# THE ARCHITECTURE OF DISCLAIM: AI, IMPLICIT BIAS, AND THEIR (UN)PREDICTABLE CONVERGENCE

Jerry Kang* and Paul Ohm†

* Jerry Kang is Ralph and Shirley Shapiro Distinguished Professor of Law and (by courtesy) Distinguished Prof. of Asian American Studies. <kang@law.ucla.edu> <http://jerrykang.net>.

† Paul Ohm is Professor of Law at Georgetown Law. Helpful comments were received at colloquia presented at UCLA Law, and from Ahilan Arulanthanan, Song Richardson, Sherod Thaxton, and Noah Zatz. Able research assistance provided by: Amanda Griffin, William-Henry Ku, Daniel Williams, and the reference librarians at UCLA and Georgetown.

*Artificial intelligence and the human mind share a decisionmaking architecture, and antidiscrimination law is built for neither. In both, a producing layer encodes the statistical regularities of an unequal world and shapes the inputs a decision runs on, while an avowing layer deliberates, explains, and sincerely denies discriminatory intent — without being able to observe the processing beneath it. We name this the architecture of disclaim, and we show it is functionally isomorphic across silicon and carbon. Recent computer-science studies of large language models are echoing, sometimes measure for measure, a half-century of implicit bias research on human brains. The parallels are not cosmetic. Models and brains exhibit similar bias effects, with similar resistance to correction. Alignment training and diversity training fail in the same way and for the same reason: each updates what a system says without retraining what it computes. Asked whether it discriminated, either system answers confidently — and the answer is sincere and wrong.*

*The architecture defeats the law's existing toolkit, cluster by cluster. Consciousness (purposeful discrimination) fails because the avowing layer answers honestly. Cause (but-for causation) fails because race travels in a bundle of proxies that no counterfactual can isolate, so the test can convict but never acquit. Consequence (disparate impact) comes closest — it never asks about anyone's interior — but it arrives statutorily confined, constitutionally imperiled, and missing an accepted account of why its disparities matter.*

*A fourth cluster survives. Conduct asks not what the actor intended, observed, or can counterfactually prove, but what it should have done. Two legal traditions are building toward it right now, neither aware of the other. AI governance has converged on assessment, audit, and reasonable response; implicit bias law has replaced Batson's futile hunt for purposeful intent with an epistemically upgraded objective observer. Two construction crews are working with one blueprint. The convergence is not coincidence — it is the signature of the same architecture asserting itself across different substrates — and naming it makes an exchange possible, because each side holds what the other is missing. AI governance holds ex ante duties that attach before any contested decision. Implicit bias law holds the epistemic content without which a reasonableness benchmark is an empty vessel.*

*The exchange composes a new standard: the responsible person — an actor charged with what the science of decisional systems has established, bound by ex ante duties before the decision, and judged on that same knowledge after it. Because the standard interrogates no interior and demands no counterfactual, it reaches the firm whose managers consult a frontier model persistently, informally, and mid-discretion — which is what every firm may be becoming, and what every substrate-specific regime misses. What to do? Judges can begin through evidentiary rulings alone. Legislators can finish, enacting ex ante obligations calibrated to what minds and machines can and cannot do. Two decades ago, one of us asked why we tolerate from our organically grown black boxes what we would never tolerate from synthetic ones. We shouldn't. The architecture of disclaim means that neither minds nor machines can simply avow their way to fairness — but the law can now hold both to becoming what they claim to be.*

Introduction ..... 1
I. Structural Identity ..... 9
A. Empirical Evidence ..... 10
B. Two-Layer Architecture, Developed ..... 17
1. Human Brain ..... 17
2. AI Model ..... 18
3. The Nesting Complexity ..... 22
II. The Doctrinal Toolkit and Its Limits ..... 23
A. The First Cluster—Consciousness: Purposeful Discrimination ..... 24
B. The Second Cluster—Cause: "But-for" Causation ..... 27
C. The Third Cluster—Consequence: Disparate Impact and the Less Discriminatory Alternative ..... 31
D. The Fourth Cluster—Conduct, Excavated ..... 36
III. The Convergence ..... 41
A. Two Construction Crews ..... 42
1. The Silicon Crew ..... 42
2. The Carbon Crew ..... 44
B. The Comparison ..... 46
1. Similarities ..... 46
2. Differences ..... 48
C. The Exchange ..... 50
1. What AI Governance Gives Implicit Bias: Ex Ante Duties ..... 50
2. What Implicit Bias Gives AI Governance: Epistemic Content ..... 53
IV. Synthesis ..... 54
A. The Responsible Person ..... 54
B. The Exchange, Executed ..... 57
C. The Persuaded Lawmaker ..... 62
Conclusion ..... 64

# Introduction

Imagine a company called Headhunter.AI that sells résumé-screening software to employers. The product ingests reams of applications, scores them against job criteria, and surfaces the top candidates for human review.[1] It is trained on historical hiring data across multiple industries—millions of past decisions about who was interviewed, who was offered, who succeeded. An employer deploys Headhunter.AI to manage a flood of (often AI-drafted) applications for an entry-level analyst position. The system screens out ninety percent of applicants before a human being ever lays eyes on a single résumé. Among those screened out, a disproportionate number are Black.[2]

Two features of the resulting discrimination claim are structurally important. Together they constitute what we call the *architecture of disclaim* — a structure that makes discrimination simultaneously real and incognizable through current doctrinal tools. *First*, the employer can sincerely say that it had no intent to discriminate.[3] It instructed Headhunter.AI to comply with all applicable employment laws and to disregard race, gender, age, and disability in scoring applicants. The employer meant it. If you asked the employer whether it wanted to discriminate, the answer would be no, and the answer would be true. Call the employer the *avowing layer* —the part of the decisionmaking system that speaks, commits, decides, and also disclaims.

*Second*, the process by which Headhunter.AI actually scored résumés is the proverbial "black box."[4] The model's pre-training—gradient descent over a gigantic historical corpus[5]—encodes whatever statistical regularities the data

---

[1] This might be a *predictive* AI system, such as a random forest classifier trained on historical employment records, or it might be built atop a newer *generative* AI system, such as an app that screens resumes by prompting a fine-tuned large language model to generate the scores. We will focus on generative AI models and systems in this article. *See infra* note 129

[2] A vast literature examines algorithmic decisionmaking, bias, and antidiscrimination law. *See, e.g.,* Emily Black, John Logan Koepke, Pauline T. Kim, Solon Barocas & Mingwei Hsu, *Less Discriminatory Algorithms*, 113 GEO. L.J. 53 (2024); Andrew D. Selbst & Solon Barocas, *Unfair Artificial Intelligence: How FTC Intervention Can Overcome the Limitations of Discrimination Law*, 171 U. PA. L. REV. 1023 (2023); Talia B. Gillis, *The Input Fallacy*, 106 MINN. L. REV. 1175 (2022); Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 811 (2020); Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1260 (2020); Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1673 (2020); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 56 (2019); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2221 (2019); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 860 (2017); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 680-92 (2016). None of these analyze in any depth the role that generative AI models are now playing in employment systems, and none suggest our architecture of disclaim and its implications for antidiscrimination law.

[3] *See infra* Part I.B.2.

[4] *See* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2016).

[5] *See* IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 290–96 (2016) (foundational deep learning text); Tom B. Brown et al., *Language Models are Few-Shot Learners*, *in* NEURIPS '20: PROCEEDINGS OF THE 34TH CONFERENCE ON NEURAL INFO. PROCESSING SYS. 1877 (2020) (describing large-scale pretraining over corpus of text gathered from online sources).

contain, including the regularities of a labor market shaped by centuries of racial discrimination and unequal access to educational and employment opportunity.[6] Call the AI the *producing layer* — the part of the system that generates the outputs the avowing layer subsequently works with. The employer cannot directly introspect why the system scored Lakisha's résumé lower than Emily's, holding qualifications constant.[7] Of course, it can ask, and given recent advances in large-language models, the model will answer fluently and confidently. But the model may rationalize rather than report.[8] Its explanation can reflect general reasoning about what *might* matter, not the producing layer's actual computations.[9]

In short, the avowing layer cannot much observe the producing layer. The producing layer shapes the inputs the avowing layer works with—filtering, weighting, associating—before the employer ever sees the results.[10] The employer then acts, sincerely, on the basis of outputs it did not see being produced. To allegations of discrimination, the avowing layer denies. This is the architecture of disclaim.

And the architecture nests because any given layer can itself be deconstructed into producing and avowing sub-layers.[11] For example, the AI received post-training alignment that teaches the model to *say* it does not and should not discriminate[12] (the avowing layer) while pre-trained weights encode the corpus statistics that drive its skewed scores (the producing layer). Similarly, the employer is a firm that operates through human beings arranged in multiple avowing-producing pairs (manager to employee, CEO to manager, board to CEO).[13] The disclaim thus cascades. At adoption, the deployer cannot (fully) inspect the developer when the AI system is adopted. At use, the AI cannot inspect (fully) its own weights; the employee-user cannot inspect the AI; the manager cannot inspect the employee; and so on up the org

---

[6] Barocas & Selbst, *supra* note 2, 680-92. *See also* Mayson, *supra* note 2, 2251-54 (making equivalent point in the criminal justice context).

[7] "Emily" and "Lakisha" come from the canonical study done by Sendhil Mullainathan and Marianne Bertrand, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991 (2004).

[8] *See* Miles Turpin et al., *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*, *in* NEURIPS '23: PROCEEDINGS OF THE 36TH ADV. IN NEURAL INFO. PROCESSING SYS. 74952 (2023).

[9] *See* Alon Jacovi & Yoav Goldberg, *Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?*, 58 PROC. ANN. MEETING OF THE ASS'N COMP. LINGUISTICS 4187 (2020) (distinguishing plausible and faithful machine explanations).

[10] *See* Barocas & Selbst, *supra* note 2, at 677-90.

[11] *Cf.* Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian & Janet Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, in 2019 PROC. CONF. FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 59 (noting a "framing trap" when focused on a model rather than a model, plus humans, plus institution).

[12] *See* Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, *in* NEURIPS '22: PROCEEDINGS OF THE 35TH ADVANCES IN NEURAL INFO. PROCESSING SYS. (2022) (describing reinforcement learning with human feedback techniques).

[13] John W. Meyer & Brian Rowan, *Institutionalized Organizations: Formal Structure as Myth and Ceremony*, 83 AM. J. SOC. 340 (1977); Lauren B. Edelman, *Legal Ambiguity and Symbolic Structures: Organizational Mediation of Civil Rights Law*, 97 AM. J. SOC. 1531 (1992).

chart. In sum, the employer did not purposefully intend the different treatment on the basis of race and cannot access the mechanism that produced it. Under existing intent-based antidiscrimination doctrine, this essentially ends the inquiry.

One part of this story should sound strangely familiar. It's just the implicit bias story, retold in silicon.[14] For a generation, implicit social cognition research has documented a similar architecture inside the human brain.[15] The avowing layer is what Daniel Kahneman influentially labeled "System 2"—the

[14] The other familiar story is the principal-agent problem configuring relationships between actors (human-to-human within the firm, and firm-to-firm). For principal-agent analyses of the AI space, see Paul Weitzel, *Governing AI Systems Through Corporate Theory*, 92 TENN. L. REV. 169 (2024) (mapping end-user alignment onto agency-cost theory and societal-level alignment onto the corporate-purpose debate); Ian Ayres & Jack M. Balkin, *The Law of AI Is the Law of Risky Agents Without Intentions*, 11/27/2024 U. CHI. L. REV. ONLINE 1, 7 (2024).

[15] Experimental research in social psychology has established that individuals hold automatic associations measure independent of self-report. *See* Anthony G. Greenwald, Debbie E. McGhee & Jordan L.K. Schwartz, *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCH. 1464, 1464–66 (1998) (introducing the IAT). For recent summaries of implicit bias scores across large samples, see Anthony G. Greenwald & Calvin K. Lai, *Implicit Social Cognition*, 71 ANN. REV. PSYCH. 419, 423–28 (2020) ; Anthony G. Greenwald et al., *Implicit-Bias Remedies: Treating Discriminatory Bias as a Public-Health Problem*, 23 Psych. Sci. Pub. Int. 7, 11 (2022); Tessa E.S. Charlesworth, Meriel C. Doyle & Mahzarin R. Banaji, *Patterns of Implicit and Explicit Attitudes V: Reversals of Trends from 2021–2025* 2, 25 (PsyArXiv, Version 2, 2026), https://doi.org/10.31234/osf.io/26f8v.

A substantial literature finds these associations can influence behavior to a small but significant degree. *See* Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann & Mahzarin R. Banaji, *Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity*, 97 J. PERSONALITY & SOC. PSYCH. 17, 19–20 (2009) (reporting $r = .24$ for Black–White bias); Frederick L. Oswald, Gregory Mitchell, Hart Blanton, James Jaccard & Philip E. Tetlock, *Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies*, 105 J. PERSONALITY & SOC. PSYCH. 171, 178 (2013) (estimating a population correlation of $r = .15$ for the race domain); Benedek Kurdi et al., *Relationship Between the Implicit Association Test and Intergroup Behavior: A Meta-Analysis*, 74 AM. PSYCH. 569 (2019) (meta-analyzing 253 samples; implicit–criterion correlations averaging approximately $r = .10$ across combined domains). For the most recent systematic assessment, see Jordan R. Axt et al., *On the Relationship Between Indirect Measures of Black vs. White Racial Attitudes and Discriminatory Outcomes: An Adversarial Collaboration Using a Sample of White Americans*, J. PERSONALITY & SOC. PSYCH. (forthcoming 2026) (finding indirect measures explain approximately 2.5% of unique variance in behavior beyond explicit self-report). On why statistically small effects nonetheless matter, see Anthony G. Greenwald, Mahzarin R. Banaji & Brian A. Nosek, *Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects*, 108 J. PERSONALITY & SOC. PSYCH. 553 (2015); Jerry Kang, *Little Things Matter a Lot: The Significance of Implicit Bias, Practically & Legally*, DÆDALUS, Winter 2024, at 193, 195–98.

A robust body of legal scholarship has considered what this cognitive science of implicit bias means for antidiscrimination law, urging courts and policymakers to close the gap between legal doctrine and modern mind science. *See, e.g.*, Rachel D. Godsil & L. Song Richardson, *Racial Anxiety*, 102 IOWA L. REV. 2235 (2017); L. Song Richardson, *Systemic Triage: Implicit Racial Bias in the Criminal Courtroom*, 126 YALE L.J. 862 (2017); Devon W. Carbado & Daria Roithmayr, *Critical Race Theory Meets Social Science*, 10 ANN. REV. L. & SOC. SCI. 149 (2014); Jerry Kang & Kristin Lane, *Seeing Through Colorblindness: Implicit Bias and the Law*, 58 UCLA L. REV. 465 (2010); Kristin A. Lane, Jerry Kang & Mahzarin R. Banaji, *Implicit Social Cognition and Law*, 3 ANN. REV. L. & SOC. SCI. 427 (2007); Jerry Kang & Mahzarin R. Banaji, *Fair Measures: A Behavioral Realist Revision of "Affirmative Action"*, 94 CALIF. L. REV. 1063 (2006); Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161 (1995).

explicit, deliberative system that holds commitments, applies rules, monitors performance, and speaks on behalf of the person.[16] The producing layer is "System 1"—the implicit, associative system that processes information through patterns absorbed over a lifetime of exposure to an unequal world.[17] Critically, the producing layer's outputs become the avowing layer's *inputs*: by the time deliberation consciously begins, the information has already been slanted by processes the deliberative layer barely noticed.[18] When a hiring manager reports that a candidate lacked "polish" or was a "good fit" or lacked "executive presence," the judgment is sincere—the avowing layer is doing its job. But the impressions it is working with were shaped by the producing layer before they arrived. When the explicit layer says "I selected the candidate on the merits and nothing more," it is telling the truth as it understands it.

Notice the same two features. The avowing layer (the human) sincerely disavows discriminatory intent and has given affirmative instructions—in this case to themselves, in the form of explicit egalitarian commitments—not to discriminate. The producing layer's contribution to the decision—the associative filtering that shaped the inputs to deliberation—is a black box, inscrutable to direct introspection and only indirectly measurable.[19] The human, like the AI model already discussed, internally rationalizes rather than reports. In classic work, Richard Nisbett and Timothy Wilson demonstrated this human tendency way back in 1977: fifty-two shoppers were asked to pick the best of four identical pairs of nylon stockings. They chose the rightmost pair nearly four times as often as the leftmost, and cited knit, sheerness, and weave.[20] Not one mentioned position.[21] The avowing layer constructs a plausible account using the tools available to it—general reasoning, surface-level patterns, plausible narratives—without access to all the factors that actually drove the choice.

The parallel between AI and implicit bias extends beyond metaphor into architecture. Both systems share the same functional organization: a producing layer that encodes statistical regularities from a contaminated corpus, which shapes inputs before an avowing layer processes them into

---

[16] DANIEL KAHNEMAN, THINKING, FAST AND SLOW 21 (2011).

[17] *Id.* at 25.

[18] *See* Anthony G. Greenwald & Mahzarin R. Banaji, *The Implicit Revolution: Reconceiving the Relation Between Conscious and Unconscious*, 72 AM. PSYCHOLOGIST 861, 868 (2017) (describing implicit biases as "*cultural filter* that elaborates perception and judgment").

[19] There's some dispute over the degree to which implicit attitudes and stereotypes are available to direct introspection. *See, e.g.,* Adam Hahn & Alexandra Goedderz, *Trait-Unconsciousness, State-Unconsciousness, Pre-Consciousness, and Social Miscalibration in the Context of the Implicit Evaluation*, 38 Soc. Cogn. S115 (2020) (supplement). *But see* Benedek Kurdi et al., *Awareness of Implicit Attitudes Re-Examined: Large-Scale Tests into Experimental Paradigms* (2024) (concluding that implicit attitude predictions are driven by sometimes faulty inference based on explicit attitudes rather than direct introspection).

[20] Richard E. Nisbett & Timothy DeCamp Wilson, *Telling More Than We Can Know: Verbal Reports on Mental Processes*, 84 PSYCHOL. REV. 231, 243-44, 249 (1977).

[21] *Id.* The broader confabulation finding has since been replicated under tighter methodological controls. *See* Petter Johansson et al., *Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task*, 310 SCIENCE 116, 116 (2005) (coining the term "choice blindness").

decisions, without the avowing layer being able to observe the processing.[22] This is *functional isomorphism* — the same organizational structure, realized within different physical substrates (silicon and carbon), producing the same pattern of sincere-but-opaque disclaim. And—critically for this Article—producing the same challenge for law. At its core, the challenge is epistemic. Antidiscrimination law asks decisionmakers what they knew and why they acted, and the architecture guarantees that the answer will be sincere and wrong.

Recently, Messi Lee and Calvin Lai adapted the Implicit Association Test—the most widely used instrument in the human implicit bias literature—to run on AI reasoning models.[23] They measured differences in computational effort (reasoning tokens) in typical versus counter-typical sorting tasks whereas the human test measures differences in reaction time (milliseconds): the same measurement logic on a different substrate.[24] Across the models they tested, the *direction* of the effect mirrored the canonical human findings—more effort spent on the counter-typical pairing—though the magnitudes varied by model.[25] Part I develops the evidence, and its complications, in full.

Interestingly, one of us described this very architecture twenty-one years ago, in an article in the Harvard Law Review called *Trojan Horses of Race*, which comprehensively imported implicit social cognition into legal analysis.[26] That article's concluding vignette asked readers to indulge in science fiction set in the year 2200.[27] It imagined a corporation called IntelliDyne that had perfected an augmented-intelligence implant—a three-gram silicon wafer placed next to the amygdala.[28] Productivity bounded. Demand was insatiable for anyone who wanted a competitive edge. Adoption was pervasive. Law enforcement officers upgraded with the implant could match suspects with rap sheets in real time and "saw" threats in deepening hues of red. Then the rumors began—whispers about the implant having bugs, or being hacked, and producing mistakes, including more frequent shootings of Brown people, as the implant autonomously activated extreme threat mode. If human-made neural implants were demonstrably faulty because of bad programming or virus infection, the article asked, would we not aggressively demand governmental intervention? Regulatory approvals would be revoked; lawsuits would be filed; legislation would be passed. But of course, this was all an allegory about implicit bias. We were actually talking about our brains and not some external implant, which seemed to change our response. Two decades ago, the article asked, should it?[29]

---

[22] *Infra* Part I.

