The Quiet Battlefield:

Algorithmic Governance and the Violence of the Mundane

Flyxion

Independent Researcher

May 2, 2026

Abstract

Public anxiety about artificial intelligence and lethal violence concentrates almost exclusively on the figure of the autonomous weapon: the drone that selects its own targets, the robotic system that pulls its own trigger. This anxiety is not irrational, but it is radically misdirected. The delegating of life-and-death decisions to algorithmic systems is not a future danger to be forestalled; it is a present condition to be reckoned with. Through the management of food subsidies, housing eligibility, medical coverage, labor allocation, bail determinations, sentencing recommendations, child welfare investigations, and predictive policing, governments and institutions already delegate decisions of survival and confinement to automated systems operating largely beyond democratic accountability. This essay makes four related arguments. First, the concept of autonomy in the AI ethics debate should be decoupled from its exclusive association with kinetic weapons systems: a system is autonomous in the morally relevant sense whenever the effective decision boundary lies upstream of any human who could meaningfully alter it. Second, even when attention is directed at the correct domain, the vocabulary used to describe its harms is analytically insufficient: the term "bias" collapses coverage error, optimization asymmetry, decision miscalibration, and institutional delegation into a single charge, misdirecting both diagnosis and remedy. Third, the conditions under which critique circulates—constrained formats, presumed audiences, the premium

on communicability—systematically favor the vocabulary of bias over more precise accounts, so that the medium itself reproduces the misdirection it nominally opposes. Fourth, the demand for neutral or "unbiased" systems is conceptually incoherent: any system capable of inference must exhibit tendencies, and the question is not whether those tendencies exist but how they are governed and by whom. Like the Roman public in 217 BC who demanded that Fabius Maximus engage Hannibal in open battle while the real attrition was conducted in the hills and supply lines of Italy, we are fixated on the dramatic confrontation while the consequential war is fought elsewhere—against populations whose vulnerability is rendered invisible by the language of optimization, and whose claims for accountability are deflected by the language of technical necessity.

1 The Specter and the System

In 217 BC, after Hannibal had inflicted two catastrophic defeats on Roman armies at Trebia and Lake Trasimene, the Senate appointed Quintus Fabius Maximus as dictator with a mandate to save the republic. Fabius refused to do what Rome demanded: he would not give the Carthaginian general a pitched battle. Instead, he kept his legions in the hills, harassed foraging parties, denied Hannibal supplies, and waited. The Roman public was furious. His political opponents called the strategy cowardice and his nickname—*Cunctator*, the Delayer—was meant as an insult. The people wanted to see the enemy confronted, defeated in the open field, made to feel Roman steel. Instead, Fabius let Italian farms burn and villages go undefended, and explained, with a patience the Senate found maddening, that the real war was being won in the logistics and the attrition, not in the spectacular clash of armies.

The contemporary AI ethics debate is structured by a remarkably similar confusion. It has a preferred villain—the autonomous lethal weapon, the drone that selects its own targets, the robotic system that pulls its own trigger—and it demands a Cannae: a decisive engagement with that specific threat. International campaigns have formed to ban autonomous weapons. Academic journals have published comparative definitional analyses to clarify exactly what properties would make a weapons system "autonomous" in the morally relevant sense

(Taddeo and Blanchard, 2022). The implicit premise is that the crossing of some decisive threshold—the moment a machine kills without a human in the loop—would constitute a genuinely new and historically significant moral event, one that must be prevented before it occurs.

Meanwhile, as Hannibal's army did in southern Italy, the actual harm advances through other means. Algorithms already decide who receives food, who receives housing, who receives medical care, who remains imprisoned, and who loses their children. These decisions are not framed as matters of life and death; they are described as administrative efficiency, resource optimization, evidence-based policy. The substantive difference between that framing and the framing applied to autonomous weapons is difficult to defend. What it reflects, rather than defends, is the same impulse that sent the Senate's chosen replacement for Fabius—Gaius Terentius Varro—marching his legions into the catastrophe at Cannae: a preference for the legible confrontation over the diffuse, unglamorous, strategically decisive struggle that is actually underway.

The interrogation of this misdirection requires naming its mechanism. The distinction is not merely between two types of algorithmic decision-making; it is between two modes of violence. Spectacular violence is immediately legible as violence because it is embodied, kinetically concentrated, and temporally compressed into a recognizable event. Administrative violence operates differently: it is temporally extended across months and years of denied benefits, prolonged incarceration, and family separation; it is institutionally mediated through layers of policy, procurement, and bureaucracy that distribute moral visibility until it dissipates entirely (Galtung, 1969; Bellanova et al., 2021). That mismatch in legibility is not a natural feature of the harms involved; it is a political achievement, the product of framings that have become so familiar they appear inevitable. The farm burning in Campania was real harm; it was just not the kind of harm that Roman political culture knew how to see as the decisive theater of the war.

