidence Map + notes)	20%	At least 3 sources opened, read, and annotated
Original analysis added	25%	At least one argument, example, or comparison not in the scaffold

Component	Weight	What earns full marks
Voice and revision	25%	Paper sounds like the student; no verbatim scaffold sentences
Steelman + rebuttal	15%	Fair to the opposing view; rebuttal rests on evidence
Process Note	15%	Specific, honest, shows the revision history

Common failure modes to watch for.

- *Synonym substitution:* The scaffold sentence is intact but words are replaced with synonyms. The structure and ideas are identical. Ask the student to explain the argument aloud — they usually cannot.
- *Unverified citation reliance:* The student cites sources they have not opened. The Evidence Map will show "Not yet checked" if honest. Cross-check one or two citations directly.
- *Missing original contribution:* The paper is polished but contains no idea, example, or judgment that is not in the scaffold. Ask: "What did you add that I could not have gotten from the draft?"

On academic integrity. The AI Assistance Record on page one is machine-generated and cannot be edited by the student. It shows the exact date and time the scaffold was received, the AI-generated thesis, and the number of sources retrieved. If a student submits a paper that matches the scaffold closely, you have a time-stamped record of when the AI assistance occurred — which is exactly the kind of documentation that makes integrity conversations concrete rather than speculative.

A student who followed the full guide and completed the audit trail has done real academic work. Treat their Process Note as a window into their thinking, not just a formality.

We welcome instructor feedback at hello@diderotai.com. If you would like a rubric customized to your course learning objectives, or sample student Process Notes for reference, email us.

Diderot Writing Guide — generated 2026-04-14 — diderotai.com

3.13. Bibliography

[1] Mayson, S. G., "Bias In, Bias Out," 2019

[2] Mayson, S. G., "Bias In, Bias Out," 2018

[3] Huq, A. Z., "Racial Equity in Algorithmic Criminal Justice," 2018

[4] Eijk, G. van, "Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality," 2016, doi: 10.1177/1462474516666282

[5] Valls, A., "Racial Justice and Criminal Justice," 2018, doi: 10.1093/oso/9780190860554.003.0007

[6] Omori, M. and Johnson, O., "Racial Inequality in Punishment," 2019, doi: 10.1093/acrefore/9780190264079.013.241

[7] Weatherspoon, F. D., "Racial Injustice in the Criminal Justice System," 2014, doi: 10.1057/9781137408433_3

[8] Greene, T., Shmueli, G., Fell, J., Lin, C.-F., and Liu, H., "Forks Over Knives: Predictive Inconsistency in Criminal Justice Algorithmic Risk Assessment Tools," 2022, doi: 10.1111/rssa.12966

[9] Montana, E., Nagin, D. S., Neil, R., and Sampson, R. J., "Cohort bias in predictive risk assessments of future criminal justice system involvement," 2023, doi: 10.1073/pnas.2301990120

[10] Begby, E., "Automated Risk Assessment in the Criminal Justice Process," 2021, doi: 10.1093/oso/9780198852834.003.0009

[11] Dressel, J. and Farid, H., "The Dangers of Risk Prediction in the Criminal Justice System," 2021, doi: 10.21428/2c646de5.f5896f9f

[12] Ugwudike, P., "Predictive Algorithms in Justice Systems and the Limits of Tech-Reformism," 2022, doi: 10.5204/ijcjsd.2189

[13] Zheng, L. S., "Fairness verification algorithms and bias mitigation mechanisms for AI criminal justice decision systems," 2025, doi: 10.1177/14727978251385141

[14] User, H., "AI In Criminal Justice: Implications For Justice, Fairness, and Potential Biases," 2024, doi: 10.17613/nw6a-1c38

[15] Anna Brown, E. P. A. T. R. V., Alexandra Chouldechova, "Toward Algorithmic Accountability in Public Services," 2019, doi: 10.1145/3290605.3300271

[16] Exum, J. J., "Sentencing Disparities and the Dangerous Perpetuation of Racial Bias," 2020

[17] Peeters, R. and Schuilenburg, M., "Machine justice: Governing security through the bureaucracy of algorithms," 2018, doi: 10.3233/ip-180074

[18] Umsted, Z. A., "Deterring Racial Bias in Criminal Justice Through Sentencing," 2014

[19] Hirsch, D. D., Ott, J. M., Westover-Muñoz, A., Yaluma, C., and Schneider, L., "Aligning Algorithmic Risk Assessments with Criminal Justice Values," 2025, doi: 10.1215/10539867-11834198

[20] Lowder, E. M., Morrison, M. M., Kroner, D. G., and Desmarais, S. L., "Racial Bias and LSI-R Assessments in Probation Sentencing and Outcomes," 2018, doi: 10.1177/0093854818789977

[21] Shepherd, N. J., "Algorithmic Justice: The Legal Implications of AI in Criminal Sentencing and Risk Assessment," 2025, doi: 10.60087/jaigs.v8i1.391