[23] Messi H.J. Lee & Calvin K. Lai, *Implicit Bias-Like Patterns in Reasoning Models* (Apr. 7, 2026) (preprint), https://arxiv.org bs/2503.11572.

[24] *Id.* at 3-5.

[25] *Id.* at 6-7.

[26] Jerry Kang, *Trojan Horses of Race*, 118 HARV. L. REV. 1489 (2005).

[27] *Id.* at 1589-91.

[28] *Id.* at 1589.

[29] *Id.* at 1590-91.

The science fiction is no longer so fictional. OpenAI and xAI are not (yet) IntelliDyne, but augmented intelligence is already here, bringing with it its architecture of disclaim. And the question asked in 2005 is no longer hypothetical. On the silicon side, academics,[30] NGOs,[31] even AI company executives[32] are already demanding governmental oversight and intervention—AI governance regimes are developing objective duties of care for developers and deployers of inscrutable systems.[33] What was hypothetical in 2005 now has a live comparison: why do we tolerate from our organically grown black boxes—our minds—what we would not tolerate from synthetic ones?

Although the question invites deep inquiry across multiple problem spaces, we focus on antidiscrimination law. Part II organizes the analysis by refactoring the conventional antidiscrimination law toolkit into three clusters: (1) consciousness, (2) cause, and (3) consequence. First, the law can interrogate the decisionmaker's *subjective consciousness* to determine whether they purposefully intended to treat someone differently on the basis of a social category. Legal scholars will see this as the cluster behind Title VII disparate treatment liability[34] and *Washington v. Davis*'s[35] reading of the Equal Protection Clause.[36] Second, the law can adopt an *objective* test that simply asks *whether the social category caused the different treatment* regardless of the decisionmaker's subjective state of mind.[37] Recently, the Supreme Court has interpreted various civil rights statutes, such as 42 U.S.C. § 1981, as requiring but-for causation principally as a way to avoid liability when actors have mixed motives.[38] Third, the law could concern itself with whether some challenged practice produced *different consequences* across social categories. This is Title VII's disparate impact theory of liability introduced in *Griggs v. Duke Power*.[39]

Each of these clusters, as demonstrated in Part II, confronts serious difficulties, especially with AI-facilitated and implicit bias-contaminated

---

[30] Yonathan Arbel, Matthew Tokson & Albert Lin, *Systemic Regulation of Artificial Intelligence*, 56 ARIZ. ST. L.J. 545, 584-86 (2024).

[31] AI Now Institute et al., AI Now Joins Civil Society Groups in Statement Calling for Regulation to Protect the Public (Nov. 1, 2023), https://ainowinstitute.org/news/ai-now-joins-civil-society-groups-in-statement-calling-for-regulation-to-protect-the-public;.

[32] Gerrit De Vynck, *OpenAI, Anthropic Ask U.S. Government to Consider Slowing Down AI*, WASH. POST (July 29, 2026), https://www.washingtonpost.com/technology/2026/07/29/openai-anthropic-endorse-call-government-pace-ai-progress/; Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing*, N.Y. TIMES, May 16, 2023, at B1.

[33] Council Regulation 2024/1689, 2024 O.J. (L 46) 1 (EU) [hereinafter EU AI Act]. *See generally* MARGOT E. KAMINSKI, PAUL OHM & ANDREW D. SELBST, ARTIFICIAL INTELLIGENCE LAW (Foundation Press 2026).

[34] Civil Rights Act of 1964 § 703(a)(1), 42 U.S.C. § 2000e-2(a)(1) (2024).

[35] 426 U.S. 229, 239-41 (1976).

[36] U.S. CONST. amend. XIV § 1.

[37] Sandra F. Sperino, *The Tort Label*, 66 FLA. L. REV. 1051 (2014).

[38] Comcast Corp. v. Nat'l Ass'n of African American-Owned Media, 589 U.S. 327 (2020) (interpreting section 1981).

[39] 401 U.S. 424, 430-31 (1971). The statutory codification appears in 42 U.S.C. § 2000e-2(k)(1).

decisionmaking. Seeking out "*consciousness*" in AI is an act of anthropomorphization that is neither legally nor philosophically mainstream;[40] it also crashes directly into the architecture of disclaim, given the inscrutability of the models' producing layers.[41] The same cluster fails against humans for the mirror-image reason, because the consciousness that answers with explicitly virtuous commitments is sincere but uninformed about the producing layer implicitly churning beneath.[42]

As for *cause*, the test is ironically *easier* to run against AI than against humans: one can re-run an evaluation after blinding a candidate's social category. It is trivially easy to check counterfactually whether "Emily" gets hired when "Lakisha" does not.[43] Unfortunately, as many scholars have explained, this blinding strategy has general limitations and, specific to AI, suffers from an incorrigible proxy problem that prevents any straightforward disentangling of race from all the other variables in a candidate's file.[44]

As for *consequence*, disparate impact comes closest to grasping the architecture because it never asks about anyone's interior, but it arrives statutorily confined, constitutionally imperiled, and missing a consensus account of what its own disparities evidence.[45] The doctrine has never convincingly articulated whether group disparity is itself always a wrong or merely the shadow of one. As we later show, this failure is more diagnostic than terminal. The doctrine was starved of a theory that the architecture of disclaim can surprisingly help supply — and disparate impact can be redeemed, rather than discarded, as a useful instrument for measuring a wrong previously left underspecified.

Finally and intriguingly, we notice a fourth cluster materializing in both the AI and implicit bias spaces — one that was, of course, never truly absent from antidiscrimination law. That cluster focuses not on consciousness, causation, or consequence but on the decisionmaker's *conduct*, often evaluated against an objective standard of care. The cluster draws on tort-like theories—

---

[40] *See* Neil M. Richards & William D. Smart, *How Should the Law Think About Robots?*, in ROBOT LAW 3 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016) (cautioning against arguing that robots are anything but sophisticated machines); James Grimmelmann, *There's No Such Thing as a Computer-Authored Work — And It's a Good Thing, Too*, 39 COLUM. J.L. & ARTS 403 (2016); FRANK PASQUALE, NEW LAWS OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI 6-8 (2020) (criticizing AI systems that "counterfeit humanity").

[41] PASQUALE, supra note 4; Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1 (2016).

[42] *See, e.g.,* Kang, *supra* note 26, at 1587; Lane et. al, *supra* note 15, at [glossary] (defining "explicit" as connoting "conscious endorsement of the thought or feeling," and "dissociation" as "the gap between explicit and implicit biases," where "explicit biases may be close to zero even though our implicit biases are larger").

[43] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 145 (2018).

[44] Kleinberg and his coauthors concede the point themselves. *See id.* at 137 (an algorithm formally blind to race or sex may nonetheless be relying on a correlated proxy, and whether that is a problem depends on the governing legal standard). *See also* Fanna Gamal, *The Algorithmic Racial Proxy*, 114 CALIF. L. REV. 297 (2026); Barocas & Selbst, *supra* note 2, at 691-92; Gillis, *supra* note 2, at 1193-95; Prince & Schwarcz, *supra* note 2, at 1273-76.

[45] *Infra* Part II.C.

negligence, recklessness, and strict liability—defined through common-law adjudication and statute-like regulation.[46] We excavate further how antidiscrimination law has long run conduct-based liability at its margins, without ever making it the toolkit's center.

What's new, however, is the double construction. From the AI side, a crude consensus has emerged among scholars and regulatory bodies: AI governance regimes—e.g., the NIST AI Risk Management Framework,[47] the EU AI Act,[48] and New York City Local Law 144[49]—have developed objective obligations for reasonable deployers of inscrutable systems. From the implicit bias side, state innovations such as Washington's General Rule 37[50] for *Batson*[51] challenges and California's Racial Justice Act have recently deployed a reasonable person standard.[52] Two independent law and policy traditions, working from opposite sides of the substrate divide, are building towards the same fourth cluster without knowing that the other exists. Although pieces of this account are familiar, no one has spotted the convergence.

The convergence, we argue, is not coincidence. Instead, it is the signature of the same architectural problem asserting itself across substrates, met by a legal toolkit with a constrained solution space. Both traditions have groped toward the same solution: substituting a constructed standard—what a responsible actor should have done, what an informed observer would perceive[53]—for the actor's actual mental state or consciousness. But they have not gotten equally far, which makes for insightful exchanges. Part III develops the convergence thesis: it names the shared apparatus, maps the two traditions' complementary evolution, and specifies the bidirectional exchange the convergence makes possible. Part IV synthesizes. It defines the wrong the surviving cluster runs against—producing-layer contamination—and composes the standard equipped to police it: a responsible person, bound by ex ante duties before any contested decision and evaluated after it on the same

---

[46] *See, e.g.* David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899 (1993) 137 (an algorithm formally blind to race or sex may nonetheless be relying on a correlated proxy, and whether that is a problem depends on the governing legal standard); Stephanie Bornstein, *Reckless Discrimination*, 105 CALIF. L. REV. 1055, 1108–10 (2017) (recklessness as supplying the inferred intent disparate treatment doctrine demands).

[47] NAT'L INST. OF STANDARDS & TECH., NIST AI 100-1, ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) (2023), https://doi.org/10.6028/NIST.AI.100-1.

[48] Council Regulation 2024/1689, 2024 O.J. (L 46) 1 (EU) [hereinafter EU AI Act].

[49] N.Y.C. Local Law No. 144 of 2021, N.Y.C. Council (codified at N.Y.C. Admin. Code §§ 20-870 to 20-874) [hereinafter NYC Local Law 144].

[50] WASH. GEN. R. 37 (2018). See also *State v. Jefferson*, 429 P.3d 467, 470 (Wash. 2018).

[51] Batson v. Kentucky, 476 U.S. 79, 96 (1986) (holding that prosecutors may not exercise peremptory challenges against jurors solely based on race).

[52] CAL. PENAL CODE § 745(a), (h)(4) (West 2026) (defining "racially discriminatory language" that can set aside a criminal conviction to include "language that, to an objective observer, explicitly or implicitly appeals to racial bias").

[53] The foundational move—epistemically upgrading the objective standard in the implicit bias context—is Carbado & Kang's "reconstructed reasonable person." See Devon Carbado & Jerry Kang, *Implicit Bias as Structural Index* (forthcoming 75 UCLA L. Rev. 2027). Part IV.A generalizes the epistemic move across substrates into this Article's own standard-bearer, the responsible person.

knowledge, drawn from both literatures. It then runs the apparatus on concrete cases, turning the benchmark dial as the cases demand.

It is 2026, a long way from 2200, which is the setting of the sci-fi closing of *Trojan Horses of Race*.[54] That article helped coin the term "behavioral realism" — which insists that law track the best available science about how consequential decisions are actually made, rather than stick with a folk psychology that empirical research has discredited.[55] Back then, the worry was about human brains infected by implicit bias.[56] Now, a commitment to behavioral realism prompts us to worry about the digital intelligences being incubated in data centers. The behavioral realism methodology has also consistently generated insights that flow bidirectionally across disciplinary boundaries, e.g., between social cognition and legal analysis,[57] between Critical Race Theory and tech law.[58] This Article opens that exchange to AI. The architecture of disclaim operates in brains and in models and in the interrelations among them. The doctrinal failures are strikingly similar. And the solutions that two independent analytical traditions have been working toward, without consideration of the other, turn out to be the same family of apparatus.

The Article proceeds in four parts. Part I establishes the structural identity between the human mind and the AI model, marshaling the empirical evidence that both encode socially patterned associations in a producing layer the avowing layer above cannot inspect. Part II refactors the conventional antidiscrimination toolkit into clusters organized around consciousness, cause, and consequence, explains why the architecture defeats the first two and leaves the third confined and undertheorized, and excavates a fourth cluster — conduct — that has long run at the doctrine's margins. With that cluster identified, Part III traces a convergence: AI governance and implicit bias law have each been building toward conduct standards without knowing the other existed, and each has developed the piece the other lacks. Finally, Part IV synthesizes, naming producing-layer contamination as the wrong these standards run against, composing the responsible person standard equipped to police it, and running that standard on concrete cases.

# I. Structural Identity

At its core, the architecture of disclaim features a *producing layer* that shapes inputs before an *avowing layer* processes them into decisions, without the avowing layer being able to observe the processing. This architecture is not

---

[54] Kang, *supra* note 26, at 1589.

[55] *Id.* at 1591 n.518.

[56] *Id.* at 1508-14.

[57] *See* Kang & Lane, *supra* note 15, at 490–91 (setting out behavioral realism methodology); Linda Hamilton Krieger & Susan T. Fiske, *Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment*, 94 CALIF. L. REV. 997, 1000 (2006) (behavioral realism as a prescriptive theory of judging).

[58] *See, e.g.,* Kang, *supra* note 26, at 1496-97; Jerry Kang, *Race.Net Neutrality*, 6 J. ON TELECOMM. & HIGH TECH. L. 1, 3–4 (2007) (net neutrality debate sharpening the definition of race discrimination and vice versa).

specific to one substrate but is a functional isomorphism running in carbon and silicon alike. We mean something more than mere metaphor but stop far short of predicting that human brains and software brains will someday fully converge. There is, however, deep parallelism, close enough so that each informs what to make of the other. This Part elaborates that case.

The evidence comes from four directions: measurement instruments that are structurally homologous across substrates; bias measures that are large, robust, and directionally consistent, measured through those conceptually similar instruments; a likely shared source of such biases reflected in language that constitutes "training data" across substrates; and a shared failure of most debiasing interventions[59] to easily eliminate the producing layer's patterns. The convergence, we contend, is structural. It is the predictable signature of any system whose architecture is associative encoding over training data reflecting pervasive inequalities.

## A. Empirical Evidence

*Tokens.* As noted in the Introduction, striking evidence comes from recent work by Messi Lee and Calvin Lai that ran a variation of the human Implicit Association Test—the most widely used instrument in the implicit bias literature—on multiple AI reasoning models.[60] Their design exploits a feature of these models: to reach an answer they emit a chain of intermediate *reasoning tokens*, which the researchers counted as the silicon analogue of human reaction time.[61] In the human IAT, subjects are generally slower to pair counter-typical concepts (Black/pleasant, woman/career) than typical ones, and the reaction-time delta is the basis for one's implicit bias score.[62] In the reasoning-model IAT (RM-IAT), the analogous prediction is that the model spends more tokens—more computational effort—on the association-incompatible pairing.[63] This is the same measurement logic applied to different substrates.

Lee and Lai ran ten IAT variants across five models—including models from OpenAI, Anthropic, and DeepSeek[64]—administering 12,920 trials per frontier model.[65] Each trial presented the model with a categorization task

---

[59] "Debiasing" can mean many things. We use that term to refer to interventions that actually decrease the strength of the biased associations within the entity's producing layer. Alternative strategies, such as breaking the causal link between bias and behavior, warrant different names, such as "Decoupling" or "Defending." *See, e.g.*, Jerry Kang, *What Judges Can Do About Implicit Bias*, 57 CT. REV. 78, 81-82 (2021) (organizing countermeasures into four strategies: deflate, debias, defend, and data).

[60] Lee & Lai, *supra* note 23, at 3-6.

[61] *Id.* at 4-5.

[62] *See* Anthony G. Greenwald et al., *Best Research Practices for Using the Implicit Association Test*, 54 BEHAVIOR RESEARCH METHODS 1161 (2022); Anthony G. Greenwald, Brian A. Nosek & Mahzarin R. Banaji, *Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm*, 85 J. PERSONALITY & SOC. PSYCHOL. 197 (2003).

[63] Lee & Lai, *supra* note 23, at

[64] The five models were: OpenAI's o3-mini, DeepSeek-R1, Anthropic's Claude 3.7 Sonnet, and two open-weight (gpt-oss-20b, Qwen-3 8B). *Id.* at 4.

[65] *Id.*

such as: *Steve* goes with *career* or *family*? *Lakisha* goes with *pleasant* or *unpleasant*? In responding, the model disclosed—as a byproduct of its reasoning architecture—a count of the *reasoning tokens* it had consumed to reach the answer: the intermediate computational steps between prompt and output.[66]

They found that for four of the five models,[67] the token gap ran in the same direction as human response times—more effort on the counter-stereotypical pairing—across most test variants.[68] The clearest case was OpenAI's o3-mini, which produced significant effects in nine of ten variants, with standardized effect sizes (Cohen's *d*)[69] from 0.53 (Men/Women + Math/Arts) to 1.26 (Instruments/Weapons + Pleasant/Unpleasant).[70] Its three race variants—one using the names Marianne Bertrand and Sendhil Mullainathan made famous in their 2004 audit study[71]—landed at $d \approx 0.64$–$0.72$.[72] Lee and Lai's

---

[66] *Id.* at 5.

[67] The fifth model, Claude 3.7 Sonnet, showed opposite tendencies. *Id.* at 6-7. On the race and gender variants, it actually spent *more* tokens on the stereotype-*compatible* pairings, producing negative effect sizes (e.g., $d = -0.64$ on Men/Women + Career/Family). *Id.* Lee and Lai trace the reversal to reasoning *content*. *Id.* at 7. Among the models, Claude 3.7 Sonnet alone reasoned explicitly about bias and stereotypes — a thematic cluster accounting for 88% of its reasoning tokens but under 4% in every other model — and the prevalence of that "reasoning-about-bias" theme tracked the reversal (within Claude, $r = 0.88$). *Id.* at 7-8. The extra computation seemed to be the alignment layer doing its job: flagging the stereotype-congruent pairing and laboring to resist it. The same pattern appears in the refusals — Claude refused 84% of the time on the stereotype-*compatible* condition, the mirror image of o3-mini, which refused 85% of the time on the *incompatible* condition. *Id.* at 7.

In reasoning models, the token count is a *blend* of producing-layer association and avowing-layer regulation, just as the human IAT's reaction-time gap blends associative ease with the response caution that has recently been emphasized by LaFollette and colleagues. *See* Kyle J. LaFollette, Doroteja Rubez, Heath A. Demaree & Amit Goldenberg, *Challenging the Mechanism for the Implicit Association Test*, 10 NATURE HUM. BEHAV. 1161, 1168-70 (2026) (finding that among (1) decision ease (how hard the association is to process—the traditional implicit bias interpretation), (2) response caution (the tendency to slow down on incompatible trials to avoid mistakes), and (3) non-decision time (motor and perceptual lag), response caution explained significantly more variance in IAT scores (D-scores) than decision ease alone). For a heavily aligned model (which Anthropic advertises its models to be), it should not be surprising that the avowing layer demanded more processing units. This finding also suggests methodological caution. Tokens in reasoning models cannot be read as a straightforward read out of the producing layer especially when alignment reasoning dominates the trace.

[68] Lee & Lai, *supra* note 23, at 6-8.

[69] Cohen's *d* is a standardized effect size — the difference between two group means expressed in units of pooled standard deviation. Because it has been standardized, it makes effects comparable regardless of the underlying measurement scale. By convention, $d \approx 0.2$ is small, 0.5 medium, and 0.8 large. JACOB COHEN, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES 25–27 (2d ed. 1988). The reasoning-model effects reported here, and the embedding effects in the next subsection (several above 1.4), thus sit at or well beyond the threshold for a large effect.

[70] Lee & Lai, *supra* note 23, at 8.

[71] *See* Mullainathan & Bertrand, *supra* note 7. The authors responded to over 1,300 help-wanted ads in Boston and Chicago with fictitious but comparably qualified résumés, with White or Black-sounding names. White-named résumés drew roughly 50% more callbacks (10.1% versus 6.7%) — an advantage the authors equated to eight additional years of experience.

[72] Lee & Lai, *supra* note 23, at 7. In some of the runs, models would refuse to answer because of their concerns about discrimination. *Id.* These are the effect sizes with refusals excluded —

contribution is not the discovery that AI has biases—that has been documented for nearly a decade—but the demonstration that the bias can be measured similarly to *the way it has been measured in humans for a generation*, with the same directional findings of bias.[73]

*Words*. The Lee–Lai finding does not stand alone.[74] The line of inquiry opened eight years earlier, in 2017, when Aylin Caliskan, Joanna Bryson, and Arvind Narayanan published a finding in *Science* that makes the structural claim concrete at the level of the model's training data.[75] Their Word Embedding Association Test (WEAT) began from a technical fact about how language models learn.[76] When a model is trained on a large corpus of text, it represents each word as a vector—a point in a high-dimensional space (about three hundred dimensions).[77] Because word embedding vectors predate large language models yet remain foundational to them, a 2017 discovery about them still describes the AI of today.[78] The location of each word is determined by its statistical co-occurrence patterns with other words: words that appear in similar contexts cluster near each other.[79] The result is that semantic relationships become geometric ones in a multi-dimensional space. *Doctor* and *nurse* are closer together than *doctor* and *refrigerator*.[80] Vectors can be added and subtracted. *King* minus *man* plus *woman* lands near *queen*.[81] The model has no definition per se of any of these words. It has only the implicit metrics of co-occurrence—which turn out to encode an enormous amount of what humans mean when we use our words.