2 Decoupling Autonomy from Kinetics

Fabius understood something his critics did not: that the decisive question was not where the fighting happened but where Hannibal's capacity to sustain his campaign was being eroded. The spectacular event—the pitched battle—was the wrong unit of analysis. What mattered was the structural condition: supply

lines, foraging ranges, ally loyalty, the slow exhaustion of men who were far from home. Once that structural condition was identified, the appropriate response followed from it, and it had nothing to do with what the public, or the Senate, demanded to see.

A structurally equivalent reframing is required in the AI ethics debate. A system is autonomous in the morally relevant sense when the effective decision boundary—the point at which outcomes are determined—is located upstream of any human actor who could meaningfully alter them. This formulation is more precise than the language of "meaningful human involvement" that dominates the weapons debate, because it specifies what is doing the moral work: not whether a human is nominally present somewhere in the process, but whether that human is positioned to make a difference. A judge who receives a risk-assessment score moments before a bail hearing, without the time or tools to interrogate the model's reasoning, is in the loop in a purely formal sense; the effective decision boundary has already been crossed upstream of the courtroom. Empirical research confirms how rarely nominal oversight constitutes genuine control: even when judges are presented with algorithmic recommendations and retain formal authority to override them, the overwhelming majority follow the recommendation, and the small fraction who deviate do not consistently outperform the model (Angelova et al., 2023; Kleinberg et al., 2018).

Santoni de Sio and van den Hoven (2018) offer a philosophical account of meaningful human control over autonomous systems that is not restricted to weapons, arguing that what matters morally is the locus of decision logic. The structural criterion proposed here sharpens that account: meaningful control requires not just that a human be present but that a human be positioned upstream of the effective decision boundary, equipped with understanding and authority sufficient to alter the outcome. Most deployments of automated decision systems in civil governance fail this test in ways that are structural rather than incidental—features of how such systems are designed, procured, and deployed, not bugs that more careful implementation would eliminate (Ada Lovelace Institute et al., 2021).

If autonomy is defined by the location of the effective decision boundary rather than by the presence of a trigger mechanism, then the threshold that the autonomous weapons debate treats as a future danger has already been crossed in civil governance, repeatedly and largely without public acknowledgment. Bloch-Wehba (2022) documents the extent to which automated decision systems

have been adopted at subnational levels of government in the United States, often in the absence of meaningful regulatory frameworks and frequently in domains where outcomes directly determine access to the means of survival. The institutional assemblage—legislators who authorized the procurement, administrators who configured the system, vendors who trained the model, and the datasets that encoded the training signal—that determines eligibility for food subsidies exercises a power over life that is not metaphorically lethal but materially so. Chronic food insecurity produces measurable excess mortality, and the determination to deny or delay benefits is a causal factor in that mortality regardless of whether it is made by a bureaucrat or a scoring model. Hannibal's army starved Italian villages visibly, with soldiers and fire. The algorithmic denial of food assistance produces no comparable image, which is precisely why it does not appear on the moral ledger of autonomous harm.

3 The Enemy Is Not One Thing: A Layered Account of Algorithmic Harm

Hannibal's army was not a unified force. It comprised Carthaginian officers, Spanish mercenaries, and Gaulish allies, each with distinct motivations, distinct vulnerabilities, and distinct modes of failure. The Spanish mercenaries fought for pay and plunder; the Gaulish allies wanted quick raids and disliked long sieges; only the Carthaginian core was bound by genuine strategic commitment. Fabius understood this differentiation even when the Roman public did not. The public saw a single threatening enemy requiring a single decisive response. Fabius saw a coalition whose components would dissolve under different pressures at different rates, and calibrated his strategy accordingly.

The critical discourse about algorithmic harm makes a structurally equivalent error. The dominant explanatory term is "bias," applied as though it named a single thing requiring a single remedy. In fact, what is called bias is a compression of four distinct failure modes, each operating at a different layer of the system, each arising from different causes, each requiring a different intervention directed at a different actor.

The first layer is *coverage error*: when a model is trained on a sample that underrepresents certain regions of the input space, it generalizes poorly there, producing elevated error rates for the populations associated with those regions.