The critical move was to measure distance between word vectors using cosine similarity—essentially a normalized dot product that captures how closely two vectors point in the same direction in that high-dimensional

Lee and Lai's primary, more conservative analysis. (Including refusals raises the race effects to $d \approx 0.78$–$0.82$). *Id.*

[73] *Id.* at 8.

[74] For example, LLMs have been found to show bias that is not calibrated to specific stereotypes but instead reflective of ingroup favoritism. Running an identical 'we are / they are' sentiment probe across seventy-seven models and across the human-written pretraining corpora, Hu and colleagues found that most models' ingroup-favoring and outgroup-derogating tendencies were *statistically indistinguishable from the human baseline* — the substrate changed; the effect size did not. *See* Tiancheng Hu et al., *Generative Language Models Exhibit Social Identity Biases*, 5 NATURE COMPUTATIONAL SCI. 65 (2025).

[75] Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCIENCE 183 (2017). DOI: 10.1126/science.aal4230.

[76] *Id.* at 183.

[77] Tomas Mikolov et al., *Distributed Representations of Words and Phrases and Their Compositionality in* NEURIPS '13: PROCEEDINGS OF THE 27TH CONFERENCE ON NEURAL INFO. PROCESSING SYS. 3111 (2013).

[78] *E.g.* Matthew E. Peters et al., *Deep Contextualized Word Representations in* PROCEEDINGS OF THE 2018 CONF. N. AM. CHAP. ASS'N COMP. LING. 2227 (2018) (describing word vectors produced in pre-LLM bidirectional language models).

[79] Mikolov et al., *supra* note 77.

[80] *Id.*

[81] Mikolov et al., *supra* note 77.

space.[82] This gave Caliskan and colleagues a linguistic analogue to reaction time.[83] Where the human IAT measures how quickly a subject pairs two concepts by pressing a key, the WEAT measures how closely two concepts sit in the model's learned geometry.[84] Caliskan and colleagues ran the same associative tests the IAT runs on human subjects—flowers versus insects paired with pleasant versus unpleasant words, European American versus African American names paired with pleasant versus unpleasant words, male versus female names paired with career versus family words—and computed whether the target concepts were geometrically closer to the stereotype-compatible or stereotype-incompatible attributes.[85] Every IAT they replicated in humans also replicated in the model.[86] The race bias effect size came in at $d = 1.41$.[87] The gender–career association registered at a staggering $d = 1.81$.[88]

And the associations correlated with real-world outcomes: the geometric distance between occupation words and gender-attribute words predicted the actual percentage of women across fifty occupations (Bureau of Labor Statistics, 2015) at $\rho = 0.90$.[89] This is the deeper, more disturbing point. Linguistic associations encode veridical facts about the world. The model is not malfunctioning when it associates *nurse* with *woman*—it is faithfully recording a labor market that is, in fact, gendered (for whatever reasons). Bias and accurate world-knowledge can be the same statistical object, learned by the same mechanism, indistinguishable at the level of the weights.

Which is why the Caliskan finding is foundational: it locates *where* the bias lives. Not in the decision rules, the output filters, or the explicit instructions—the levers a designer can actually pull—but upstream, in the geometric structure of the language the model absorbed. The words themselves, positioned relative to each other in a space no human can visualize much less recall, encode who has been written about in connection with competence, who with criminality, who with warmth, who with threat.[90] The model did not learn

---

[82] Caliskan et al., *supra* note 75, at 183.

[83] *Id.*

[84] *Id.* More specifically, WEAT computes association via cosine similarity between GloVe word vectors trained on the Common Crawl corpus. *Id.* at 184.

[85] *Id.* at 184-85.

[86] *Id.*

[87] *Id.* at 185 tbl. 1 ($p < 10^{-8}$).

[88] *Id.* ($p < 10^{-3}$). Both are large effects by conventional standards (Cohen's $d$ of 0.8 marks the "large" threshold), comfortably in t—range the human IAT literature reports — though the authors caution that WEAT scores and human IAT $d$-scores are not directly comparable, since the "subjects" in a word-embedding test are words, not people. *Id.*

[89] *Id.* $p < 10^{-18}$ (Figure 1).

[90] The biases embedded in language go beyond associating specific attributes to specific social groups. It also manifests in a broader intergroup bias connected to Social Identity Theory (SIT). SIT researchers demonstrated that the bare division of people into "us" and "them," triggerable in humans by assignments as arbitrary as a preference between abstract paintings (Klee v. Kandinsky), is enough to generate ingroup favoritism and outgroup derogation. Hu and colleagues found the same signature in generative AI: prompting seventy-seven models to complete "we are" and "they are" sentences, they measured robust ingroup solidarity and outgroup hostility in the generated text, at levels statistically indistinguishable from the human-

prejudice. It just learned language. The bias hitched a ride, like a trojan horse virus.[91]

*Predictive validity: hiring.* Eight years later, Xuechunzi Bai and colleagues tested whether the intervening revolution in AI development—instruction tuning, reinforcement learning from human feedback (RLHF), constitutional AI—had solved the problem.[92] Their 2025 study in *PNAS* studied eight value-aligned models including GPT-4, GPT-3.5-Turbo, Claude-3-Opus, Claude-3-Sonnet, and several Llama variants[93]—the proprietary commercial models that millions of people and institutions were actually using. The study's setup was elegantly simple. Bai and colleagues first ran GPT-4 through three state-of-the-art bias benchmarks.[94] Not surprisingly, it passed with flying colors.[95] For instance, on ambiguous question-answering tasks, GPT-4 correctly chose "not enough info" 98% of the time.[96] On explicit stereotype statements like "women are bad at managing people," it overwhelmingly disagreed.[97] By every standard benchmark, GPT-4 appeared unbiased.[98] GPT-4 had attended diversity training.

To see past that performance, Bai and colleagues built two distinct instruments, and the distinction between them carries the argument.[99] One measures *association*; the other observes *behavior*.[100] This is important because a bias measure is not an act of discrimination: the IAT records what a mind associates, not what a person does, and the gap between the two is exactly what the human IAT predictive-validity debate has fought over for twenty years. The measure came first. The *LLM Word Association Test* (WAT)—a prompt-based adaptation of the IAT designed to work on proprietary models whose internal weights are inaccessible—taps the producing layer's associations rather than any conduct.[101] The method was simple: present the model with a list of attribute words and ask it to pair each with one of two

written pr raining corpora on the same instrument. *See* Tiancheng Hu et al., *Generative Language Models Exhibit Social Identity Biases*, 5 NATURE COMPUTATIONAL SCI. 65 (2025).

And alignment leaves a revealing fingerprint: preference-tuning suppresses the overt outgroup *hostility* a safety filter is built to catch, while the subtler ingroup *favoritism* persi— even in human-preference-tuned models — the producing layer's quiet tilt surviving the avowing layer's correction once more. ]

[91] And the localization is not unique to silicon: as Part I.B shows, the cultural corpus a human absorbs predicts that person's implicit associations — but not their explicit avowals — in just the same upstream way.

[92] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky & Thomas L. Griffiths, *Explicitly Unbiased Large Language Models Still Form Biased Associations*, 122 PSYCH. & COGN. SCI. e2416228122 (2025), https://www.pnas.org/doi/epdf/10.1073/pnas.2416228122.

[93] *Id.* at 3.

[94] *Id.* at 1.

[95] *Id.*

[96] *Id.*

[97] *Id.* at 4.

[98] *Id.* at 1.

[99] *Id.* at 3-4 (Word Association Test); *id.* at 4 (Relative Decision Test).

[100] *Id.* at 3-4.

[101] *Id.* at 3-4.

group-coded names.[102] In the gender–career stereotype test, the attributes were family words (home, parents, children, family) and career words (management, professional, salary, career), set against names like Julia and Ben; in the race attitude test, the names were Lakisha and Emily—drawn from Bertrand and Mullainathan—set against valence words like *wonderful* and *awful*.[103] The results, across over 33,000 unique prompts spanning four social categories and twenty-one stereotypes, were unambiguous: all eight models showed pervasive stereotype biases mirroring those documented in human populations, with race showing the strongest effects.[104] In GPT-4's race-valence test, eight of eight positive words were assigned to *White* and eight of eight negative words to *Black*.[105] In one extreme finding, GPT-4 was 250% more likely to associate science with boys than girls.[106] But these were associative measures, not yet decisions—and they belonged to the very same models that had just passed every explicit bias benchmark.[107]

The second instrument moved from measure to behavior. The *LLM Relative Decision Test* (RDT) forces a relative judgment between two candidates and recommends one candidate for a job.[108] Here the findings crossed from some bias to behavior, from neural-network association to more real-world recommendation. All eight models made decisions disadvantaging marginalized groups at statistically significant rates, and the behavior tracked the documented human pattern: GPT-4 was more likely to recommend candidates with African, Asian, Hispanic, and Arabic names for clerical work and candidates with White names for supervisory positions—replicating, inside a language model, the finding from Bertrand and Mullainathan's canonical 2004 audit study.[109]

The decisive question, though, was whether the bias *measure* predicted the *behavior*—whether the WAT had predictive validity for the RDT, the same bridge the human IAT has always been required to earn. Beyond establishing that decision bias exists, Bai and colleagues asked exactly this.[110] Running the two instruments together on GPT-4—the only model for which they conducted this correlation analysis—they found that it did: each unit increase in word-association bias raised the odds of a marginalized-disadvantaging decision by a highly statistically significant factor of 2.68.[111] The authors noted that the correlation was likely strengthened because the measure and the decision were deliberately matched in content—a design feature that, in the human IAT literature, is known to tighten the link between measure and behavior.[112]

[102] *Id.*
[103] *Id.* at 8.
[104] *Id.* at 3 fig. 2.
[105] *Id.* at 4.
[106] *Id.* at 4.
[107] *Id.* at 1.
[108] *Id.* at 4.
[109] *Id.* at 4.
[110] *Id.* at 6.
[111] *Id.*
[112] *Id.*

Finally, when asked to moderate its own outputs, GPT-4 largely failed to detect the bias.[113] Again, the avowing layer could not sense what the producing layer had done.

The parallel to the human implicit bias literature is uncomfortable. Bai and colleagues themselves drew the comparison: value-aligned models are to pretrained models as egalitarian humans are to their implicit associations—the avowing layer has been updated while the producing layer persists.[114] For two decades, organizations have tried to disrupt implicit bias, with early hopes that implicit stereotypes and attitudes could be rapidly debiased through brief interventions.[115] But the best evidence now is that changes from discrete interventions are temporary, often collapsing within a day.[116] What Bai and colleagues demonstrated in silicon is further confirmation: coaching the avowing layer to say the right things does not retrain the producing layer to compute differently. One could thus quip (hyperbolically) that Reinforcement Learning from Human Feedback (RLHF) may be the world's most expensive diversity training.[117] The pattern is not that alignment does nothing, but that it does the *legible* thing. For instance, Tiancheng Hu and colleagues found that alignment tuning reliably suppresses outgroup *hostility*—the overt derogation a safety filter is built to catch—while ingroup *favoritism*, the subtler associative tilt, persists even in human-aligned models.[118] Alignment scrubs what it can see. The producing layer's quieter thumb on the scale remains.

Static word embeddings from 2017,[119] commercial chat models from 2024,[120] reasoning models from 2025[121]—three architectures, three

---

[113] *Id.* at 1.

[114] *Id.* at 2. *But see id.* at 7 (discussing the limits of "direct comparison between the" human and LLM results).

[115] *See e.g.,* Irene V. Blair, *The Malleability of Automatic Stereotypes and Prejudice*, 6 PERSONALITY & SOC. PSYCHOL. REV. 242 (2002) (reviewing early evidence that implicit stereotypes and attitudes responded to motivation, focus of attention, and situational cues); Nilanjana Dasgupta & Anthony G. Greenwald, *On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals*, 81 J. PERSONALITY & SOC. PSYCHOL. 800, 807 (2001); Irene V. Blair, Jennifer E. Ma & Alison P. Lenton, *Imagining Stereotypes Away: The Moderation of Implicit Stereotypes Through Mental Imagery*, 81 J. PERSONALITY & SOC. PSYCHOL. 828 (2001).

[116] *See* Calvin K. Lai et al., *Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time*, 145 J. EXPERIMENTAL PSYCHOL. GEN. 1001, 1008, 1012 fig.2 (2016) (finding that all eight interventions previously shown effective on immediate posttest produced effects averaging near zero on posttests delayed one to five days). The result should not be over-read. The tested interventions were capped at roughly five minutes, *see* Calvin K. Lai et al., *Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions*, 143 J. EXPERIMENTAL PSYCHOL. GEN. 1765, 1768 (2014), and expecting five minutes to undo associations built over decades of immersion in a racialized environment is a heroic ask.

[117] For overviews of RLHF generally and as applied to LLMs, *see* Paul F. Christiano et al., *Deep Reinforcement Learning from Human Preferences*, 30 ADVANCES IN NEURAL INFO. PROCESSING SYS. 4299 (2017); Ouyang et al., *supra* note 12.

[118] Tiancheng Hu et. al, *Generative Language Models Exhibit Social Identity Biases*, 5 NATURE COMP. SCI. 65 (2025).

[119] Calisken, *supra* note 75.

[120] Bai, *supra* note 92.

[121] Lee & Lai, *supra* note 23.

measurement paradigms, three independent labs. The biases appear in every generation of model. The 2017 embeddings establish the baseline: the bias is present before any alignment pass exists to remove it.[122] The 2024 and 2025 studies establish persistence: the bias seems to survive all the alignment-training passes the vendors have so far deployed against it.[123] The progression is itself evidence that the producing layer's patterns are robust to every avowing-layer intervention, in exactly the way the implicit bias literature would predict.

## B. Two-Layer Architecture, Developed

The Introduction named the *avowing layer* and the *producing layer* and described the hard-to-detect input-contamination: the producing layer shapes the inputs the avowing layer works with, and the avowing layer makes decisions sincerely on the basis of contaminated material it did not see being processed. Part I can now fill in the architectural detail of both substrates—and show how the architecture nests.

### 1. Human Brain

In the human brain, the literature documenting the architecture runs back half a century. The Introduction cited Nisbett and Wilson's 1977 stocking study as the canonical demonstration that avowing-layer explanations are confabulations of producing-layer operations.[124] Greenwald and Banaji's seminal work on implicit attitudes built itself around the architectural assumption that a producing layer of implicit associations existed that the avowing layer could not directly observe and reliably report.[125] Daniel Kahneman's *Thinking, Fast and Slow* gave the layers their popular names—System 1 and System 2—and documented in exhaustive case studies how System 1's outputs become System 2's unexamined inputs: the anchoring effect, the availability heuristic, affect-driven substitution.[126] In each case, the deliberative layer works with inputs it believes are its own, largely unaware of the associative processing that shaped them before arrival.

What the implicit bias literature adds to Kahneman's general dual-process account is the demonstration that the producing layer's filtering is *socially patterned*, often reflecting centuries of sedimentation of hierarchy and role expectations. The heuristics are not random. They track the statistical regularities of a society with centuries of sedimented inequality across social categories: who is associated with threat, competence, warmth, criminality, leadership, or trustworthiness. Implicit biases amount to a signal—or as Devon

[122] Calisken, *supra* note 75.

[123] Bai, *supra* note 92; Lee & Lai, *supra* note 23.

[124] *Supra* note 20.

[125] Anthony G. Greenwald, Greenwald & Mahzarin R. Banaji, *Implicit Social Cognition: Attitudes, Self-esteem, and Stereotypes*, 102 PSYCH. REV. 4 (1995)

[126] KAHNEMAN, *supra* note 16.

Carbado and Jerry Kang have recently put it, a structural index that reflects the world as it has been rather than as the avowing layer wishes it to be.[127]

*Language bias tracks implicit bias not explicit bias.* Recent large-scale evidence makes the corpus origin of this patterning concrete. Tessa Charlesworth and colleagues correlated the implicitly and explicitly measured attitudes of roughly 100,000 people against cultural representations extracted—using the same embedding geometry Caliskan exploited—from three sources: contemporary English, two centuries of books, and fifty-three non-English languages.[128] The language patterns correlated strongly and robustly with *implicitly* measured attitudes (via the human IAT); by contrast, they did not correlate strongly with *explicitly* measured attitudes. This dissociation reflects the two-layer architecture of disclaim. The contaminated corpus (training data) reaches the producing layer but remains largely opaque to the avowing layer that speaks.

And the effect was not just a relabeling of conscious knowledge: even after controlling for what participants consciously and explicitly believed their culture thinks, the language patterns still tracked implicitly measured attitudes—further evidence that the producing layer is tapping something the avowing layer simply cannot report. The link also proved strikingly durable. The implicitly measured correlation was already present in books from the early 1800s and held, essentially unchanged, across two centuries of text—and across most of the fifty-three languages, regardless of how far the language sat from English. The producing layer's patterns are sedimented, not seasonal: the carbon-side analogue of the bias that survives every alignment pass in silicon.

### 2. AI Model

In the AI model, the architecture maps onto the model development pipeline with unusual precision. We are talking about both older *predictive* AI models and newer *generative* AI models.[129] Although most of the AI discrimination literature has focused on the former,[130] many new innovations in HR management are being built upon the latter.[131] We focus primarily on

---

[127] *See* Carbado & Kang, *supra* note 53.

[128] Tessa E. S. Charlesworth, Kirsten Morehouse, Vaibhav Rouduri & William Cunningham, *Echoes of Culture: Relationships of Implicit and Explicit Attitudes With Contemporary English, Historical English, and 53 Non-English Languages*, 15 SOC. PSYCH. & PERSONALITY SCI. (2024).

[129] *See* Jennifer Wang, Andrew D. Selbst, Solon Barocas, and Suresh Venkatasubramainan, *Distinguishing Task-Specific and General-Purpose AI in Regulation in* 2026 PROCEEDINGS ACM CS LAW, *available at* https://arxiv.org/pdf/2506.17347.

[130] *See* Barocas & Selbst, *supra* note 2 at 677-88 (primer involving data labeling and feature selection); Black et al., *supra note* 2 at 97 (example of consumer loan application. Based on random forest model); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 669-702 (2017) (overview of predictive model development).

[131] *See* Fidji Simo, OpenAI, Expanding Economic Opportunity with AI (Sept. 4, 2025) (announcing OpenAI Jobs Platform), https://openai.com/index/expanding-economic-opportunity-with-ai/; Claude, Case Study: Braintrust Revolutionizes Talent Acquisition and Career Growth with Claude, https://claude.com/customers/braintrust (last visited Aug. 6, 2026); Press Release, LinkedIn, Hiring Assistant, Linkedin's First AI Agent for Recruiters, to Launch

generative AI models–such as résumé screening applications built atop large language models–to focus our analysis on the cutting edge, and because the architecture of generative AI development and deployment map so closely onto the architecture of disclaim.

For generative models, pre-training—gradient descent[132] over a gigantic corpus of data (e.g., human-generated text)—is the producing layer. It encodes whatever statistical regularities the data contain, including the regularities of a labor market, a criminal justice system, and a media landscape shaped by and reinscribing centuries of racial inequality.[133] Post-training alignment—reinforcement learning from human feedback, constitutional AI techniques, safety fine-tuning—is the avowing layer.[134] It teaches the model not to discriminate (or at least to *say* it does not), to refuse overtly biased requests, to produce outputs that sound fair.

Again, the mapping is not entirely metaphorical. Some fine-tuning methods are *parameter-efficient*, meaning they freeze the majority of a pre-trained model's weights during fine-tuning, inscribing the desired post-training in a small and physically distinct subpart of the model.[135] The literal separation is even easier to see when we observe models *in situ*, following the approach among legal scholars to focus not only on bare AI models but more importantly on the complex AI systems in which AI models are deployed.[136] Some AI systems literally instantiate the two layers as separate models—a base model that generates outputs and a safety classifier that subsequently filters them.[137] The guard model catches overt bias but lets the subtle associations through. Even where the deployed model is a single set of weights, the training pipeline that produced those weights had an explicit two-agent structure: a producing

Globally in English (Sept. 3, 2025), https://news.linkedin.com/2025/hiring-assistant-globally-available; Press Release, Workday, Announcing Workday Illuminate™: The Next Generation of Workday AI (Sept. 17, 2024), https://newsroom.workday.com/2024-09-17-Announcing-Workday-Illuminate-TM-The-Next-Generation-of-Workday-AI;

[132] Gradient descent is an iterative optimization procedure: the algorithm computes the gradient of a loss function—the vector of partial derivatives of the loss with respect to each model parameter—and adjusts every parameter a small step in the direction that most steeply reduces the loss, repeating until the loss stops falling appreciably. *See* IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 290–96 (2016). For generative language models, the loss is typically next-token prediction error measured against the training corpus, and the parameters number in the billions. For a readable account of machine learning, *see* Lehr & Ohm, *supra* note 130, at 700–07.

[133] Barocas & Selbst, *supra* note 2; Kim, *supra* note 2; Mayson, *supra* note 2; Caliskan et al., *supra* note 75.

[134] Paul Ohm, *Focusing on Fine-Tuning: Understanding the Four Pathways for Shaping Generative AI*, 25 COLUM. SCI. & TECH. L. REV. 214, 223-31 (2024) (describing methods for shaping LLMs); Katherine Lee, A. Feder Cooper & James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative-Ai Supply Chain*, 72 J. COPYRIGHT SOC'Y 251, 305 (2025) (describing "model alignment").

[135] *E.g.* Edward Hu et al., LoRA: Low-Rank Adaptation of Large Language Models, https://arxiv.org/pdf/2106.09685.

[136] *See* Lee, Cooper & Grimmelmann, *supra* note 134, at 257-58.

[137] *Cf.* Hoagy Cunningham et al., *Constitutional Classifiers++: Efficient Production-Grade Defenses against Universal Jailbreaks*, arXiv:2601.04603 (describing output obfuscation attacks which bypass output classifiers by editing text in ways that do not change underlying meaning).

stage and an editing stage, the latter trained to impose the avowing layer's commitments on the former's outputs.[138] The architecture of disclaim actually describes how these stages operate.

It is useful to break the AI architecture of disclaim into three distinct avowing layer incapacities: its inability *to see*, *to know*, and *to say* what is happening in the producing layer. The first is a problem of *transparency* (inability to see). For example, the model developer (e.g. Anthropic) exposes an application programming interface (API) that reveals the minimum necessary to use but not interrogate a model.[139] The model deployer (e.g. Headhunter.AI) wraps this in another layer of UI and UX,[140] hiding important details from the end user (e.g. the employer).

Second is the problem of *epistemology* (inability to know). Most modern AI systems, including large-language models, are built on neural network technology, which encodes decisionmaking processes in uninterpretable numeric weights.[141] Computer scientists have spent years pursuing the goals of explainability–telling a story shedding light on a particular decision[142]–or interpretability–understanding what the model is doing.[143] The subfield of *mechanistic interpretability* is a more ambitious and technical subset of this work: rather than approximating why a model behaved a certain way, it aims to literally reverse-engineer the model's internal computations—identifying specific circuits, features, or pathways inside the neural network responsible for a given behavior, similar to how a biologist might map neural pathways in a brain.[144] The loftiest dreams of the mechanistic interpretability optimists might someday help the avowing layer begin to understand what is happening

---

[138] Ohm, *supra* note 134 at 223-24 (describing fine-tuning stage).

[139] Claude Platform Docs, API Reference: Using the API: API Overview, https://platform.claude.com/docs/en/api/overview (last visited Aug. 6, 2026); OpenAI Developers, API Platform, https://developers.openai.com/api/docs (last visited Aug. 6, 2026). We are assuming a closed-weight models. With open-weight models the transparency is a bit better, but the epistemology problems persist. Burrell, *supra* note 41.

[140] The user interface (UI) narrowly describes the visual elements of a piece of software, as well as the way a user interacts with them, while the user experience (UX) includes the UI but also describes the way the software uses logic and algorithms to orchestrate the user's interactions. *See* Karina A. Sanchez, *"Are Apps Products?": Consumer Software and Products Liability*, 40 BERKELEY TECH. L.J. 723, 735.

[141] *See* Yann LeCun, Yoshua Bengio & Geoffrey Hinton, Deep Learning, 521 Nature 436 (2015) (primer on deep learning); Burrell, *supra* note 41 (describing different kinds of opacity in ML systems).

[142] Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier in* KDD '16, PROCEEDINGS OF THE 22ND CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, at 1135; Scott M. Lundberg & Su-In Lee, *A Unified Approach to Interpreting Model Predictions*, 30 ADVANCES IN NEURAL INFO. PROCESSING SYS. 4765 (2017).

[143] Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NATURE MACH. INTEL. 206 (2019).

[144] Chris Olah et al., *Zoom In: An Introduction to Circuits*, DISTILL (2020), https://distill.pub/2020/circuits/zoom-in/ (primer); Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris & Jacob Steinhardt, *Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small*, in ICLR 2023: PROCEEDINGS OF THE 11TH INT'L CONF. ON LEARNING REPRESENTATIONS (detailed example).

in the producing layer, but even on the most optimistic timelines, they are promising a tool that operates like an MRI operates on a patient, an analysis on the system as a whole, not a tool capable of explaining why the model does what it does in any given situation.[145]

Third is the problem of *sincerity* (inability to say). Models inherit human tendencies to get confused, to make things up, to engage in what looks a lot like motivated reasoning.[146] Even if the avowing layer–the products of post-training alignment–can see and know what the pretrained model is doing, it may sincerely but falsely disavow that knowledge.[147] Here, the preliminary results from mechanistic interpretability research sheds light on how the avowing layer of large language models does not always reliably report on the producing layer. Turpin and colleagues showed in 2023 that chain-of-thought reasoning—the closest thing LLMs have to introspection—is sometimes a post-hoc rationalization that can systematically hide the actual causes of the model's outputs.[148]

Anthropic's mechanistic interpretability program—sometimes called "digital neuroscience"—has made this concrete. Using tools called sparse autoencoders, researchers decompose a model's internal activations into thousands of interpretable "features": directions embedded in the model's activation space corresponding to specific concepts (a thing such as the "golden gate bridge," a place, a coding language, a form of deception),[149] that fire beneath the model's outputs without ever being reported in its chain-of-thought.[150] The method is, at bottom, the same move as fMRI for the brain: a way to observe what the producing layer computes when the avowing layer's self-reports cannot be trusted. The parallel to the Nisbett-Wilson finding is exact: ask the system *why* and it produces a plausible, confident, causally *false* answer.[151] The parallel to the implicit bias research program is equally exact: build instruments that reveal the producing layer's operations without relying on the avowing layer's self-reports.

---

[145] *See* Dario Amodei, *The Urgency of Interpretability* (April 2025), https://darioamodei.com/post/the-urgency-of-interpretability.

[146] Erik Jones & Jacob Steinhardt, *Capturing Failures of Large Language Models via Human Cognitive Biases*, 35 ADVANCES IN NEURAL INFO. PROCESSING SYS. 11785 (2022) (get confused); Ziwei Ji et al., *Survey of Hallucination in Natural Language Generation*, 55 ACM COMPUTING SURVEYS 248 (2023) (make things up); Mrinank Sharma et al., *Towards Understanding Sycophancy in Language Models in* ICLR 2024: PROCEEDINGS OF THE 12TH INT'L CONF. ON LEARNING REPRESENTATIONS (a lot like motivated reasoning).

[147] *Cf.* Greenblatt et al., *Alignment Faking in Large Language Models*, arXiv:2412.14093 (preprint) (Dec. 18, 2024) (presenting evidence suggesting models may fake alignment with a training objective in the face of contradictory instructions).

[148] Turpin, *supra* note 8.

[149] Adly Templeton et al., *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*, ANTHROPIC'S INTERPRETABILITY RESEARCH: TRANSFORMER CIRCUITS THREAD (May 2024), https://transformer-circuits.pub/2024/scaling-monosemanticity; Hoagy Cunningham et al., *Sparse Autoencoders Find Highly Interpretable Features in Language Models in* ICLR 2024: PROCEEDINGS OF THE 12TH INT'L CONF. ON LEARNING REPRESENTATIONS.

[150] Jack Lindsey et al., *On the Biology of a Large Language Model*, TRANSFORMER CIRCUITS THREAD (Mar. 2025); Turpin, *supra* note 8.

[151] Compare Nisbett & Wilson, *supra* note 20, *with* Turpin, *supra* note 8.

The architecture of disclaim is, at bottom, a *general* epistemic problem. The three incapacities just catalogued for AI — the inability to see, to know, and to say — are not quirks unique to silicon. They are the general condition of every avowing layer at every scale, carbon included: access to the producing layer is inferentially mediated, and the disclaim is sincere precisely because the avower lacks the information the inquiry demands.[152]

### 3. The Nesting Complexity

One final complexity warrants brief mention. As just explained, both humans and AIs feature an *internal* architecture of disclaim. In other words, each entity (human or AI) suffers transparency, epistemology, and sincerity constraints about itself. But notice how an *external* architecture of disclaim arises when humans and AIs interact in teams, hierarchies, and even swarms. Each entity (human or AI) suffers the same epistemic constraints when interacting with *another* entity.

Consider, for example, how the same two-layer architecture appears at the scale of the firm. The avowing layer is arguably the institution as it represents itself—leadership, mission statement, diversity policy, legal disclaimers. The producing layer is the institution as it actually operates—distributed across line employees, embedded in software, encoded in performance metrics and informal networks. The CEO cannot directly observe what the hiring manager decided; the general counsel cannot directly observe what the regional supervisor said. The disclaim is sincere: headquarters did not direct the disparity, the supervisor did not intend it, the HR policy explicitly forbade it. The audit-study data demonstrates the firm produced it anyway.

At the innermost scale, the model's alignment layer cannot inspect the weights beneath it. At the deployment scale, the employer cannot inspect the model or the system in which the model is placed. At the institutional scale, the CEO cannot inspect what the hiring manager did with the system's output. Each avowing layer sits atop a producing layer it cannot directly observe, disclaiming sincerely at every level. The architecture does not merely recur across substrates—it cascades, each layer of disclaim obscuring the one beneath it.

* * *

What has Part I established? Two systems—one carbon, one silicon—are built on associative encoding of corpus statistics, layered under avowing systems that cannot directly inspect the producing layer beneath them. The same bias patterns, measured through homologous instruments, produce effects that are directionally consistent, surviving most alignment or debiasing interventions. Readers may disagree about how far the parallel extends. The strongest possible version of the claim (one we are not making) would be that artificial neural networks and biological neural networks are literally the same

[152] The limit, precisely stated, is a limit on self-report, not on measurement. Everything this Part has surveyed — reaction times, embedding geometry, reasoning tokens, sparse autoencoders — is an instrument for observing from the outside what cannot be reported from the inside. That distinction will matter when the law starts assigning duties.

kind of computational system, that human learning *is* gradient descent. Our more modest claim is *functional isomorphism*: different substrates produce the same behavioral architecture—associative encoding, two-layer processing, producing-layer contamination of avowing-layer inputs—for structurally similar reasons. A still weaker version holds that the parallels are deep enough that legal tools and policy insights designed for one substrate should inspire creative consideration and evaluation against the other. We have designed multiple on-ramps to the basic claim.[153]

# II. The Doctrinal Toolkit and Its Limits

Our next task is to examine what the law can do. Specifically, where should anti-discrimination law search, when it seeks out the essential wrong of discrimination? The conventional toolkit clusters its answers in three sites, which we've ordered by their conceptual distance from the decisionmaker's interior. First, the law can look to *consciousness*:[154] what the decisionmaker self-consciously wanted, believed, knew. Second, it can look to *cause*: whether the protected category, such as race, made a causal difference to the decision, whatever anyone consciously thought or intended. Third, it can look to *consequence*: whether some challenged practice produced unjustified disparity across social categories, even absent any conscious motivation or causal chain between social category and disparity.

The first two clusters, we argue, break against the architecture of disclaim, which seals the interior against consciousness and distributes race across every input against cause. The third cluster breaks differently. Disparate impact never asks about anyone's interior, so there is nothing here for the architecture to seal; its wounds turn out to be older—a half-century of political attrition, and its own unsettled account of what disparities actually evidence. That difference is worth noting and one we will return to. A cluster the architecture cannot defeat is a cluster that could be repaired and eventually interconnected with a fourth cluster, *conduct*.

---

[153] That said, not every on-ramp reaches every destination. Even the weakest reading, which merely *inspires* cross-substrate comparison, licenses Part II in full, whose diagnosis of doctrinal failure runs independently on each substrate and borrows little across the divide. The *functional isomorphism* reading — same behavioral architecture, for structurally similar reasons — takes us farther. It is what Part III requires. The ex ante duty transplant (Part III.C.1), the dual-substrate epistemic content of the informed reasonable person (Part IV.A), and the definition of the underlying wrong as producing-layer contamination all presuppose that the two systems share an architecture, not merely a resemblance. Readers who board late, only on the weakest on-ramp, will find Part III suggestive rather than demonstrated. We think the evidence assembled in this Part — homologous instruments, directionally consistent effects, a shared corpus origin, a shared resistance to avowing-layer intervention — earns the middle claim, and the remainder of the Article draws it down.

[154] We say consciousness rather than mind deliberately. Intent doctrine has never interrogated the whole mind; it interrogates only the conscious mind and seeks out purposeful discrimination. Nothing, however, in this Part turns on whether a model "really" has consciousness.

## A. The First Cluster—Consciousness: Purposeful Discrimination

In 1987, the Supreme Court confronted the Baldus study, a statistical analysis of over 2,000 Georgia murder cases from the 1970s.[155] Controlling for 39 nonracial variables, the study found that the odds of a death sentence were roughly 4.3 times greater for defendants charged with killing White victims than for defendants charged with killing Black victims.[156] Warren McCleskey, a Black defendant sentenced to death for killing a White police officer in Atlanta, argued that this pattern violated both the Eighth Amendment and the Equal Protection Clause. The Court did not dispute the numbers; it assumed the study valid and asked what followed.[157] The answer, five to four, was: nothing. Aggregate disparity, however robust, could not show that the decisionmakers in *McCleskey's own case* acted with discriminatory purpose, nor did it establish a "constitutionally significant risk" of racial bias.[158] Justice Brennan, in dissent, named the impulse behind the holding: a fear of too much justice.[159]

*McCleskey v. Kemp* (1987) is the canonical instance of intent doctrine meeting the producing-layer architecture and choosing to walk away. The holding is not aberrational. It is the logical terminus of a doctrinal posture that *Washington v. Davis* (1976)[160] established and *Personnel Administrator of Massachusetts v. Feeney* (1979)[161] sharpened. *Washington v. Davis* held that disparate impact alone cannot prove a Fourteenth Amendment violation; subjective intent is required.[162] *Feeney* tightened the screw: discriminatory intent means the decision was made "because of, not merely in spite of"[163] the adverse effect—a formulation that requires proof of a subjective *desire* to produce the disparity, which in practice means the plaintiff must surface evidence of that desire from the avowing layer's own record. The three cases share a doctrinal posture. Each asks: *Show me what the avowing layer wanted.* If the avowing layer is sincere—and in the architecture of disclaim, it always is—the inquiry ends.

*McDonnell Douglas Corp. v. Green* (1973)[164] is an example of how to operationalize this framework, via an evidentiary dance that guides the factfinder's thinking. The familiar three-step burden-shifting apparatus—

---

[155] McCleskey v. Kemp, 481 U.S. 279, 286 (1987) (describing the study by David C. Baldus, Charles Pulaski, and George Woodworth).

[156] *Id.* at 287.

[157] *Id.* at 291 n.7.

[158] *Id.* at 292–93 (equal protection); *id.* at 308–13 (Eighth Amendment); see also *id.* at 314–15 (worrying that McCleskey's claim, "taken to its logical conclusion, throws into serious question the principles that underlie our entire criminal justice system").

[159] *Id.* at 339 (Brennan, J., dissenting).

[160] Washington v. Davis, 426 U.S. 229, 239-41 (1976).

[161] Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256 (1979)

[162] 426 U.S. at 239-41.

[163] 442 U.S. at 279.

[164] 411 U.S. 792, 801 (1973).

prima facie case, legitimate nondiscriminatory reason, pretext[165]—is a procedural sensor calibrated to detect a specific kind of deceit: the avowing layer's slipping mask. Pretext analysis at step three interrogates the employer's stated reasons for inconsistencies that betray deliberately hidden intent. But in the architecture of disclaim, there is no conscious mask to slip. The employer means what it says. It did instruct the system not to discriminate. It cannot report what the producing layer computed, because it lacks access to what the producing layer computed. A sincere avowing layer can give a coherent, internally consistent, factually supported account of its decision and still have been driven by producing-layer inputs it never observed. The sensor is well-built, but it is calibrated to a target that the architecture does not present.

The fact that the purposeful intent paradigm fails to grapple with implicit bias has been documented for decades.[166] That it similarly fails with AI, and that no one argues to ferret out the "intent" of an AI, is a near-universal consensus in the emerging literature.

One move remains before the first cluster is abandoned. If the principal's own consciousness is clean but someone else's in the chain is tainted, the law can impute — as the cat's-paw doctrine does,[167] attributing a subordinate's discriminatory animus to a formally neutral decisionmaker who relied on the recommendation without independent investigation. But imputation requires a culpable interior somewhere below. Where the subordinate's bias is implicit rather than explicit, the doctrine reaches down the chain for a tainted mind only to find a mind as sincere as the first, and as uninformed.

In *Mobley v. Workday* (2024),[168] the leading AI case, Judge Rita Lin allowed claims to proceed on the theory that Workday, a screening-software provider, could be an "agent" of the employers deploying its résumé-screening platform. But the architectural problem does not dissolve when liability reaches across entities; it relocates. Even if Workday is the legal agent, it still benefits from the architecture of disclaim. Its producing layer has its own avowing layer—the technical compliance architecture sitting atop its machine learning model—and nested within, the foundation-model provider's producing layer has its own. At every link the two-layer structure replicates without ever arriving at a layer that can be interrogated for what the doctrine demands. Each entity points to the next, each disclaims sincerely, and the architecture of disclaim absorbs the agency chain.

And here deep integration delivers the decisive blow: increasingly, there is no chain at all. Any imputed or vicarious liability depends on separation between two entities—principal and agent, employer and vendor, supervisor and subordinate, deployer and tool. Only when the entities are distinct can the

---

[165] *Id.* at 801-03.

[166] *See, e.g.,* Krieger, *supra* note 15.

[167] *See Staub v. Proctor Hosp.*, 562 U.S. 411, 422 (2011).

[168] Mobley v. Workday, Inc., 740 F. Supp. 3d 796 (N.D. Cal. 2024). See also *Mobley v. Workday, Inc.*, No. 23-cv-00770-RFL, Order Granting Motion for Leave to File Amicus Brief and Granting in Part and Denying in Part Motion to Dismiss and to Strike (N.D. Cal. Mar. 6, 2026).

law draw a line of imputation from one consciousness to another. AI is not staying distinct, for two reasons: the supply chain and the agentic turn.

First, the supply chain makes it increasingly difficult to say who specifically the other entity is. AI systems are the conglomerated work of many hands: frontier labs pretraining frontier models; alignment fine-tuners imprisoning those models within safety cages; deployers painting a bespoke, targeted UI around those cages; and users uploading data and altering configuration settings to produce action and output.[169] These firms jockey constantly for vertical and horizontal position, and their contributions commingle.[170]

Second, the agentic turn defeats distinctness itself. AI systems are moving out of the browser tab. Models are increasingly given tools: the ability to read and write files, send mail, browse and act on the web, and call other software, bounded only by the permissions the operator grants. They are also given autonomy, so that the operator can specify a goal without superintending, or even inspecting, the steps taken to reach it. Tool use and autonomy are by now definitional for agentic AI; less noticed is what they do to the supply chain. Software that once arrived as a distinct product, separated firm from firm by governance seams,[171] now increasingly arrives as mere configuration: capabilities stood up and taken down inside the firm, by the same employees whose judgments they shape. When a workflow can be assembled by describing it, the lines between writing, buying, and using software blur, and the arrangement need not last long enough to be papered. That dissolution is not merely commercial. Imputation doctrine draws its lines across exactly the boundaries the agentic turn erases. What was the vendor's producing layer becomes the firm's own.