This is a mathematical fact about learning under finite samples, not a moral property of the model or its designers. The appropriate intervention is at the data collection stage, and the appropriate accountability lies with those who specified and funded the training corpus.

The second layer is *optimization asymmetry*: even a model trained on perfectly representative data can produce unequal error burdens if the objective function encodes the wrong tradeoffs. A system optimized for aggregate accuracy will systematically sacrifice performance on minority subgroups when that sacrifice reduces aggregate loss, because minority populations contribute less weight to the aggregate metric (Barocas and Selbst, 2016). The objective function is a statement of values—it encodes judgments about whose outcomes matter and how much—and it is chosen by designers who are in principle accountable for those choices. The impossibility of a single neutral optimization target follows directly from the structure of these tradeoffs: different definitions of fairness are mathematically incompatible, so that choosing any objective is choosing whose interests to prioritize (Friedler et al., 2021).

The third layer is *decision miscalibration*: the conversion of a probabilistic score into a consequential action requires a threshold, and the choice of threshold determines how errors are distributed across populations in ways that are independent of both the training data and the objective function (Corbett-Davies and Goel, 2018). The decisive moment is not the score itself but the threshold at which it is converted into action: detention or release, benefit approved or denied, investigation opened or closed. Threshold design is a deployment decision made by humans who can be identified and held accountable. A well-specified model can produce threshold-triggered harm if the deployment decision is poorly calibrated to the actual cost structure of errors in context.

The fourth layer is *institutional delegation*: the harms critics most frequently document arise not from models alone but from institutional assemblages that treat probabilistic outputs as authoritative determinations, extend predictions into high-consequence domains without oversight, and lack the correction mechanisms that would catch and address errors (Busuioc, 2021; Ada Lovelace Institute et al., 2021). Technical fixes confined to data or model architecture are insufficient when the failure is institutional: this is as true in hiring, where auditing the model does not address the organizational practices that determine how its outputs are used, as in any other domain (Raghavan et al., 2020). This is the layer that corresponds most directly to the Fabian theater of the war: it is where

the attrition actually happens, it is the least visible, and it is the layer that the discourse of algorithmic harm is least equipped to name.

The taxonomy matters because Fabius did not respond to Hannibal's cavalry, his infantry, and his supply vulnerability with the same tactic. Each component of the coalition required a different response calibrated to its specific weakness. Treating all algorithmic harm as instances of a single phenomenon produces the same strategic failure as treating Hannibal's army as a monolith: the response is misdirected, resources are misallocated, and the components that are actually determining the outcome go unaddressed.

4 The Carceral Web

Fabius's critics accused him of allowing Rome's allies to be butchered and their property burned without intervention. The accusation had some justice: people were suffering in the path of Hannibal's army, and the Fabian strategy required accepting that suffering rather than risking a battle to stop it. What the critics failed to see was that the alternative—engagement on Hannibal's terms—would have been catastrophic, and that the suffering they witnessed in the countryside, real as it was, was not where the war was decided.

The carceral application of algorithmic governance presents a precisely analogous structure. The visible harm—an individual imprisoned, a family separated, a community subjected to intensified surveillance—is real and serious. But the mechanism that produces it is not the dramatic confrontation that political attention gravitates toward; it is the slow, structural operation of risk-assessment instruments and prediction models that most people never see. The four-layer account developed above applies at every level simultaneously, which is precisely why the accountability gap in carceral algorithmic governance is so resistant to reform.

At the coverage layer, instruments trained predominantly on records from over-policed communities generalize poorly for populations whose contact with the carceral system reflects enforcement intensity rather than behavior. At the optimization layer, instruments calibrated to minimize aggregate prediction error sacrifice accuracy for populations that contribute less to the aggregate metric. At the decision layer, the conversion of a continuous risk score into a binary detention or release decision involves a threshold that is rarely designed with explicit

attention to the error distribution across demographic groups, so that threshold-triggered harm can arise even from a technically well-specified model (Corbett-Davies and Goel, 2018). At the institutional layer, the accountability problem is double-structured: there is epistemic opacity, where the model's reasoning is unavailable to the judge, defendant, or counsel; and there is institutional diffusion of responsibility, where the judge who follows the recommendation, the vendor who sold the instrument, and the jurisdiction that adopted it each claims only partial ownership of the decision (Busuioc, 2021; Kleinberg et al., 2018).

This gap is structurally identical to what Taddeo and Blanchard (2022) and others identify as the central moral problem of autonomous weapons. In the weapons case, theorists worry about a future in which no one can be held responsible because the machine made the decision. In the carceral case, that future is already the present. Fabius's allies were being harmed by a force that no Roman commander claimed direct responsibility for stopping. Rome's inability to see that harm as the decisive theater had political, not military, causes.