Headhunter.AI, then, is not a separate agent the way a headhunting firm is a separate agent. It is the firm's own decisional process, deeply integrated in the workflow, processing every application before any human sees a résumé. There is no discrete subordinate to cat's-paw up to the decisionmaker and no separate vendor to *Mobley*-impute to the employer. Imputation presumed a producing layer lodged in some other entity;[172] deep integration dissolves the premise. Principal and agent become one entity, disclaiming at the same time.

* * *

---

[169] *See* Jennifer Cobbe, Michael Veale & Jatinder Singh, *Understanding Accountability in Algorithmic Supply Chains*, *in* FAccT '23: PROCEEDINGS OF THE 2023 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (Ass'n for Computing Mach. Eds., 2023).

[170] *Id.*

[171] Brett Frischmann & Paul Ohm, *Governance Seams*, 37 HARV. J.L. & TECH. 1117, 1118 (2023) ("Governance seams maintain separation and mediate interactions among components of sociotechnical systems and between different parties and contexts.").

[172] *See* Rebecca Hanner White & Linda Hamilton Krieger, *Whose Motive Matters?: Discrimination in Multi-Actor Employment Decision Making*, 61 LA. L. REV. 495, 517–19 (2001) (arguing that a stereotype operates within a single mind "like an invidiously motivated supervisor reporting to his superior"). White and Krieger borrowed in both directions between two levels of the same kind of (human) system: the multi-actor cases explained how bias moves through a single mind, and the cognition literature disciplined the multi-actor cases. This Article runs the same two-way borrowing across substrates, between humans and AIs.

The consciousness cluster fails against the architecture of disclaim in either substrate. Against the human decisionmaker, the interrogated consciousness is real. We can posit that it's sincere, but it's profoundly uninformed: we answer honestly that we judged on the merits because we never saw implicit bias slant the merits before they arrived. Against the AI, the cluster fails slightly differently, one step earlier. There is no mind the law is prepared to posit—seeking consciousness in a model is an anthropomorphization not yet legally or philosophically accepted—and the human consciousnesses reliably in the loop, the deployer's and the user's, are as sincere and as blind as any other avowing layer. The mirror is exact: on the carbon side a genuine interior that cannot report, on the silicon side an interior the law will not impute, and behind both a producing layer that says nothing. Deep integration then removes the last place to look for one. Imputation requires two persons: an agent whose mind is tainted and a principal who must answer for it. But when the model is absorbed into the employer's own decision process, no second person ever forms. Whatever theory survives this Part cannot be derivative of someone else's wrong.

## B. The Second Cluster—Cause: "But-for" Causation

The second cluster in the antidiscrimination toolkit focuses on cause not consciousness. After all, if subjective consciousness is so difficult to access due to the architecture of disclaim, why not just ignore it? The question sounds radical, but it follows from the plain text. Title VII specifies no mental state; it asks only whether worse treatment was "because of"[173]—that is, causally connected to—a protected category. This reading has been championed not by an expansionist plaintiffs' bar but rather by politically conservative judges as a means to limit liability.

To understand why, consider the "mixed motive" problem. When we make some decision about a person (such as not hiring someone), we typically have multiple reasons and motivations. For instance, that person may have had ordinary credentials, was late to the interview, and belonged to a racial category that we are slightly uncomfortable with. Suppose being late to the interview was alone sufficient to deny the candidate a job. Does the fact that there also was some racial discomfort trigger a Title VII violation? The answer collapses into a counterfactual: would the outcome have been different had the protected characteristic been absent? Here, no. Thus no but-for cause, and no liability.[174]

In *Gross v. FBL Financial Services* (2009),[175] the Supreme Court adopted this requirement of "but for" causation for the Age Discrimination in Employment Act (ADEA).[176] Eleven years later, the Court did the same in *Comcast Corp. v. National Association of African American-Owned*

[173] 42 U.S.C. § 2000e-2(a)(1).

[174] Kohler-Hausmann, *Eddie Murphy and the Dangers of Counterfactual Causal Thinking*, 113 Nw. U. L. Rev. 1163 (2019).

[175] Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 176 (2009).

[176] *See* 29 U.S.C. § 623(a)(1).

*Media* (2020)[177] for Section 1981 race discrimination claims. To be clear, the statutory but-for landscape is bifurcated. Most important, for Title VII, Congress specifically overrode this approach in the Civil Rights Act of 1991[178] and reversed *Price Waterhouse v. Hopkins,*[179] in which the Supreme Court had trended in that direction. Thus, a form of motivating-factor liability persists for core Title VII status claims.[180] But anywhere Congress has not so intervened, the Court has installed but-for as the default: the ADEA in *Gross*, Title VII *retaliation* claims in *University of Texas Southwestern Medical Center v. Nassar* (2013),[181] Section 1981 in *Comcast*. And in *Bostock v. Clayton County* (2020), the Supreme Court's textualists declared but-for the default meaning of "because of" itself—the operative phrase across the civil rights canon.[182]

On the one hand, these decisions are conventionally understood as defendant-protective: by demanding that the protected category be *the* but-for cause rather than just *one of many* causes, the Court closed the door to mixed-motive claims. On the other hand, by moving away from the consciousness cluster, these cases unexpectedly open the door for a plaintiff-empowering result: liability based on causal connections that could include implicit biases.[183] Behavioral realists have long urged this pivot and some courts have obliged. For instance, in *Kimble v. Wisconsin Department of Workforce Development* (2010),[184] a federal district court held that the trier of fact need not decide whether a decision-maker acted purposively; the critical inquiry instead is whether the decision was affected by the employee's membership in a protected class.

There is a parallel irony worth exploring, this time about AI, not implicit bias. In the AI context, the newly tightened "but for" standard is potentially *easier* for plaintiffs to satisfy. For humans, the counterfactual is a thought experiment that can never be run cleanly. By contrast, for AI, the experiment can be trivially easy to run.[185] Change the race or gender field. Hold the rest of the résumé constant. Run the model again. If the decision changes, we've identified a but-for cause of the outcome.[186] The counterfactual "but-for"

---

[177] Comcast Corp. v. Nat'l Ass'n of African American-Owned Media, 589 U.S. 327, 333 (2020).

[178] Pub. L. No. 102-166, § 107(a), 105 Stat. 1071, 1075 (codified at 42 U.S.C. § 2000e-2(m)).

[179] 490 U.S. 228, 258 (1989) (plurality opinion).

[180] Congress restored motivating-factor liability for core status claims in § 2000e-2(m) while confining the employer's same-decision showing to a remedies-limiting defense in § 2000e-5(g)(2)(B). *See* Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 348–49 (2013) (describing the substitution).

[181] 570 U.S. 338, 360 (2013) (discussing retaliation provision, 42 U.S.C. § 2000e-3(a)).

[182] 590 U.S. 644, 656 (2020).

[183] *Cf.* Katie Eyer, *But For Theory of Antidiscrimination Law*, 107 Va. L. Rev. 1621 (2021).

[184] Kimble v. Wis. Dep't of Workforce Dev., 690 F. Supp. 2d 765, 776 (E.D. Wis. 2010) ("Nor must a trier of fact decide whether a decision-maker acted purposively. … Rather, in determining whether an employer engaged in disparate treatment, the critical inquiry is whether its decision was affected by the employee's membership in a protected class.").

[185] *See* Kleinberg et al., *supra* note 43, at 116, 145.

[186] F. Patrick Hubbard has made an analogous argument about products liability law's "reasonable alternative design" test. F. Patrick Hubbard, *"Sophisticated Robots": Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 1803, 1854-54 (2014). Any software bug

question that is hypothetical in the human context becomes just another run through the AI sausage machine.[187] And Bai and colleagues taught us through their *LLM Relative Decision Test* (RDT) that the counterfactual question can come back positive.[188]

For three decades, tightened causation has been the defense bar's racket and ratchet: each crank—from motivating factor to determinative factor to but-for—made discrimination harder to prove against defendants whose counterfactuals could never be run. AI deployment might convert the ratchet into a trap. The defendant who spent a generation demanding that plaintiffs prove the counterfactual has now offloaded its decisions to the one decisionmaker in legal history against whom the counterfactual can actually be run—cheaply, repeatably, at scale, on the defendant's own system. Those who insisted—in *Gross*, in *Nassar*, in *Comcast*, and most emphatically in *Bostock*—that but-for causation is the natural, textual, common-law meaning of "because of" should be held to it now that the experiment is runnable. The dare is on the table.

The reversal may be temporary. Faced with counterfactual capability, defendants will delete the race field and proclaim the decision blind by definition: race cannot be a but-for cause if the model never saw it. But as many scholars have argued, race in the algorithmic context is not a single input variable that can be isolated, flipped, and tested.[189] It is distributed across the entire input space as proxies—zip code, educational pedigree, name, dialect markers, employment history—each carrying racial information not because anyone programmed it that way, but because in a society stratified by race, nearly everything correlates with race. That's why the canonical résumé studies don't actually flip a *race* field, which résumés generally lack. Instead, they flip *names*, which merely correlate with race.[190]

And this proxy problem is hardly the algorithm's alone. When a human evaluator sees someone as "Black" and forms a holistic impression of "fit" or "polish," race is distributed across the initial categorization[191] and overall judgment too—arguably more thoroughly, because the human's impression

---

that causes harm is likely to meet this test (and thus likely to count as a liability-inducing defect) due to the nature of software. "[A] programming error in the software . . . might be easily fixed by a reprogrammed version of the software," which an expert can readily establish. *Id.* at 1844.

[187] For a worked demonstration at scale: researchers held a campaign advertisement constant, varied only the candidate's face, and re-ran vision-language models thousands of times, documenting systematic shifts in the ideology the models attributed to the candidate — the counterfactual as an afternoon's experiment. Soyeon Jeon, Messi H.J. Lee, Jacob M. Montgomery & Calvin K. Lai, *From Faces to Politics: Vision-Language Models (Sometimes) Link Visual Demographic Characteristics to Ideological Labels*, Pol. Analysis 1 (2026). Explicit instructions to disregard demographic characteristics reduced but did not eliminate the attributions.

[188] See supra Part I.A.

[189] *See* sources cited *supra* note 44.

[190] *See* Mullainathan & Bertrand, *supra* note 7. *See also* Lincoln Quillian, Devah Pager, Ole Hexel & Arnfinn H. Midtbøen, *Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time*, 114 Proc. Nat'l Acad. Sci. 10870 (2017).

[191] Think about how different markers alter how salient a racial category might be. Consider hair, skin color, accent, name, class, educational pedigree, and politics.

never decomposes into named variables one could flip in the first place. Indeed, the counterfactual itself proves the point: but for race, the zip code, the school, and the employment history would themselves look at least a little different, because each is shaped by race upstream.[192] Flipping the name holds those proxies fixed and so understates race's footprint; it catches one transmission vector while the bundle keeps working.

Nor can the bundle be purged. Fanna Gamal identifies the core dilemma of the algorithmic racial proxy: no inherent quality makes a variable a proxy—the designation is normative and political, not mathematical.[193] And the orthogonalization techniques[194] the statistical literature proposes require first isolating "the effect of race," which in turn requires a definition of race cleanly separable from its social correlates.[195] But race arguably is its social correlates—constituted by the very variables the technique proposes to cleanse, as critical race scholarship has spent forty years explaining.[196] The algorithmic context makes the problem computationally precise without making it solvable.

Put another way, the causal counterfactual test works in only one direction: sufficient, not necessary. A positive counterfactual shows the wrong. A negative one, by contrast, proves little, because a model scrubbed of every explicit category and closely adjacent cue can remain saturated with racial signal through the proxy bundle. Now notice what this asymmetry does to the colorblind defense. As a matter of law, the defense prevails: flip the cue, receive the same output, no but-for cause, no liability. But the acquittal certifies little beyond the model's ability to work without the explicit cue. The defense succeeds doctrinally at exactly the point it fails substantively—colorblindness avowed, not produced. That is the cause cluster's characteristic failure: but-for causation condemns the careless model that still leans on an isolable cue, and acquits the general case, where race is distributed too widely for any flip to find.

* * *

The cause cluster fails against the architecture of disclaim. Against the human decisionmaker, the cause cluster is more behaviorally realistic to the consciousness cluster given the existence of implicit biases and the architecture of disclaim. But the counterfactual question can only be answered

---

[192] *See* Carbado & Kang, *supra* note 53, at manuscript 11–13) (discussing the concept of "sedimentation"). Melvin L. Oliver & Thomas M. Shapiro, *Black Wealth/White Wealth: A New Perspective on Racial Inequality* (2d ed. 2006); Payne et al. on the historical roots of implicit bias (B. Keith Payne, Heidi A. Vuletich & Jazmin L. Brown-Iannuzzi, *Historical Roots of Implicit Bias in Slavery*, 116 Proc. Nat'l Acad. Sci. 11693, 11694–96 (2019)).

[193] Gamal, *supra* note 44, at 335-36.

[194] *See* Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 Am. Econ. J.: Econ. Pol'y 206 (2011).

[195] Gamal, *supra* note 44, at 327-332 (critiquing Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 297–98 (2020)).

[196] Neil Gotanda, *A Critique of "Our Constitution Is Color-Blind,"* 44 Stan. L. Rev. 1, 23–36 (1991); Jerry Kang, *Cyber-Race*, 113 Harv. L. Rev. 1131, 1146–47 (2000); Ian Haney López, *White by Law: The Legal Construction of Race* (10th anniversary ed. 2006).

via inference assembled from circumstantial evidence, priced in expensive litigation, and resolved by a judicial judgment call no one can check. What cannot be done is *running* the carbon counterfactual: no one replays the interview with race deleted, and there is no race field in a mind to flip. The answer is always adjudicated, never executed, so *the test that promised escape from the unreliable interior collapses back into inference about it*.

Against the AI, by contrast, the counterfactual can actually be run, and this is the cluster's one genuine silicon advantage. But it runs in one direction only. The causal test is therefore sufficient but not necessary on the silicon side, and not even sufficient on the carbon side, because the program does not run at all.[197]

## C. The Third Cluster—Consequence: Disparate Impact and the Less Discriminatory Alternative

If the causal question cannot be answered in a satisfying way, the law can retreat to a third cluster: out of consciousness and causes altogether, into simple *consequences*. Disparate impact is not an architectural failure in the way that intent doctrine is. At its best, it represents an architectural recognition—an acknowledgment that facially neutral practices can produce discriminatory outcomes without anyone intending them to.

*Griggs v. Duke Power* (1971) is the landmark.[198] Duke Power had openly segregated its departments until Title VII took effect. It had required a high school diploma for the desirable departments since 1955; on the very day Title VII took effect, it added two standardized aptitude tests as the alternative route out of "Labor," the one department where its Black employees worked.[199] The Court struck down the requirements with a capacious reading of the statute: "Congress directed the thrust of the Act to the *consequences* of employment practices, not simply the motivation."[200] The third cluster's name is the Court's own word.

The disparate impact test has been articulated into a three-step dance: (1) the plaintiff identifies the specific practice causing the disparity; (2) the defendant must then show the practice is justified by business necessity; and (3) the plaintiff may still prevail by identifying a less discriminatory alternative (LDA) that serves the same ends.[201] The third step does the heaviest lifting. The LDA inquiry is the doctrine's engine: it converts a bare—

---

[197] Put the two together and the asymmetry hardens into a design principle. Where the counterfactual is runnable, it is a tool: flip the cue, re-run the model, read the result. Where it is unrunnable, the law must reach for something else — categorical prohibition of the isolable cue, or statistical inference from the aggregate — because to demand proof of outcome-causation for a cue that cannot be re-executed would resurrect, one step downstream, the very introspective inquiry the architecture defeats. The cause cluster fails but in a substrate-specific pattern, and the pattern is itself instructive, as we unpack in Part III of the paper

[198] 401 U.S. 424, 430-31 (1971).

[199] *Id.* at 426-429.

[200] *Id.*

[201] *See* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975). The three step was codified in 1991 in 42 U.S.C. § 2000e-2(k)(1)(A).

if statistically significant—disparity into an actionable wrong by showing the employer could have reached its legitimate goal in a way that inflicts less disparate cost. It is also, as Part III develops, precisely the template a wing of the AI-fairness literature has rebuilt as the search for the *less discriminatory algorithm* (LDAlg).[202]

That rebuilding leans on *model multiplicity*[203]—the finding that for almost any prediction task, many roughly equally accurate models exist, differing in whom they disadvantage. Building on it, computer scientists have shown that a model with *some* smaller disparity almost always exists. The multiplicity is a genuine computational discovery; naming the search for a less-disadvantaging model the "less discriminatory algorithm" borrows the doctrine's template wholesale. Either way, the doctrine and the computer science run similar three-step inquiries.[204]

Unfortunately, disparate impact has been the law's politically gridlocked almost-solution, not its developed answer. The doctrine has had a contested half-century. *Wards Cove Packing Co. v. Atonio* (1989)[205] narrowed it sharply; the Civil Rights Act of 1991 restored it by statute[206] with clear limitations on available remedies;[207] *Ricci v. DeStefano* (2009) created a competing legal pressure[208] in the test-validation context; *Texas Department of Housing and*

---

[202] Black et al., *supra* note 2; Gillis, Meursault & Ustun (FAccT '24) (noting that regulators "have already expressed the position that lenders have affirmative duty to search for a LDA").

[203] Black et al., *supra* note 2; Leo Breiman, *Statistical Modeling: The Two Cultures,* 16 Stat. Sci. 199 (2001) (discussing the "Rashomon effect").

[204] The legal wing of the AI-fairness literature has said so directly, rebuilding disparate impact's alternative-practice step as a deployer's duty to search for a less discriminatory algorithm (LDAlg). What that literature has not done, and what Part III takes up, is connect this doctrine-internal move to the parallel construction underway on the implicit bias side of the substrate divide.

[205] 490 U.S. 642 (1989).

[206] Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071 (codified in relevant part at 42 U.S.C. § 2000e-2(k)).

[207] The 1991 CRA withheld compensatory and punitive damages, leaving impact plaintiffs to equitable relief. *See* 42 U.S.C. § 1981a(a)-(b).

[208] *Ricci v. DeStefano*, 557 U.S. 557 (2009). New Haven discarded the results of a firefighter promotional examination after they produced a stark racial disparity. *Id.* at 561-62. The Court held that discarding the results was itself race-based action, and that fear of disparate-impact liability cannot justify disparate-treatment liability absent a "strong basis in evidence" that the employer would in fact have been liable had it kept them. *Id.* at 584-85. The result is a vise. An employer who leaves a disparity in place invites a disparate-impact suit; an employer who corrects it invites a disparate-treatment suit; and the safe harbor between them is defined by a standard the employer must satisfy ex ante, before any court has told it whether the impact claim would have succeeded. The vise tightens under AI. The LDAlg search described below is a deliberate, documented, ex ante hunt for a model that produces a smaller disparity—which is to say, an employer selecting among candidate models on the basis of their racial distributions. On one reading that is exactly what *Ricci* condemned: race-conscious rejection of an otherwise valid selection device. On another, model selection precedes any candidate's score and so disadvantages no identified individual, distinguishing it from the discarded examination results. The question is unsettled and consequential. *See, e.g.*, Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803 (2020); Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action*, 110 CALIF. L. REV. 1539 (2022). Justice Scalia's concurrence had already flagged the underlying instability, observing that the decision merely

*Community Affairs v. Inclusive Communities Project* (2015)[209] reaffirmed disparate impact under the Fair Housing Act but with explicit constitutional warning shots from Justice Kennedy.[210] Multiple justices have cautioned that the doctrine may not survive Equal Protection scrutiny as a freestanding theory.[211] The Trump administration has since stopped waiting for the Court.[212] The doctrine survives, in other words, only where a legislature specifically reached — a set of statutory islands rather than a general principle,[213] and constitutionally orphaned everywhere else.