The feedback mechanism underlying the worst pathologies of carceral algorithmic governance can now be stated with precision. Any system that treats prior state contact as a proxy for underlying risk simultaneously commits a coverage error—treating enforcement records as representative samples of the phenomenon being modeled—and an optimization asymmetry—building a loss function that compounds existing inequality rather than correcting for it. The result is self-reinforcing: predictions authorize enforcement, enforcement generates contact, contact becomes training data, data validates predictions. This is the algorithmic equivalent of Hannibal's foraging problem in reverse: rather than a force slowly starved by the depletion of its supply lines, a population is slowly criminalized by the compounding of its contact history. The mechanism operates at the household level in family policing, where risk-prediction algorithms trained on historical surveillance data produce more investigation, more intervention, and more data, in a cycle whose consequences include the permanent termination of parental rights (eScholarship.org, 2025; Brown et al., 2023). It operates at the community level in predictive policing, where models trained on historical arrest data direct future enforcement toward communities that were over-policed in the past (Florida Law Review, 2023). In both cases, the system is not learning about the phenomenon it purports to measure; it is learning about the history of state attention to that phenomenon.

5 The Myth of the Clean Policy

Fabius was accused, by his political rival Marcus Minucius Rufus, of dilatoriness and inactivity—of allowing the enemy to advance through Roman negligence. The accusation reframed a deliberate strategy as a failure of will, making it legible as weakness rather than calculation. The language of inactivity obscured the fact that Fabius was doing something very specific and very demanding: managing the conditions under which the eventual decisive engagement would be fought. Calling it inactivity transferred the appearance of agency to those doing something dramatic, regardless of whether that something was strategically sound.

The political function of algorithmic framing in governance performs the same displacement. When a jurisdiction adopts an algorithm to determine infrastructure investment priorities, housing voucher eligibility, or labor market program enrollment, the language of optimization frames what is in fact an intensely political decision as a technical one. But the act of formalization is itself a political decision of the deepest kind—one that precedes any data and any model. To instantiate a policy as a learned system is to collapse a contested social reality into a tractable representation: a set of variables, an encoding scheme, an objective function (Scott, 1998). The choices involved in that collapse determine outcomes. Which variables to include, how to operationalize contested constructs, what weight to assign each input, what loss to minimize, what threshold to apply—each is a decision made by specific actors who are in principle accountable for it. Attributing the resulting outcomes to "the algorithm" is not a description of how the system works; it is a rhetorical move that launders agency from human actors into a technical artifact, dissolving accountability in the same way that calling the Fabian strategy "inactivity" dissolved the political responsibility of those who preferred Cannae.

What appears as the algorithm's decision is in fact the compounded effect of coverage limitations in the training corpus, tradeoffs encoded in the objective function, threshold selection at the deployment stage, and institutional choices about where and how the output is used. Each of these is a human decision made by an identifiable actor. The optimization asymmetry and decision miscalibration layers are particularly susceptible to this rhetorical displacement. When a model produces disparate error burdens because its objective function was calibrated for aggregate performance, those burdens appear as a technical property of the

model rather than as the consequence of a design decision. When a threshold produces racially disparate detention rates, the disparity appears as a statistical artifact rather than as the result of a threshold decision that people could have made differently. Santoni de Sio and van den Hoven (2018) note that meaningful human control requires not just a human in the loop but a human capable of understanding what the system is doing and why, equipped with the knowledge and authority to intervene. Proprietary systems shielded by trade secrecy, and technically complex systems whose opacity resists lay scrutiny, fail this test not incidentally but by design (Ada Lovelace Institute et al., 2021).

6 The Misdiagnosis of Bias

The contemporary critique of algorithmic systems has correctly identified a range of harms: systems that produce asymmetric outcomes across populations, that reproduce historical inequities in resource allocation and confinement, that distribute error burdens unevenly across groups. These findings are empirically serious and morally significant (Bellanova et al., 2021; Barocas and Selbst, 2016). What is less clear is whether the dominant explanatory term is capable of explaining them.

The word "bias" is doing three incompatible jobs simultaneously. In the architecture of a learned model, bias names the parameter offset—the additive constant in an affine transformation that allows the model to represent functions that do not pass through the origin. This is a structural necessity; without it, the model's representational capacity collapses. In the theory of statistical estimation, bias names the expected deviation of an estimator from the true population quantity it approximates—the quantity at the center of the bias-variance tradeoff in learning theory, where introducing deliberate bias can reduce variance and improve generalization. In the discourse of social critique, bias names a social disparity: the systematic tendency of a deployed system to produce results that differ in quality or consequence across demographic groups. These three meanings are not interchangeable. The architectural necessity of parameter offsets is silently available as a defense against the empirical finding of outcome disparity; the finding of outcome disparity is treated as evidence that the architecture is corrupt. Both moves are mistakes, and both are made possible by the terminological collapse. The public discourse collapses these three meanings into a single term

and transfers properties between them, which is how the conversation about genuine harm gets absorbed into a debate about mathematical necessity that it cannot win.

The collapse is not merely imprecise. It has the same political effect as the language of optimization examined in the previous section. To say that a system exhibits asymmetric outcomes is already more precise: it names a measurable property of outputs without implying anything about architecture. To say it is "biased" attributes its behavior to a property of the system itself, rather than to the assemblage of coverage decisions, objective design, threshold selection, and institutional use that produces its outputs. The term launders agency from the designers who chose the training corpus, the engineers who specified the objective, the administrators who set the threshold, and the institutions that deployed the output without oversight—into a technical artifact that appears to have acted on its own.

This misdiagnosis has consequences for what remedies are proposed. A politics that demands fairer data without attending to objective functions, or that audits models without attending to deployment thresholds, or that calls for algorithmic reform without attending to the institutional assemblages that convert probabilistic outputs into authoritative decisions, is engaging the wrong component of the coalition (Ada Lovelace Institute et al., 2021; Raghavan et al., 2020). It is, to return to the Fabian frame, skirmishing with the foraging parties while the strategic supply lines remain intact.

7 The Mechanics of Structural Prior: What Precise Critique Would Require

The claim that "bias" is too coarse to name the mechanisms of algorithmic harm is not merely a philosophical objection. It has a precise technical content, and spelling that content out reveals exactly what a more adequate vocabulary would need to do. Two bodies of work illuminate this from complementary directions: the theory of sparse and redundant representations (Elad, 2010), and the spectral analysis of zero-shot posterior sampling in diffusion models (Benita et al., 2026). Neither is presented here as a direct account of algorithmic governance; rather, they make visible the structure of well-posed constraint-closure problems—the kind of inference that algorithmic governance systems nominally perform—

and thereby clarify what goes wrong when that structure is obscured by the vocabulary of bias.

The Sparse-Land model begins with a foundational observation about underdetermined linear systems (Elad, 2010). Given an overcomplete dictionary $D \in \mathbb{R}^{n \times m}$ with $m > n$ atoms, any signal y is consistent with infinitely many coefficient vectors α satisfying $D\alpha \approx y$. The inference problem is to select the *sparsest* consistent solution—the one using the fewest atoms—by solving

$$\min_{\alpha} \|\alpha\|_0 \quad \text{subject to} \quad D\alpha = y.$$

The striking result is that this solution is *unique* whenever the true sparsity satisfies $\|\alpha\|_0 < \text{spark}(D)/2$, where the spark is the size of the smallest linearly dependent subset of columns. Below that threshold, the structural prior—sparsity—is strong enough to select a unique solution from the infinite consistent set. Above it, the problem remains genuinely underdetermined and additional information is required. This is not a failure of the model but a precise characterisation of the conditions under which a structural commitment resolves an ill-posed problem. The mutual coherence $\mu(D) = \max_{i \neq j} |\langle d_i, d_j \rangle| / (\|d_i\|_2 \|d_j\|_2)$, the maximum inner product between distinct normalised atoms, provides a computable proxy: uniqueness is guaranteed when $\|\alpha\|_0 < (1 + 1/\mu)/2$.

This matters for the present argument in the following way. A risk-assessment instrument that predicts recidivism, a benefits-eligibility model that determines food access, or a child-welfare algorithm that flags families for investigation is, in each case, performing inference over an underdetermined system: the measurement (the available data about a person or household) does not uniquely determine the outcome (the risk, the eligibility, the probability of harm). The system *commits* to an outcome by applying a prior—encoded in its training distribution, its objective function, and its threshold—that selects one solution from the consistent set. The spark condition in Sparse-Land names the precise requirement for that commitment to be well-defined: the prior must be strong enough relative to the ambient dimensionality to collapse the solution space to a point. When this condition is violated—when the model is applied in regions of feature space where the training data is sparse, or when the objective encodes tradeoffs that leave multiple outcomes equally plausible—the commitment is arbitrary. The model picks a solution, but that solution is not determined by the evidence; it is determined by whichever asymmetries happen to be encoded in

the prior.