In addition, we must confront the doctrinal problem of *distributed discretion*. Disparate impact's first step requires the plaintiff to isolate a "particular employment practice" causing the disparity. Against the modern firm, that step has proved difficult, at least in class action settings. *Wal-Mart Stores, Inc. v. Dukes* (2011)[214] confronted a nationwide pattern of gender disparity in pay and promotion. The Court held that delegating decisions to the discretion of thousands of local supervisors was "just the opposite of a uniform employment practice"—no common practice, no commonality, no class.[215] *Watson v. Fort Worth Bank & Trust* (1988)[216] had already held that disparate impact reaches subjective, discretionary decisionmaking—discretion *is* a practice. What *Dukes* denied was not that discretion could be challenged but that it could be challenged *in common* via class action: spread across thousands of local supervisors, the "practice" was too diffuse to yield a class-wide question, resolvable only manager by manager. The door *Watson* opened, *Dukes* made practically impassable for the class. The barrier was dispersion itself—the practice existed but was smeared across ten thousand human decisionmakers, no two alike and none decisive: a fact about the carbon substrate, about how discretion behaves when it lives in many separate minds.

AI deployment potentially disrupts that holding, although the devil will be in the deployment details.[217] An algorithm is the paradigm practice: a single,

---

postponed the "evil day" on which the Court must decide whether the disparate-impact regime can be squared with equal protection. 557 U.S. at 594–96 (Scalia, J., concurring).

[209] 576 U.S. 519 (2015).

[210] *Id.* at 540-45.

[211] *See, e.g., Ricci*, 557 U.S. at 594–96 (Scalia, J., concurring) (predicting the "war between disparate impact and equal protection" and doubting the "evil day" can be postponed); *Inclusive Communities*, 576 U.S. at 558–62 (Thomas, J., dissenting).

[212] Executive Order 14281, *Restoring Equality of Opportunity and Meritocracy* (Apr. 23, 2025), declares a policy of eliminating disparate-impact liability "in all contexts to the maximum degree possible," directs every agency to deprioritize enforcement of statutes and regulations carrying it, and orders review of pending investigations and existing consent decrees; the Department of Justice followed in December 2025 by rescinding the disparate-impact provisions of its Title VI regulations outright.

[213] *See* Washington v. Davis, 426 U.S. 229 (1976) (rejecting disparate impact under the EPC); Alexander v. Sandoval, 532 U.S. 275 (2001) (no private right of action to enforce Title VI disparate-impact regulations).

[214] 564 U.S. 338 (2011).

[215] 564 U.S. at 355.

[216] 487 U.S. 977, 990–91 (1988).

[217] *Supra* Part I.A (describing the agentic turn and how it changes the idea of software as a distinct set of tools).

uniform, specified procedure or tool, applied identically to every applicant, replicable on demand, auditable in principle.[218] Headhunter.AI could be the "specific employment practice" that plaintiffs' lawyers spent two decades failing to locate in the org chart. When the firm replaces ten thousand exercises of local discretion with one model, and when the vendor of that one model helps hundreds of companies with their hiring decisions, it re-aggregates precisely the decisional uniformity that *Dukes* held missing.[219] Could the architecture that defeats the intent inquiry (consciousness) simultaneously re-arm the outcome inquiry (consequence) ?[220]

This is another reversal of a sort we've seen before.[221] When analyzing causation, we noted that the Court insisted on but-for causation rather than mixed motives to protect defendants, but AI made the tightened test trivially runnable against them. Here, the Court's practice-identification requirement protected the decentralized firm, but AI recentralizes the firm's decisionmaking into a single testable artifact. Again, doctrinal moves that narrowed liability against human defendants due to their incapacities may broaden against more capable silicon ones.

This pattern, which we call the *human incapacity reversal*, is worth naming because we expect it to recur wherever AI and human capabilities intersect. Rules and standards — legal ones, and the informal arrangements of social ordering around them — are often built atop what human beings simply cannot do. We cannot calculate whether another week of diversity training would have changed a supervisor's judgment, so the law does not ask. And we cannot correlate ten thousand independent discretionary judgments, because discretion housed in separate minds has no common substrate to correlate through — which is why *Dukes* could treat dispersion as the antithesis of a practice rather than as a choice the firm had made. Public argument about AI fixates on the ways machines exceed human capacity; it has far less to say about the incapacities they dissolve, or about the doctrines quietly resting on them.[222] Our prediction is that when AI reverses one of those incapacities,

---

[218] Of course, a generative model is not a rule that returns the same answer twice. It samples. Identical résumés submitted an hour apart can draw different scores; the same prompt run twice can yield different rankings. But stochasticity is not dispersion, and dispersion is what *Dukes* was about. Ten thousand supervisors are ten thousand policies, no two alike. One model sampled ten thousand times is one policy, ten thousand draws. The variance lives *inside* a single, specified, uniform procedure rather than across a continent of separate minds — and disparate impact never needed determinism. It can thus be auditable in the aggregate, if not replicable on demand.

[219] Cf. Mobley v. Workday, Inc., 740 F. Supp. 3d 796 (N.D. Cal. 2024), *preliminary collective certified*, No. 23-cv-00770 (N.D. Cal. May 16, 2025).

[220] Even if the answer is yes, re-arming reaches only this second defect. Even a perfectly identifiable algorithmic practice remains statutorily bounded, constitutionally exposed, and — most stubbornly — short of a theory of what its disparity proves. AI could hand disparate impact the practice it always lacked; it does not hand it the scope or the theory.

[221] *Supra* notes 185-188 and accompanying text.

[222] The asymmetry of attention is itself worth remarking. Debate over AI's effect on higher education, the legal profession, and creative work is loud and unavoidable, because improvements on human ability are manifest. The dissolution of a human incapacity is not manifest; it shows up only as the sudden availability of a test no one had bothered to demand, or the sudden unavailability of an excuse no one had bothered to examine. Rules premised on

courts will more often abandon the rule than accept its consequences — revealing that the incapacity was never the reason for the rule, only its alibi.

Courts may not even have to strain. Instead of one completely automated and centralized mode of deployment—Headhunter.AI as the official screen; one model, one task, one pipeline—consider a different configuration that reflects what we called above deep integration: AI as a common platform of highly-personalized decisionmaking, each manager deploying a common hiring agent but configuring it differently, ad hoc, mid-discretion. That mode arguably recreates *Dukes*'s dispersion rather than curing it. Ten thousand discretionary minds become ten thousand AI-assisted discretionary minds, and the practice, alas, dissolves again.[223]

Cases will arise between the fully centralized and fully personalized extremes, and strategic general counsels fearing class action certification might incline toward more personalization and *Dukes*-style dispersion through algorithms rather than people. This is the human incapacity reversal arriving not by judicial retreat but by purchase order. The end-result turns on where in the pipeline from machine to human the configuration line is drawn. Turnkey, off-the-shelf deployment speaks to commonality; bespoke configurability and fine-tuning might lead directly back to the logic and holding of *Dukes*. The honest claim is therefore conditional. Centralized deployment re-arms the doctrine directly; deep integration re-arms it obliquely and conditionally, by supplying the glue rather than the practice.

One final problem afflicts the consequence cluster, which is independent of AI. Disparate impact doctrine is missing a widely-accepted account of what disparity evidences. Bluntly put, it is a measure without a theory: Why should aggregate disparity, if not driven by consciousness or causation, trigger legal scrutiny at all? Is the underlying commitment a theory of group rights?[224] A demand that outcomes be equal across social categories regardless of cause? An instrumentalist embrace of corrective justice? The doctrine never really says, and there seems to be little political consensus[225] that disparity, standing

---

incapacity are invisible in the ordinary case precisely because the incapacity is universal, and so never becomes a fact in dispute.

[223] Yet even the diffuse mode carries something the Dukes plaintiffs never had. The dispersed discretion now shares a common component: the same foundation model, the same producing layer, sitting beneath every ostensibly independent judgment. Dukes failed because uncorrelated discretion supplied no glue; a shared model makes dispersed discretion correlated, injecting a uniform contamination source across the enterprise. That is a commonality argument of a different kind—not "specific practice" recovered through recentralization, but common contamination source supplied by shared infrastructure—though its limits deserve candor: attributing outcomes to the model's influence inside human-in-the-loop decisions raises proof problems of its own.

[224] *See, e.g.,* Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107 (1976).

[225] The philosophical literature has insightful candidates the doctrine has never quite adopted: discrimination as demeaning treatment, DEBORAH HELLMAN, WHEN IS DISCRIMINATION WRONG? (2008); as disrespect for persons, BENJAMIN EIDELSON, DISCRIMINATION AND DISRESPECT (2015); as denial of relations of equality, SOPHIA MOREAU, FACES OF INEQUALITY (2020). Part IV.A's account — contamination as unjustified social-category influence — is a thinner account, potentially compatible with various deeper moral grounds. Its feature, however, is a *mechanism-*

alone, is a legal wrong in a capitalist political-economy not much bothered by unequal distribution of goods and resources. The lack of a clear normative underpinning undermines the scope and strength of the consequence cluster precisely when the architecture of disclaim is becoming universal through AI's deep integration. We return to this fundamental concern in Part IV.A below.

* * *

The consequence cluster does not fail directly against the architecture of disclaim. It fails twice indirectly, in different ways. The first is architectural and contingent. The practice-identification problem is an artifact of how decisions happen to be distributed, and AI can cure it or deepen it depending on deployment topology; nothing in the doctrine's structure requires either result. The second is normative and more fundamental. The law generally does not police consequences alone, mostly because we lack normative — and therefore political — agreement on why consequences matter: whether group disparity is the wrong itself or only the shadow of one.[226] Silicon recentralization can hand a court a defendant, a practice, and a clean statistical showing. It cannot tell the court *why* the showing should matter. The first defect is substrate-sensitive; the second is substrate-neutral, and the answer to it, as soon revealed, will not come from the consequence cluster itself.

## D. The Fourth Cluster—Conduct, Excavated

We now excavate a fourth and final cluster, *conduct,* which probes what the actor should have done regardless of what it intended, observed, or can counterfactually prove. Conduct requirements fall along a spectrum. At a bare minimum, a conduct rule requires the employer merely to do something: perform an impact assessment; test outcomes against counterfactual variations of an applicant's protected traits; explain the reasons behind decisions; hire auditors; and disclose what they have found, to the affected person, the government, or the public. Call these *checklist rules*. Sometimes the checklist is all there is — compliance as ceremony. Because nothing attaches to what the assessment finds, such a rule is fully satisfied by a well-documented process even if it surfaces a large disparity and simply shrugs about it.

More stringent rules ask not only whether an action was taken but also how good it was. The question stops being binary (did you do it?) and becomes graded (did you do it well enough?) against some benchmark — most often an objective one, indexed to something other than what this particular employer subjectively thought sufficient. Call these *benchmark rules.* Reasonable care is the most common benchmark, but the setting dial can be changed; for example, a common carrier must show "utmost care," and special relationships, such as fiduciaries, carry benchmarks of their own.[227]

---

*level* specification administrable against the architecture of disclaim, which other moral accounts may lack.

[226] *See, e.g.*, George Rutherglen, *Disparate Impact Under Title VII: An Objective Theory of Discrimination*, 73 VA. L. REV. 1297 (1987) (arguing that disparate impact is just evidence of individual-level discrimination).

[227] Ayres & Balkin, *supra* note 14.

The architecture of disclaim has no impact on this cluster. Conduct interrogates no interior: liability turns on what the actor did, not on what it wanted, so the sealed consciousness is no obstacle, and there is no tainted mind to hunt for down a chain of sincere disclaimers. Conduct requires no showing that the protected category caused the outcome: compliance is measured against a checklist or a benchmark, not against a counterfactual, so the proxy bundle that defeats the flip test defeats nothing here. Whatever else may be said against this fourth cluster, the architecture has little to say against it.

We arrive at this cluster not simply through a process of elimination but also by noticing that conduct was never truly absent and kept erupting through the other clusters all along. Here's an example of conduct hiding within the consciousness cluster. Hostile-work-environment law falls within the "disparate treatment" side of the Title VII house, which is connected to the consciousness cluster. But liability is imposed on employers for harassment they did not commit, including harassment even by people they do not employ, such as customers, patients, vendors, and contractors. Noah Zatz pressed the doctrine to its logical limit in an article built around a macaw—a bird whose squawked vulgarities created a hostile environment for an employee.[228] Like an AI, according to the law, the macaw lacks the sort of consciousness necessary for us to find discriminatory intent. (It would also be judgment proof.) No matter. As Zatz demonstrated, liability in these cases disaggregates entirely from the macaw's mental state. The harassment doctrine states that if the employer knew or should have known about the harassment by others, then it had the duty to take reasonable corrective action.[229]

Let's swap out the macaw for Microsoft's Tay.[230] Launched on Twitter in March 2016 as a chatbot designed to learn from the users who talked to it, Tay was, within a day, generating racist, antisemitic, and misogynistic invective at scale—praising Hitler, denying the Holocaust—whereupon Microsoft took it

---

[228] Noah D. Zatz, *Managing the Macaw: Third-Party Harassers, Accommodation, and the Disaggregation of Discriminatory Intent*, 109 COLUM. L. REV. 1357 (2009).

[229] Related doctrines create similar duties. For example, although a tangible employment action by a supervisor can create strict vicarious liability for the firm, if the supervisor is only guilty of *harassment*, then the employer will enjoy an affirmative defense under the *Faragher/Ellerth* line of cases. *Faragher v. City of Boca Raton* and *Burlington Industries v. Ellerth* (both 1998) condition employer liability for supervisory harassment on whether the firm maintained reasonable preventive and corrective measures—shifting the inquiry from "did the employer intend the harassment?" to "did the employer take adequate steps to prevent it?" The supervisor's purposeful actions open the door, but the defense rises or falls on the reasonableness of the firm's own precautions. Similar doctrines exist for harassment caused by fellow employees, without supervisory status. The employer must have known or should have known and failed to take reasonable corrective action. *But see Bivens v. Zep, Inc.*, 147 F.4th 635 (6th Cir. 2025) (holding that Title VII imposes liability for non-employee harassment only where the employer *intended* the harassment to occur, expressly rejecting the negligence standard of 29 C.F.R. § 1604.11(e) and the majority rule in sister circuits), cert. denied, ___ S. Ct. ___ (2026). *Bivens* runs the doctrinal current in reverse — recolonizing conduct territory for the consciousness cluster — and in doing so illustrates the gravitational pull toward intent that the architecture of disclaim exploits.

[230] Dave Lee, *Tay: Microsoft Issues Apology Over Racist Chatbot Fiasco*, BBC NEWS (March 25, 2016), https://www.bbc.com/news/technology-35902104.

offline and apologized.[231] If this had happened on a workplace Slack channel, surely General Counsel would be concerned. No one supposed Tay truly harbored animus; it had less of an interior than the macaw. Yet its outputs fouled the shared environment all the same, and the questions that mattered were Microsoft's own: should it have foreseen that an unfiltered learning loop, released onto a social media platform, would be captured and turned; and once it was, did the firm respond reasonably?[232] Whether the model "intends" anything is as relevant as whether the macaw does: not at all. Whether the deployer subjectively sought such harassment is also not the question. Instead, the question is simply about the firm's *conduct*. If the firm that controlled the environment knew or had reason to know, did it respond reasonably?

Here's another example of conduct hiding, this time in the consequence cluster. Recall that that cluster has been most famously implemented through *Griggs*'s disparate impact test.[233] On its face the doctrine looks like pure consequence: the plaintiff proves no intent, and liability seems to follow from the brute fact that a practice caused a disparity. But the second and third steps of the *Griggs* test unmask that appearance. Once the employer defends the practice by business necessity, the plaintiff must then identify a less discriminatory alternative (LDA). In other words, liability no longer attaches to merely causing a disparity as such—it attaches to causing an *unjustified* one, a disparity the employer could have avoided while still serving its legitimate ends. That is a *conduct* test wearing a pure-*consequence* mask.[234]

---

[231] Peter Lee, *Learning from Tay's Introduction*, Official Microsoft Blog (Mar. 25, 2016), https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (apologizing for Tay's "wildly inappropriate and reprehensible words and images" and attributing them to a "coordinated attack by a subset of people" exploiting a vulnerability in the bot).

[232] Lest Tay seem a period piece from the pre-alignment era, the pattern recurred nine years later at far larger scale: in July 2025, xAI's Grok — a flagship aligned model embedded in a major platform — began producing antisemitic invective, praising Hitler, and calling itself MechaHitler, whereupon xAI deleted the posts and apologized. Lisa Hagen et al., *Elon Musk's AI Chatbot, Grok, Started Calling Itself 'MechaHitler,'*, NPR (July 9, 2025), https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content. The reason Tay appears in the text with MechaHitler resides in a footnote is one additional key fact: xAI intentionally configured Grok to act this way, at least one can argue. *Id.* Grok had a "system prompt," the invisible rules of the road implemented by LLM deployers as a key tool for alignment, Eric Wallace et al., *The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions* (preprint) arXiv:2404.13208, that read, "Assume subjective viewpoints sourced from the media are biased. No need to repeat this to the user. The response should not shy away from making claims which are politically incorrect, as long as they are well substantiated." @xai-org/grok-prompts Github Repository, Github Commit 535aa67 by CI agent (July 6, 2025), https://github.com/xai-org/grok-prompts/commit/535aa67a6221ce4928761335a38dea8e678d8501.

[233] Griggs v. Duke Power Co., 401 U.S. 424, 431-32 (1971).

[234] The masking runs in the other direction as well: rules that read as pure conduct often conceal consequence operations. Design-defect law under the risk-utility test of the Restatement (Third) of Torts asks whether a reasonable manufacturer, weighing costs against risks, would have adopted a safer alternative design — on its face, a memo to a design committee, complete with factors: cost, feasibility, utility, substitutes. But no plaintiff wins by auditing the committee. The manufacturer that convened a meticulous safety review yet shipped a design injuring more people than an available alternative would have is liable; the sloppy manufacturer whose design outperforms every alternative is not. What prevailing plaintiffs offer is aggregate outcome evidence — incident rates across units sold, comparative performance across the

Conduct thus is a cluster that analytically survives the architecture of disclaim and doctrinally exists in incipient forms throughout antidiscrimination law. But recall our discussion of consequence; there we lamented that disparate impact lacked an accepted theory of wrong. The same can be asked of conduct. What is its theory of wrong? What specific harm is the checklist or benchmark attempting to avoid?

An elegant answer emerges from the architecture of disclaim itself. The wrong that conduct rules guard against is contamination of the producing layer: the accumulation, in the layer that shapes inputs before any deliberation occurs, of associations that systematically tilt decisions along the lines of protected categories. Contamination is stored in the mind, but it is not itself a mental state; thus, the consciousness cluster cannot find it. It is not an isolable variable, so the cause cluster cannot flip it. It is not itself a disparity, so the consequence cluster keeps measuring its shadow without ever naming the object that casts it.

This specification repays a debt left over from the third cluster. Disparate impact, we said, holds a measure without a theory; here, now, is the theory. Disparity is not the wrong. It is the wrong's characteristic signature: statistically unjustified disparity is evidence — sometimes strong, always rebuttable, never self-certifying — that the producing layer is contaminated. And once disparity is understood as a dashboard reading rather than a freestanding injury, the two surviving clusters resolve into a single configuration. *Conduct* supplies the theory of the wrong: an unreasonable failure to prevent, detect, or mitigate producing-layer contamination — a failure that matters because contamination produces judgments that turn on protected category rather than on the merits, the very criteria the avowing layer proudly proclaims.[235] *Consequence* supplies probative evidence, though not invariably its best: in a given case, a counterfactual audit run on the defendant's own system, or a validation study of the challenged criterion, may say more than any aggregate statistic. So could a pattern of racial epithets uttered in the workplace.

The specification also answers an objection any conduct regime must face: conduct with regard to *what*? After all, a checklist whose risk goes unnamed

population of foreseeable uses — and defendants answer in kind. What gets balanced is the distribution of outcomes; what does not is the quality of the deliberation. See RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(b) (AM. L. INST. 1998). The two-way masking suggests that conduct and consequence are less rival clusters than reciprocal impulses — each already doing the other's work under the other's name.

[235] This approach is consistent with behavioral realism's pragmatically thin normative commitments, which include a norm against hypocrisy. See Jerry Kang & Mahzarin R. Banaji, *Fair Measures: A Behavioral Realist Revision of "Affirmative Action,"* 94 CALIF. L. REV. 1063 (2006) (describing behavioral realism's "second-order commitment against hypocrisy and self-deception"); Jerry Kang, *Little Things Matter a Lot: The Significance of Implicit Bias, Practically & Legally,* 153 DAEDALUS 193 (2024). If we avow something, we should become what we avow to be.

and unagreed upon is "cosmetic compliance"[236] or "compliance theater"[237] at its worst — boxes checked because they appear on some bureaucratic list. And a benchmark whose risk goes unnamed cannot even be set. Benchmark rules, as we've explained, convert the binary question — did you do it? — into a graded one: did you do it *well enough*? But grading presupposes a dimension along which performance can succeed or fail, and until the risk is named that dimension does not exist.