This is what the vocabulary of coverage error names at the data layer: the training distribution is sparse in the relevant region, and the model generalises poorly there. But Sparse-Land makes clear that the problem is not merely one of data volume. The mutual coherence of the training distribution—the degree to which the features used to predict outcomes are entangled with features that should be irrelevant—determines whether adding more data of the same kind would help or merely compound the ambiguity. A distribution with high mutual coherence between, say, prior arrest history and neighbourhood of residence cannot be corrected simply by adding more observations drawn from the same distribution. The structure of the prior itself must change.

The spectral analysis of Benita et al. (2026) introduces the second dimension of this picture. They study the inverse problem $y = Hx + n$ where H is a linear degradation operator, n is noise, and the goal is to recover x using a diffusion model as a prior. Under a Gaussian prior $x_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, they derive the optimal posterior denoiser in closed form:

$$x_0^* = \left[(1-\bar{\alpha}_t)\Sigma_0 H^T H + \sigma_y^2 \bar{\alpha}_t \Sigma_0 + \sigma_y^2(1-\bar{\alpha}_t)I\right]^{-1}\left[(1-\bar{\alpha}_t)\Sigma_0 H^T y + \sigma_y^2\sqrt{\bar{\alpha}_t}\,\Sigma_0 x_t + \sigma_y^2(1-\bar{\alpha}_t)\mu_0\right].$$

When both Σ_0 and H are shift-invariant, the Fourier transform diagonalises them simultaneously, and the full S-step inference process reduces to d independent scalar transfer functions—one per frequency band—each mapping noise and degraded measurements to a clean estimate. The guidance weight ζ_s at each step controls the rate at which the reconstruction commits to the measurement at each frequency. The optimal weight is the one that closes the system to the true posterior at the right rate in each band, jointly accounting for the prior's spectral structure, the degradation operator's frequency response, and the diffusion dynamics. Heuristic weights—the current state of practice—close some frequency bands too fast and others too slow, producing a systematic mismatch between the estimated and true posterior.

The relevance to algorithmic governance is not that these systems are literally diffusion models, but that they share the same structural problem: inference over a continuous space of outcomes under a prior, with a measurement that does not fully determine the answer. The four-layer taxonomy maps directly onto the structure of the spectral analysis. Coverage error corresponds to poor estimation of the prior covariance Σ_0 in underrepresented regions—the model does not

know what the true distribution looks like there. Optimization asymmetry corresponds to the choice of the loss function that determines what "near-optimal" means—analogous to the choice of discrepancy measure D in Benita et al. (2026)'s Wasserstein-2 formulation. Decision miscalibration corresponds to the guidance weight ζ_s: the threshold at which a probabilistic score is converted into a binary action is exactly the parameter that determines how fast and at what rate the system commits to one outcome over another. And institutional delegation corresponds to the deployment context: what happens after the inference is complete, when a probabilistic output is treated as an authoritative determination by a system that lacks the apparatus to audit whether the commitment was well-founded.

Fabius's strategic genius was precisely the ability to see which components of Hannibal's position were structurally determined and which were contingent. The Spanish mercenaries' loyalty was contingent—it depended on continued plunder and quick success. The supply-line vulnerability was structural—it followed necessarily from Hannibal's position as an invading force far from home, and no tactical victory could alter it. The Fabian strategy worked because it applied pressure at the structural point, not the contingent one. The discourse of algorithmic bias applies pressure at the contingent point. It treats the manifest outcome—the disparate error rate, the discriminatory prediction, the unjust allocation—as the site of intervention, when the structural determination of that outcome lies upstream: in the prior's coverage, in the objective's encoding of tradeoffs, in the threshold's conversion of probability into action, and in the institution's treatment of probabilistic output as authoritative fact. Correcting the manifest outcome without addressing the structural determination is, in the precise technical sense illuminated by both Elad and Benita et al., closing the wrong frequency band.

8 The Compression of Critique

The misdiagnosis of algorithmic harm is not only a function of conceptual confusion; it is also a function of the media through which that confusion circulates. The contemporary discourse on algorithmic systems is produced under conditions that systematically favor compression: limited attention, constrained formats, and audiences presumed to require clarity over complexity. These con-

ditions shape not just how arguments are received, but how they are formulated in the first place.

Complex, layered explanations of system behavior do not survive intact under these constraints. A four-layer account distinguishing coverage error, optimization asymmetry, decision miscalibration, and institutional delegation is analytically precise but communicatively expensive. It requires time, technical familiarity, and conceptual patience. In contrast, the term "bias" compresses these distinctions into a single, morally legible claim that can circulate across platforms and audiences with minimal friction. The persistence of the term is therefore not only a conceptual failure but an adaptation to the communicative environment in which critique occurs. Arguments become simplified not just for clarity, but for legibility within institutional expectations: policymakers want actionable summaries, journalists require narrative resolution, and technical nuance gets stripped out at each relay.

This compression produces a structural distortion that mirrors the optimization asymmetry it nominally opposes. Just as models optimized for aggregate performance sacrifice accuracy on minority subgroups, discourse optimized for attention and interpretability sacrifices explanatory precision on difficult mechanisms. The result is a systematic preference for narratives that can be easily communicated over those that are difficult to explain. Visible failures—misidentifications, extreme classification errors, discrete injustices—circulate widely; the structural conditions that produce them—objective functions, threshold choices, institutional deployment decisions—remain obscure. The same selectivity that Galtung (1969) identifies in the social perception of physical versus structural violence operates here at the epistemic level: the immediately recognizable event displaces the slow, distributed mechanism.

The effect is not merely rhetorical. A critique that targets data representation because it is legible, while leaving objective design and deployment thresholds unexamined because they are not, reproduces the same misallocation of attention identified in the autonomous weapons debate. The medium itself reinforces the preference for spectacular over administrative harm: a falsely flagged face is the kind of failure that circulates; a systematically miscalibrated threshold affecting tens of thousands of pretrial defendants does not. The spectacular error is Hannibal crossing the Alps; the structural attrition is the foraging problem, and it does not make for dramatic news.

The result is a discourse that is, in a precise sense, miscalibrated in the same

way it claims to critique. It identifies real harm but directs intervention toward the most communicable layer rather than the most consequential one. The language of bias persists because it fits the medium. The mechanisms it obscures persist because they do not. This is why the essay now has three cascading arguments rather than one: the problem is not only where we look, but what conceptual tools we use when we look correctly, and why even those tools get simplified beyond usefulness before they can do their work.

9 Re-centering the Ethics Debate

After Cannae, the Romans learned. They returned to the Fabian strategy and applied it with the discipline that Fabius had always prescribed. Hannibal spent the remaining decade of the war in Italy unable to force a decisive engagement, steadily losing men and supplies, watching his Spanish mercenaries desert and his Gaulish allies drift back toward neutrality. He was never defeated in open battle; he simply ran out of the capacity to sustain his campaign. The lesson the Romans eventually absorbed was that the real war had always been where Fabius said it was: not in the pitched battle, but in the conditions that made pitched battle possible or impossible. And the strategic insight that made the eventual victory possible was the differentiated understanding of the enemy coalition—not a monolith to be met with a single decisive blow, but a set of components with different vulnerabilities requiring different responses at different rates.

The argument of this essay is not that the autonomous weapons debate is unimportant. The specific moral concerns about military targeting—about international humanitarian law, about the requirements of distinction and proportionality, about the accountability of states for the use of force—are genuine and pressing. The argument is that the current distribution of moral attention mirrors the Roman public's demand for Cannae, and that the consequences of that misdirection fall not on those making the demand but on those in the path of what was always the real advance. To be clear about what is being claimed: this is not an argument for equivalence of mode. A missile strike and a denied benefits claim are different events, different in immediacy, in visibility, in the kind of harm they produce. The argument is for equivalence of structure: both involve the delegation of decision logic over conditions of survival and confinement to systems whose effective decision boundaries lie upstream of meaningful human

intervention.

Nor is this an argument that the critical discourse on algorithmic harm is wrong in its findings. It is an argument that the conceptual tools that discourse employs, and the communicative conditions under which those tools are deployed, prevent it from seeing the mechanisms of harm clearly even when it looks directly at them. The problem is not only where we look, but what we think we see when we look correctly, and why even correct accounts get compressed into a form that misidentifies the decisive action.

Here the philosophical constraint becomes decisive. The demand for neutral or "unbiased" systems is not just impractical; it is conceptually incoherent. Any system capable of inference must generalize from past observations to novel cases, and generalization is the exercise of tendencies—dispositions toward certain outputs given certain inputs, formed through training and applied forward (Hume, 1748). A system with no tendencies would output nothing informative; it would be incapable of prediction. The question is therefore not whether systems exhibit tendencies but how those tendencies are shaped, where they are deployed, and who is accountable for their consequences. The problem is not that these systems have tendencies, but that those tendencies are treated as authoritative within institutions that do not account for their origins, do not audit their distribution of error, and do not own their consequences. This is the real content of what gets called bias: not a property of the model that could be removed, but a governance failure that could be addressed—if the vocabulary of the critique were precise enough to locate it, and if the medium through which the critique circulates preserved that precision rather than optimizing it away.