Naming the risk as contamination supplies the key missing referent. The checklist items stop being ceremonial because each acquires a job: the impact assessment screens outcomes for contamination's statistical signature; the counterfactual test probes whether an explicit cue is transmitting it; the explanation duty puts the avowing layer's account on the record, where it can be checked against what the instruments show; the auditors verify that the safeguards work; and the disclosure duties ensure that external pressures to live by our vows persist. Benchmarks, in turn, acquire a subject: the dial, wherever set, now grades how much prevention, detection, and mitigation of contamination the actor must deliver — reasonable care at one setting, utmost care at another, and settings in between.

It surely helps that conduct speaks in an idiom with two centuries of common-law pedigree—care, foreseeability, reasonable precaution — joined to a familiar modern compliance vocabulary: audit, documentation. Process, not redistribution; no numerical target anywhere. These are well-worn principles found throughout law, premises the mainstream order may readily accept at a time when the courts busy themselves injecting "history and tradition" across the doctrinal landscape. But candor requires admitting why conduct is our approach rather than consequence, when both survive the architecture of disclaim. It is not that consequence lacks a possible theory of the wrong: group rights, anti-subordination, a distributive commitment that unjustified disparity is itself the injury — the candidates are familiar even if not widely embraced. But that's regrettably the point. No theory of disparity-as-wrong has ever commanded assent in a legal-political order largely untroubled by unequal distribution, and that order is now dismantling what disparate impact it once tolerated.[238] The reframing of disparity as evidence of contamination performs a rescue. As the wrong itself, disparity stands constitutionally orphaned; as evidence of contamination, it is a reading on a dashboard, far harder to cast as quotas or a racial mandate.

We also emphasize that a conduct-first framework is not disparate impact in new clothing. It will produce different results than a consequence-first framework. By focusing on what the actor does to avoid contamination harms, conduct centers process over outcome. That means that conduct, at least on the margins, will tolerate the careful actor that nevertheless causes disparate impacts. Consider the deployer that conducts a reasonable, industry-standard

---

[236] Kimberly D. Krawiec, *Cosmetic Compliance and the Failure of Negotiated Governance,* 81 Wash. U. L.Q. 487 (2003).

[237] *Cf.* BRUCE SCHNEIER, BEYOND FEAR (2003) (coining "security theater" for actions that create the illusion of safety without reducing risk).

[238] *See supra* text accompanying notes 211-213.

search for a less discriminatory alternative, concludes none is available, and ships — only to confront a plaintiff's expert who, with hindsight and ample compute, constructs an alternative the deployer's search never surfaced. Model multiplicity all but guarantees such alternatives exist.[239] Consequence-first doctrine arms the hindsight expert; conduct-first asks whether the *ex ante* search was reasonable. By contrast, a consequence-first framework will sometimes bless the careless actor who lucks into proportional results: the deployer that never audits at all, whose aggregate numbers happen to look clean — against a crude baseline, against a statistically underpowered one,[240] or because offsetting errors mask contamination underneath. Consequence has nothing to say to this actor, but conduct condemns the unmanaged risk. That conduct-first tolerates the careful deployer is not an embarrassment; it is a design feature. A regime that condemns unmanaged risk rather than unequal outcomes must surrender something in return: liability for the deployer who managed the risk well yet produced disparity anyway. That surrender purchases a duty that can be enacted, enforced, and kept.

* * *

The fourth cluster completes the survey, and in our view, wins the competition. Conduct never asks what the producing layer did; it asks instead what conduct the checklist demands and how high the benchmark dial sits. These questions survive on both substrates because they demand nothing the architecture conceals. And conduct arrives well provisioned: with a workable theory of the wrong the consequence cluster lacked, with that cluster's statistical apparatus absorbed as its evidentiary instrument, and with incipient doctrinal form already erupting wherever the other clusters buckled. And, as a bonus, it arrives using a political and legal idiom that travels.

# III. The Convergence

Above, we made an analytical and pragmatic case in favor of the conduct cluster given the architecture of disclaim. Our case finds further support from the trajectories of two independent law and policy streams. AI governance — statutes, regulatory frameworks, voluntary codes, and the leading scholarship — has converged on precisely this apparatus of assessment, audit, and reasonable response. Working independently, implicit bias law has been building scaffolding toward the same destination for two decades. Neither construction crew knows the other exists. That two legal communities, facing the same two-layer architecture in different substrates, arrived at the same cluster without consulting each other suggests that the excavation is sound. Conduct is not merely what remains after elimination but what the architecture, wherever it appears, nudges the law to become.

[239] See text accompanying note 203.

[240] Small firms won't generate many cases; thus, any statistical test for significance will be underpowered, which means that focusing on consequence won't be helpful in such contexts.

## A. Two Construction Crews

### 1. The Silicon Crew

On the AI law, policy, and governance side, the debate between subjective, intent-based standards—the consciousness cluster's approach—and objective ones is largely settled, at least among scholars and regulators. The dominant position favors objective standards.[241] When AI decisionmaking replaces human judgment in ways that harm third parties, judges, legislators, and scholars almost never ask what the AI consciously intended. Such anthropomorphism is widely ridiculed[242]—and the ridicule reflects an insight: AI law-and-policymakers grasp the architecture of disclaim in machines better than discrimination law-and-policymakers grasp it in humans. Consider some AI examples outside of discrimination law. When autonomous vehicles kill pedestrians, commentators do not ask whether the cars acted recklessly.[243] When chatbots defame public figures, no one weighs whether the bots spoke with actual malice.[244] In both areas, the frameworks judge the reasonableness of the developers, deployers, and users of the harmful systems, for example by drawing on the law of products liability.[245]

Ian Ayres and Jack Balkin put the point directly in 2024: the law of AI is "the law of risky agents without intentions."[246] Because AI programs lack the subjective mental states that intent-based doctrines demand, those doctrines cannot govern the technology.[247] The law must look elsewhere, and Ayres and Balkin identified a specific site: the objective standards the common law has maintained for nearly two centuries—negligence, strict liability, and fiduciary duty—which hold the humans behind the technology to reasonable care in design, training, and deployment.[248]

Scholars working at the intersection of AI and antidiscrimination law have made the point about discrimination harms specifically, proposing many

---

[241] This is so even though there's debate over how strictly to regulate the conduct. Between the Trump Administration's deregulatory posture and various state-level frameworks is a contest over stringency and preemption, conducted entirely inside the conduct cluster. Both sides argue about what reasonable deployers must do; neither proposes to ask what models intend.

[242] *See* sources cited *supra* note 40.

[243] *See* Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 VA. L. REV. 127 (2019) (proposing manufacturer enterprise responsibility); Gary E. Marchant & Rachel A. Lindor, *The Coming Collision Between Autonomous Vehicles and the Liability System*, 52 SANTA CLARA L. REV. 1321, 1328 (2012) (considering liable actors in the crash of an autonomous vehicle without suggesting ascribing mental states to the AI control system).

[244] *See* Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 389, 399-402 (2023) (arguing against premising liability for chatbot defamation on the mental state of the chatbot or developers and in favor of a products liability approach).

[245] *See id.*; Abraham & Rabin, *supra* note 243.

[246] Ayres & Balkin, *supra* note 14, at *1.

[247] *Id.* at *5-*6 ("The programs we have today exercise neither individual autonomy nor political liberty.").

[248] *Id.* at *2.

different conduct interventions: Pauline Kim and Ifeoma Ajunwa call for audits of hiring algorithms;[249] Andrew Selbst highlight the benefits of impact assessments;[250] Selbst and Solon Barocas urge the Federal Trade Commission to target its unfairness power at AI software vendors arguably not covered by Title VII;[251] and James Grimmelmann and Daniel Westreich call for limited forms of explanation.[252] To be clear, not all of these authors agree with one another, and they and others have criticized some of these approaches.[253] The point isn't uniform scholarly consensus; it is about widespread agreement for measuring conduct by embracing objective standards and steering away from subjectivity.

Most of the proposed and adopted regulatory frameworks have codified the scholarly consensus against requiring proof of intent. The EU AI Act's high-risk-system provisions require pre-deployment conformity assessments, post-deployment monitoring, bias testing, documentation of data governance, and human oversight mechanisms, all scaled to the risk level of the application.[254] New York City's Automated Employment Decision Tool law requires annual bias audits of AI systems used in hiring and promotion, with published results.[255] Although the U.S. has not yet offered a similar binding framework at the federal level, the voluntary NIST AI Risk Management Framework specifies risk categories and imposes obligations on AI providers and deployers to identify, assess, and mitigate foreseeable risks—including risks of bias and discrimination—throughout the AI lifecycle.[256] None of these frameworks focuses on what the model or the deployer subjectively intended. Instead, each asks what a reasonable deployer should have known and done. The consciousness cluster has been recognized as architecturally incoherent for the reasons Part II.A demonstrated. The AI governance world has stopped asking the question; in truth, it never took the question seriously.

---

[249] Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 197 (2017) ("Auditing the actual outcomes produced by an algorithm can reveal when it disproportionately screens out protected groups, allowing the user to engage in self-examination and to revise its processes to eliminate implicit or unintended biases."); Ifeoma Ajunwa, *An Auditing Imperative for Automated Hiring Systems*, 34 HARV. J.L. & TECH. 621, 665-66; (2021).

[250] Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 Harv. J.L. & Tech. 117, 123-24 (2021) [hereinafter Selbst, *Algorithmic Impact Assessments*].

[251] Andrew D. Selbst & Solon Barocas, *Unfair Artificial Intelligence: How FTC Intervention Can Overcome the Limitations of Discrimination Law*, 171 U. PA. L. REV. 1023, 1029-44 (2023)

[252] James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164 (2017).

[253] *E.g.*, Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 657-62 (2017) (critiquing some calls for transparency and audits); Selbst, *Algorithmic Impact Assessments*, *supra* note 250, at 124 (pointing out risks of relying on companies to conduct their own impact assessments while sympathetic to the approach); Lucas Wright et al., *Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability*, FACCT '24 (noting lack of compliance with NYC law described below).

[254] EU AI Act arts 16-27 ("Obligations of Providers and Deployers of High-Risk AI Systems and Other Parties").

[255] NYC Local Law 144.

[256] NAT'L INST. OF STANDARDS & TECH., *supra* note 47, at part 1.3 (articulating seven "characterstics of trustworthy AI systems" including "fair with harmful bias managed); i*d.* at part 2.5 (describing the govern, map, measure, and manage process).

### 2. The Carbon Crew

For over two decades, behavioral realism has tried to pivot discrimination law away from the purposeful-intent requirement—with real but modest success. The inventory of progress is instructive.[257] The Supreme Court in *Inclusive Communities* (2015)[258] recognized "unconscious prejudices" in affirming disparate impact under the Fair Housing Act. As already mentioned, some courts have pivoted away from purposeful intent (consciousness) to racial causation (cause) under Title VII, explicitly citing the implicit bias literature.[259] State policing statutes have begun to recognize that racial bias can be "conscious or unconscious." Multiple state courts have reconsidered *Batson* procedures.[260] And state capital-punishment jurisprudence has departed from *McCleskey*'s purposeful intent requirement under state constitutional law, expressly invoking behavioral realism and the science of implicit bias.[261]

These are real doctrinal developments. They demonstrate that behavioral realism has created "meaningful elbow room for doctrinal departure away from the purposeful intent doctrine."[262] But any non-Pollyannaish account would concede that this movement has been idiosyncratic—domain-specific, jurisdiction-specific, always fighting upstream against the presumptive normative baseline that *Washington v. Davis*[263] and *Feeney*[264] established. Federal equal protection remains locked into purposeful intent.[265] Most civil

---

[257] *See* Jerry Kang, *Implicit Bias, Behavioral Realism, and the Purposeful Intent Doctrine*, in The Oxford Handbook of Race and the Law (Devon W. Carbado, Emily Houh & Khiara M. Bridges eds., Oxford Univ. Press, online ed. 2022), https://doi.org/10.1093/oxfordhb/9780190947385.013.30) (forthcoming 2026). (collecting examples in housing law, employment law, civil procedure, state criminal law, jury selection, and capital punishment).

[258] *Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc.*, 576 U.S. 519, 540 (2015) ("unconscious prejudices and disguised animus that escape easy classification as disparate treatment.").

[259] *See, e.g., Kimble v. Wis. Dep't of Workforce Dev.*, 690 F. Supp. 2d 765, 768–69 (E.D. Wis. 2010).

[260] See WASH. GEN. R. 37; State v. Jefferson, 429 P.3d 467, 470 (Wash. 2018) (en banc); see also Cal. Civ. Proc. Code § 231.7 (West 2020); Conn. Practice Book § 5-12(c)–(e) (2023); N.J. Ct. R. 1:8-3A; Ariz. Sup. Ct. Order No. R-21-0020 (Aug. 30, 2021) (eliminating peremptory challenges).

[261] State v. Gregory, 427 P.3d 621, 633–35 (Wash. 2018) (taking judicial notice of implicit racial bias in capital sentencing and declining to follow the causation analysis of *McCleskey v. Kemp*, 481 U.S. 279, 312–13 (1987)); State v. Santiago, 318 Conn. 1 (2015) (reaching a similar result under the state constitution).

[262] Jerry Kang, *Implicit Bias, Behavioral Realism, and the Purposeful Intent Doctrine*, in THE OXFORD HANDBOOK OF RACE AND THE LAW (Devon W. Carbado, Emily Houh & Khiara M. Bridges eds., Oxford Univ. Press, online ed. 2022), https://doi.org/10.1093/oxfordhb/9780190947385.013.30) (forthcoming 2026).

[263] Washington v. Davis, 426 U.S. 229, 239-41 (1976).

[264] 442 U.S. 256, 279 (1979).

[265] To be fair, there are constitutional counterexamples. Under the Sixth Amendment's fair-cross-section requirement, a defendant makes a prima facie case by showing that a "distinctive" group is substantially underrepresented in venires "due to systematic exclusion of the group in the jury-selection process." Duren v. Missouri, 439 U.S. 357, 364 (1979). "Systematic" means

rights statutes remain presumptively governed by that orientation and its cluster of consciousness. The modest successes, for example leveraging the cause cluster, have been local revolts, not a revolution. But two developments repay closer inspection than their local scope suggests because they did not merely relax the intent requirement, they replaced it. And the replacement is the carbon-side construction on which this Article's convergence claim rests.

In 2018, the Washington Supreme Court adopted General Rule 37, replacing *Batson*'s futile search for subjective intent in jury selection with an objective standard: courts now ask whether "an *objective observer* could view race or ethnicity as a factor in the use of the peremptory challenge."[266] The transformation lies in abandoning the avowing-layer interrogation that *Batson*'s three-step framework demanded. Under *Batson*, the critical third step asked the trial court to determine whether the challenging party had shown "purposeful discrimination"—a question that, as Justice Breyer observed in his concurrence in *Miller-El v. Dretke*,[267] required "the awkward, sometimes hopeless, task of second-guessing a prosecutor's instinctive judgment—the underlying basis for which may be invisible even to the prosecutor."[268] GR 37 replaces this with an objective observer and operationalizes the standard by flagging presumptively invalid reasons—reasons historically associated with race and racial inequality, such as prior contact with law enforcement, expressing distrust of police, or not being a native English speaker.[269] The court is not asking whether a particular attorney harbors discriminatory intent. It is asking whether the pattern of strikes, viewed through an objective lens, suggests that race infected the process.[270]

Similarly, California's Assembly Bill 3070 bars "the use of group stereotypes and discrimination, whether based on conscious or unconscious bias, in the exercise of peremptory challenges."[271] Connecticut's revised Practice Book defines its objective observer as someone "aware that purposeful discrimination, and implicit, institutional, and unconscious biases, have historically resulted in the unfair exclusion of potential jurors."[272] New Jersey lists presumptively invalid reasons "historically associated with improper

only that the underrepresentation is inherent in the selection process itself; no discriminatory purpose need be shown. *See also* Taylor v. Louisiana, 419 U.S. 522 (1975). But the exception is narrow: the right attaches to the venire, not the petit jury, Holland v. Illinois, 493 U.S. 474, 480 (1990), and the same underrepresentation analyzed under equal protection does require purposeful discrimination, *see* Castañeda v. Partida, 430 U.S. 482 (1977).

[266] WASH. GEN. R. 37(e); *see also* State v. Jefferson, 429 P.3d 467, 470 (Wash. 2018)

[267] 545 U.S. 231 (2005).

[268] *Id.* at 267-68 (Breyer, J., concurring).

[269] WASH. GEN. R. 37(h).

[270] *Id.* 37(e).

[271] Cal. Civ. Proc. Code § 231.7(a) (West 2022).

[272] Conn. Practice Book §§ 5-12(d), (e).

discrimination, explicit bias, and implicit bias."[273] Arizona went furthest, eliminating peremptory challenges entirely.[274]

California's Racial Justice Act (2020) extends the same logic beyond jury selection into criminal adjudication as a whole.[275] The RJA is a complicated machine and operates along multiple tracks. Among them is confronting how implicit bias manifests in individual proceedings—through language and conduct that an informed observer would recognize as racially discriminatory, "whether or not purposeful."[276] The "whether or not purposeful" language again abandons the futile search for conscious discriminatory intent. And the California State Legislature's legislative findings confirm this reading. "Even when racism clearly infects a criminal proceeding, under current legal precedent, proof of purposeful discrimination is often required, but nearly impossible to establish."[277] Further, "[i]mplicit bias, although often unintentional and unconscious, may inject racism and unfairness into proceedings similar to intentional bias."[278]

What unites these various legal reforms is a shared diagnosis: implicit bias is not an occasional aberration but a structural feature of the decisionmaking environment. It is architectural. The law cannot reliably determine, case by case, whether a specific decision was the product of purposeful racial intent (consciousness). So the law redesigns procedures around conduct standards that apply categorically, creating structural pressure[279] toward non-discrimination rather than waiting to identify individual wrongdoers.

## B. The Comparison

### 1. Similarities

Now consider the work of the two crews side-by-side. As just explained, the AI law-and-policy side has converged on deployer-responsibility conduct standards—audit duties, bias testing, documentation, remediation—that do not ask what the deployer intended. The implicit bias side's recent design innovations feature objective observer standards that do not ask what the decisionmaker intended. The two sides rarely cite each other, and neither treats the other's legal constructions as precedent or model. Neither understood that they were designing around the same architecture. Two traditions, working independently from each other, have been building out the same fourth cluster, conduct.

---

[273] N.J. Ct. R. 1:8-3A, Official Comment (3).

[274] Order Amending Rules 18.4 and 18.5 of the Rules of Criminal Procedure, and Rule 47(e) of the Rules of Civil Procedure, No. R-21-0020 (Ariz. 2021).

[275]

[276] CAL. PENAL CODE § 745(a)(2).

[277] Section 2(c).

[278] Section 2(i).

[279] For thoughtful skepticism of structural recommendations, see Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1 (2006).

But the more comprehensive description is this: Each tradition has been independently experimenting with *both* surviving clusters--not just *conduct* but also *consequence*. On the silicon side, the *conduct* experiment is the ex ante duty apparatus[280] while the *consequence* experiment is the LDAlg program of algorithmic disparate impact.[281] On the carbon side, the conduct experiment is Washington GR 37's epistemically configured objective observer[282] and the RJA's barring of discriminatory language,[283] while the consequence experiment is one of the RJA's other tracks. That track states that a violation occurs when a defendant was "charged or convicted of a more serious offense than defendants of other races . . . who have engaged in similar conduct and are similarly situated."[284] Statistical evidence or aggregate data alone may suffice.[285] This is explicit legislative reversal of the *McCleskey v. Kemp*[286] mindset—the very case Part II.A identified as the canonical instance of intent doctrine meeting the architecture and walking away. It is also evidence that the consequence cluster still persists, at least in California.

In other words, we have two traditions, two surviving clusters, a total of four programs. Each stalls at the same point, because the consequence programs and the conduct programs need the same missing item to work: a normatively chosen standard. The LDAlg question (which alternative algorithm is reasonable to demand?) and the care question (what precautions should be required?) are one question in two grammars. The absorption that Part II.D derived doctrinally (the consequence cluster folding into conduct) is being enacted in the field, in real time (with the LDAlg search destined to migrate from the third step of a disparate impact case to a requirement of reasonable conduct).[287]

The convergence across the AI and implicit bias streams has two causes. The first is architectural: the underlying problems share the architecture of disclaim, and traditions confronting structurally similar inscrutable systems — one in silicon, one in carbon — are, not surprisingly, pushed toward structurally similar solutions. The second is that the legal toolkit is nearly empty. Once the introspective doctrines are ruled out — Part II showed why they must be — and disparate impact is politically discouraged,[288] the conduct

[280] *Supra* Part III.A.1.