What is required is an accountability framework organized by the four-layer taxonomy rather than by the undifferentiated charge of algorithmic asymmetry: mandatory transparency about training corpus, objective function, and deployment threshold for all systems making consequential decisions about persons; meaningful appeal processes for individuals subject to algorithmic determinations; regular audits that distinguish coverage error from optimization asymmetry from decision miscalibration from institutional delegation, so that remedies are directed at the right level and the right actor; and legal frameworks that assign responsibility to the institutional assemblages that produce harm rather than allowing the diffusion of agency to dissolve accountability entirely (Ada Lovelace Institute et al., 2021; Busuioc, 2021). Above all, it requires a politics willing to do what Fabius did: to name the unglamorous theater as the decisive

one, to resist the demand for the legible confrontation, and to insist that the violence being absorbed quietly in the countryside of algorithmic governance—the denied claim, the extended sentence, the separated family—is the real war, and that it is being lost not because the enemy is invisible, but because the vocabulary we have inherited to describe it, and the conditions under which that vocabulary circulates, prevent us from seeing it clearly.

The battlefield where algorithmic decisions are already determining who lives, who is free, and whose family remains intact is not a speculative scenario. It is the present. The autonomous weapons debate, for all its genuine moral seriousness, risks functioning as a displacement activity: a way of being concerned about algorithmic power at a comfortable distance from the institutional assemblages that are already operating, already consequential, and already largely unaccountable. Hannibal was not defeated at Cannae. He was defeated in the hills of Campania, by attrition, over years, in a war that Rome spent a long time refusing to recognize as the real one.

References

Ada Lovelace Institute, AI Now Institute, and Open Government Partnership. Algorithmic accountability for the public sector. Technical report, Ada Lovelace Institute, 2021. URL `https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/`.

Victoria Angelova, Will S. Dobbie, and Crystal S. Yang. Algorithmic recommendations and human discretion. Working Paper 31747, National Bureau of Economic Research, 2023.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

Rocco Bellanova, Kristina Irion, Katja Lindskov Jacobsen, Francesco Ragazzi, Rune Saugmann, and Lucy Suchman. Toward a critique of algorithmic violence. *International Political Sociology*, 15(1):121–150, 2021. doi: 10.1093/ips/olab003.

Roi Benita, Michael Elad, and Joseph Keshet. Analyzing and guiding zero-shot posterior sampling in diffusion models. *arXiv preprint*, 2026. arXiv:2602.07715.

Hannah Bloch-Wehba. Algorithmic governance from the bottom up. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4054640. Available at SSRN: https://ssrn.com/abstract=4054640.

Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tomlinson, and Rhema Vaithianathan. Algorithmic harms in child welfare: Uncertainties in practice, measurement, and accountability. *ACM Journal on Responsible Computing*, 1(1), 2023. doi: 10.1145/3616473.

Madalina Busuioc. Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81(5):825–836, 2021. doi: 10.1111/puar.13293.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint*, 2018. arXiv:1808.00023.

Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, New York, 2010. doi: 10.1007/978-1-4419-7011-4.

eScholarship.org. The epistemic injustice of algorithmic family policing. eScholarship, 2025.

Florida Law Review. How premature predictive policing can lead to a self-fulfilling prophecy of juvenile delinquency. *Florida Law Review*, 2023.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021. doi: 10.1145/3433949.

Johan Galtung. Violence, peace, and peace research. *Journal of Peace Research*, 6 (3):167–191, 1969. doi: 10.1177/002234336900600301.

David Hume. *An Enquiry Concerning Human Understanding*. A. Millar, London, 1748.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *Quarterly Journal of Economics*, 133(1):237–293, 2018. doi: 10.1093/qje/qjx032.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020. doi: 10.1145/3351095.3372828.

Filippo Santoni de Sio and Jeroen van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, 2018. doi: 10.3389/frobt.2018.00015.

James C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, 1998.

Mariarosaria Taddeo and Alexander Blanchard. A comparative analysis of the definitions of autonomous weapons systems. *Science and Engineering Ethics*, 28 (5), 2022. doi: 10.1007/s11948-022-00392-3.