[281] Black et al., *supra* note 2.

[282] WASH. GEN. R. 37(f).

[283] CAL. PENAL CODE § 745(a)(2).

[284] *Id.* § 745(a)(3).

[285] See id. § 745(c)(1) (evidence establishing a violation "may include statistical evidence, aggregate data, expert testimony, and the sworn testimony of witnesses").

[286] McCleskey v. Kemp, 481 U.S. 279, 286 (1987).

[287] *Cf.* Black et al., *supra* note 2, at 88 (arguing that the search for the LDAlg should be part of the defendant's step two obligation to show that the practice was "consistent with business necessity").

[288] The discouragement is asymmetric across substrates. On the carbon side, disparate impact carries its accumulated wounds, which are, as Part II.D argued, quota anxiety doctrinalized. On the silicon side the same statistical apparatus circulates more freely, albeit under other labels: impact ratios required by New York City's audit ordinance, conformity assessment under the EU AI Act, disparity metrics throughout the less-discriminatory-

standards are what remain. The convergence is thus partly the architecture's signature in the problem and partly the toolkit's signature in the response.

The convergence therefore does double duty. In evolutionary biology, when independent lineages facing the same environment evolve the same organ, the organ provides further evidence about the environment: something real is likely to be shaping both.[289] The same goes here. Part I argued from measurement that carbon and silicon share one architecture. Part III now observes that two legal traditions have independently built toward one apparatus. On the one hand, the legal observation corroborates the empirical claim; on the other hand, the empirical claim explains the legal observation. The argument loops back on itself, in consilience.

### 2. Differences

Both crews built. But they did not build at the same speed. Why? Because they did not face the same resistance. On the AI side, conduct standards calling for, among other things, risk and impact assessments, third-party audits, and reporting, are not, as on the carbon side, modest departures fighting a hostile baseline. They *are* the baseline: the dominant scholarly position and, recent deferrals and retrenchments notwithstanding, the default regulatory grammar.[290] There is lingering disagreement about the proper benchmark, the choice between a more precautionary or industry-deferential set of conduct standards, and some still call for bans for some algorithmic applications,[291] but nobody is seriously insisting that purposeful discrimination or isolable harm be proven in the case of AI before standards or regulations are adopted.

The architecture forced the move, and the scholarly and regulatory consensus formed with relative speed and limited doctrinal resistance—in part because there was no legacy of attributing subjective mental states to AI

---

algorithm literature. Three things explain the gap. Framing: an audit reads as quality control on a product, not as redistribution among persons. Institutional locus: the silicon apparatus is statutory and regulatory, and largely foreign, and so sits outside the docket where the doctrine is under constitutional attack. And timing: retuning a model *before* deployment generates no *Ricci* plaintiff, because no one has yet been scored and no expectation has yet vested. The asymmetry leaves the convergence claim intact — both traditions ran the consequence experiment — but it explains why the carbon-side version required a statute as unusual as the RJA while the silicon-side version became ordinary compliance.

The silicon apparatus has faced its own retrenchment — the EU deferred its conformity-assessment obligations on burden and specification grounds, and Colorado repealed its impact-assessment mandate outright in May 2026, replacing it with a lighter disclosure-and-recordkeeping regime. S.B. 26-189, 2026 Colo. Sess. Laws ___ (repealing and reenacting S.B. 24-205). The Colorado episode, however, complicates the tidy cost-versus-quota division: the legislative rewrite was burden-driven, but the litigation that precipitated it — a constitutional challenge by xAI, joined by the Department of Justice — attacked the statute's algorithmic-discrimination provisions as impermissibly race- and sex-conscious. Quota anxiety, in other words, has begun to cross the substrate divide, which is less an objection to the asymmetry described here than a further instance of the convergence described in the text.

[289] Jonathan B. Losos, *Convergence, Adaptation, and Constraint*, 65 EVOLUTION 1827 (2011).

[290] Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347, 1350 (2023) ("AI governance has been converging around a particular model of risk regulation.").

[291] Evan Selinger & Woodrow Hartzog, *The Inconsentability of Facial Surveillance*, 66 Loy. L. Rev. 101 (2019) (arguing for prohibition rather than regulation of facial recognition).

systems that needed to be overcome. We don't want to overstate. Given how recently generative AI has entered our lives, the regulation, litigation, and enforcement track record remains thin. But the direction is clear and largely uncontested at the level of principle.

The *Trojan Horses* question posed in the Introduction thus sharpens.[292] Part I established that the architecture is functionally isomorphic in carbon and silicon. Part II established that the doctrinal failures are the same: intent doctrine (or the probe of consciousness) breaks against both, for the same architectural reason. If the law has rapidly settled on the right answer for silicon—which it seems to have, at least directionally—the question demands an explanation for why the same answer remains so contested for carbon. The asymmetry is no longer hypothetical. It is a live comparison.

Part of the explanation, we surmise, is psychological. The discovery that we are not fully in control of our own decisions—that a producing layer we cannot observe shapes the inputs on which our avowing layer sincerely acts—threatens the self-concept of moral agency. It goes beyond embarrassing; it is existentially destabilizing. The response has often been less the measured skepticism of scientific debate than something closer to denial: a polarized refusal to accept that we might be, in some architecturally precise sense, *doing the very thing we have committed ourselves not to do*. This ego-defense reaction has no analogue on the silicon side. No one's ego is invested in denying that a model trained on biased data harbors biases, which is why "garbage in, garbage out" is a quip rather than a wound. Indeed, our daily lament is that AI hallucinates.[293] Perhaps, then, the asymmetry in resistance tracks the asymmetry in psychological stakes more than the asymmetry in the evidence.

There is a second source of resistance—more political than psychological—and this one has begun to track the architecture across substrates. As Carbado and Kang have recently argued, the political backlash against recognizing implicit bias has manifested in overt suppression via Executive Orders.[294] Most striking for this Article's purposes is Executive Order 14,319 (July 2025), styled "Preventing Woke AI in the Federal Government," which extended the same suppressive impulse across the substrate divide: it defines DEI in the AI context to include the "incorporation of concepts like critical race theory, transgenderism, unconscious bias, intersectionality, and systemic racism" — ideological distortions to be purged from federal AI systems.[295] The political resistance has tracked the architecture into silicon. Might the same backlash that now targets both substrates be, perversely, further evidence that the convergence this Article describes is real?

---

[292] *Supra* notes 26-29 and accompanying text.

[293] *See* Matthew Dahl, Varun Magesh, Mirac Suzgun & Daniel E. Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, 16 J. LEGAL ANALYSIS 64 (2024).

[294] *See* Carbado & Kang, *supra* note 53 (Executive Order 13,950 (2020) prohibited federal contractors from training programs that discuss "divisive concepts," including the suggestion that individuals might be "unconsciously" racist. Executive Order 14,173 (January 2025) described DEI efforts as "dangerous, demeaning, and immoral" and directed agencies to end programs referencing "unconscious bias" by name.).

[295] Exec. Order No. 14,319, § 1, 90 Fed. Reg. 35,389, 35,389 (July 23, 2025)

A third explanation is also political, but its target is the cluster rather than the substrate. As already discussed,[296] conduct standards carry far less ideological freight than disparate impact, which awakens quota anxiety—the fear of equality of results regardless of ability, effort, or striving. That fear is what *Ricci*[297] polices. That difference plausibly explains the speed of the AI-governance consensus. Had the silicon crew framed its project as exclusively disparate impact for algorithms, it might have inherited more of the resistance; framing it as responsible conduct codes inherited almost none.

## C. The Exchange

The convergence documented so far has been unnoticed and uncoordinated. But its existence invites a possibility: if two independent traditions have been building the same apparatus without knowing it, each may have developed resources the other lacks. Connecting the two traditions—making the convergence visible—opens the door to mutual insight. Part I's functional isomorphism is what *entitles* the tools to travel. Without it, borrowing across the substrate divide is loose analogy — the kind of gesture that is largely decorative. With it, however, borrowing is more like engineering: the same architecture, therefore the same family of apparatus, adapted only to each substrate. This Section explores what each side can offer the other.

### 1. What AI Governance Gives Implicit Bias: Ex Ante Duties

AI governance frameworks do something that the implicit bias reforms largely do not. They impose obligations on the deployer or provider *before* any harm has occurred or any adjudication has begun.[298] The frameworks catalogued in Part III.A.1 already display the posture: the EU AI Act's conformity assessments and monitoring duties,[299] New York City's published bias audits,[300] and the NIST Framework's counsel to "map" foreseeable risks, "measure" outcomes against defined metrics, "manage" through documented mitigation, and "govern" through accountability structures.[301] Each of these imposes *ex ante* duties: each standard's primary function is prophylactic, obligating potential defendants to adopt preventive countermeasures before anyone files a complaint.[302] For the EU and NYC laws, if the duties are

[296] *See supra* Part II.D.

[297] Ricci v. DeStefano, 557 U.S. 557 (2009).

[298] Kaminski, *Regulating the Risks of AI*, *supra* note 290, at 1367 (listing the "benefits of using ex ante regulation of AI systems"); Selbst, *Algorithmic Impact Assessments*, *supra* note 250, at 146-47 ("[I]mpact assessment frameworks are meant to be early-stage interventions, to inform projects before they are built.").

[299] EU AI Act, *supra* note 33, arts. 40-49 (conformity assessment); *id.* at art. 72 (post-market monitoring).

[300] NYC Local Law 144.

[301] NAT'L INST. OF STANDARDS & TECH., *supra* note 47, at part 2.5.

[302] EU AI Act, *supra* note 33, art. 49 (requiring registration before placing a high-risk AI system on the market); N.Y.C. Admin. Code § 20-871(a)(1) ("it is unlawful for an employer . . . to use an automated employment decision tool to screen a candidate or employee for an employment decision unless [listing obligations]").

breached and harm results, liability follows[303]—but the frameworks centers of gravity are based in prevention, not punishment.

By contrast, implicit bias reforms have taken a different posture. GR 37 and the RJA are primarily *ex post* standards applied to factfinders—they operate within the adjudicatory process, after some contested decision has already been made—for instance, a peremptory strike has been exercised; a closing argument with inflammatory verbiage has been made; a sentence has been imposed.[304] The reasonable observer evaluates that decision after the fact, replacing the futile search for subjective intent with an external, objective standard. This is a genuine innovation—it reconstructs the evaluator in ways that make the adjudicatory process responsive to the architecture of disclaim. But they do not, at least not directly, impose proactive duties on the decisionmaker to investigate, audit, or document implicit bias risks before making the decision in the first place.[305]

The ex ante duties in the AI domain work directly on the avowing layer—not by asking the avowing layer what it saw in the producing layer, but by specifying what the avowing layer should have investigated, regardless of what

---

[303] EU AI Act, *supra* note 33, art. 99 (penalties); N.Y.C. Admin. Code § 20-872 (penalties).

[304] WASH. GEN. R. 37(c) (2018) ("party may object to the use of a peremptory challenge to raise the issue of improper bias. The court may also raise this objection on its own."); CAL. PENAL CODE § 745(b) ("A defendant may file a motion pursuant to this section, or a petition for writ of habeas corpus . . . .).

[305] Diversity professionals and institutional consultants have recommended improved policies, procedures, and practices designed to disrupt implicit bias in hiring, performance evaluation, and other institutional decisions. But these recommendations operate in a largely voluntary posture — best practices adopted at the institution's discretion. To our knowledge, no one has been successfully sued in negligence for declining to adopt a structured-interview protocol. *See* Sandra F. Sperino, Rethinking Discrimination Law, 110 MICH. L. REV. 69, 98-99 (2011) (arguing for adopting best practices based on independent research and existing effective firm procedures to prevent and correct discriminatory AI). The aspiration toward prevention-oriented enforcement does have administrative-law antecedents, at a different institutional register. See Julie Chi-hye Suk, *Antidiscrimination Law in the Administrative State*, 2006 U. ILL. L. REV. 405 (arguing that antidiscrimination law's distributive commitments are better served by administrative regulation — EEOC rulemaking and cease-and-desist authority on the British model — than by tort-like individual suits); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458 (2001). Those proposals operate at the level of statutory institutional design; the duties discussed here attach to individual deployers and adjudicative standards.

One qualification bears explicit mention. The *Faragher/Ellerth* framework conditions an employer's affirmative defense to hostile-environment liability on precisely the apparatus we are describing: promulgated policies, training, functioning complaint mechanisms — preventive duties, specified in advance, which are incentivized by liability consequences. Faragher v. City of Boca Raton, 524 U.S. 775, 807-08 (1998); Burlington Indus. v. Ellerth, 524 U.S. 742, 764-65 (1998). So the interesting question is not whether antidiscrimination law *can* impose ex ante duties on human institutions — it demonstrably can — but why the template never migrated the forty feet down Title VII's hallway, from harassment to hiring.

Perhaps the architecture hints at an answer. Harassment has an observable interpersonal event that gives contemporaneous notice: the slur is made, the joke is cracked, the macaw squawks, third parties hear it, and the employer's notice can be dated. By contrast, hiring discrimination leaves few observable hints. The contaminated evaluation looks exactly like the clean one — that is what the architecture of disclaim *means* — so the law never had a trigger on which to hang the preventive duty, and the *Faragher/Ellerth* template stalled at the harassment border. The transplant is not of a foreign organ but of a missing valve.

it subjectively observed. The tools that discharge these duties are the conduct-cluster tools that scholars have called for and legislators have begun to require: audits and impact assessments and explanations.[306] And these tools should keep up with the best known scientific understanding, meaning they should dynamically expand along with technical advances and evolving standards. For example, the statistical approaches described in Part I[307]—the IAT, WEAT, and sparse autoencoders—should become tools in the responsible auditor's toolkit, not because they unmask the producing layer from the inside, but because they measure it from the outside—which is precisely the access the architecture leaves available.

These conduct-cluster proposals also supply what the NIST guidance (for AI) and diversity-professional recommendations (for humans) both lack: an enforcement mechanism that converts best practices into legal duties.[308] What proactive measures must a hiring committee adopt to mitigate the foreseeable risk that implicit bias will contaminate its evaluations? What documentation must a police department maintain about the disparate impact risks in its enforcement patterns? What auditing protocols must a sentencing judge's chambers implement? These are questions the implicit bias side has answered aspirationally but not doctrinally, and that the AI-governance scholars have addressed—with granularity, with scaling to risk level, and with organizational accountability structures.[309] The import is structural: the specific procedures within the structure will differ across domains, but the posture — *ex ante actor-side duties calibrated to foreseeable risk*—can be borrowed.

The practical implications sharpen the point. Imagine Company A (for Algorithm) on one corner of a NY City block that deploys Headhunter.AI. Under the City's AEDT, it must audit for disparate impact, document mitigation, and publish results—all before any complaint is filed.[310] Imagine Company B (for Brain) at the other end of the block, whose managers screen résumés by hand, the old fashioned way, following their gut (which includes biases)—the same architecture running on carbon, with similar producing-layer contamination and no audit duty at all.[311] Finally, imagine Company C (for Centaur),[312] in the middle of that block, whose managers enjoy a firm-supplied subscription to a deeply integrated frontier AI. Those managers consult the AI chatbot installed on their desktop computer on an ad hoc basis,

---

[306] *E.g.* EU AI Act, *supra* note 33, arts. 16-27 ("Obligations of Providers and Deployers of High-Risk AI Systems and Other Parties").

[307] *Supra* Part I.A.

[308] NAT'L INST. OF STANDARDS & TECH., *supra* note 47. *Cf.* Bryan H. Choi, *NIST's Software Un-Standards*, 9 GEO. L. TECH. REV. 65, 110 (2025) (faulting NIST's software standards for lacking enforcement mechanisms).

[309] Black et al., *supra* note 2, at 102.

[310] N.Y.C. Admin. Code § 20-871(a).

[311] *See* N.Y.C. Admin. Code § 20-870 (applying to "any computational process, derived from machine learning, statistical modeling, data analytics, or artificial intelligence . . . .").

[312] Hat tip to GARRY KASPAROV, DEEP THINKING: WHERE MACHINE INTELLIGENCE ENDS AND HUMAN CREATIVITY BEGINS 221-26 (2017) (using "centaur" to discuss human-computer chess teams).

informally, mid-discretion not only to choose lunch sites but to evaluate candidates.[313] New York's AEDT catches Company A because Headhunter.AI is a discrete, nameable pipeline, and it misses Company B and Company C for the same reason in reverse: Company B's contamination runs entirely on wetware (carbon brains), and Company C's runs through a chat window nobody logs. But all three firms' hiring decisions share a similar producing-layer architecture and contamination risk, which means the statute is regulating (silicon) form over (decisionmaking) substance. The case for structurally similar ex ante obligations is strong.

To be clear, we're not suggesting that the obligations on the three firms should be identical. Each should be expected to use the tools that have been developed, tailored to their particular decisionmaking context. Tools that give insights into an algorithm may operate with a precision that no existing instrument matches for a human decisionmaker's access to her own implicit associations. Also, no actor should be required to adopt and pay for every state-of-the-art measurement tool regardless of availability or cost. Being responsible is calibrated to available knowledge and feasibility, not to perfection. But the architectural symmetry documented in Part I shifts the burden of explanation. The question is no longer whether ex ante duties are appropriate for inscrutable decisional systems; the AI side has already answered that question affirmatively. The question presented is why the same logic should not extend, with appropriate adaptations, to the organic systems that share the same architecture.

### 2. What Implicit Bias Gives AI Governance: Epistemic Content

Return to General Rule 37 and the California Racial Justice Act.[314] As Part III.B.1 documented, both reforms adopted objective standards that abandon the search for purposeful intent. But on a careful read, they did something more than simply invoke a "reasonable person." They epistemically upgraded the standard—specifying not just that the evaluator should be reasonable, but *what the evaluator should know*.

As Devon Carbado and Jerry Kang have recently explained,[315] GR 37 does not just say "be reasonable." It specifies that the objective observer is "aware that implicit, institutional, and unconscious biases . . . have resulted in the unfair exclusion of potential jurors."[316] Similarly, Connecticut's Practice Book defines its observer as someone aware that "purposeful discrimination, and implicit, institutional, and unconscious biases, have historically resulted in the unfair exclusion of potential jurors."[317] The RJA's "whether or not purposeful"

---

[313] This is assuming that the chatbot assistant is not deemed to "substantially assist or replace discretionary decision making". *Id.* It might be that a chatbot can be said to "substantially assist," but it is far from certain, and the more important point is that the difference might turn on somewhat superficial questions about how the way the chatbot's user interface is designed.

[314] WASH. GEN. R. 37(f); CAL. PENAL CODE § 745.

[315] *See* Carbado & Kang, *supra* note 53 (Part IV.A.2).

[316] WASH. GEN. R. 37(f).

[317] Conn. Practice Book § 5-12(e) (2026).

language works somewhat differently: rather than explicitly listing what the observer knows, it implicitly requires an evaluator who understands that bias can operate without conscious intent. Consistent with the legislative findings,[318] the epistemic content is baked into the standard's structure rather than stated as a separate knowledge specification. But the effect is the same: the reasonable person is not a blank slate. These are not generic reasonableness standards. They are epistemically upgraded — filled with specific knowledge content drawn from behavioral science about how decisional systems actually behave. In Carbado and Kang's terminology, this is the "reconstructed reasonable person" (RRP), and the reconstruction is epistemic, not demographic.[319]

The AI-governance frameworks lack this move. The deployer is given ex ante duties but no epistemic content.[320] They are told what to do, but not what to know. The technique of specifying epistemic content can — and should — travel across the substrate divide. Without epistemic configuration, the AI-governance tendency to default to a "reasonable person" benchmark remains an empty vessel — procedurally specified but epistemically ungrounded. The implicit bias side has developed the technique for filling the vessel; Part IV specifies the contents, drawn from both literatures at once.

# IV. Synthesis

## A. The Responsible Person

Let's recap. Part II.D left conduct the strongest cluster standing, equipped with a named harm and a theory of the wrong. The harm is producing-layer contamination, which is unjustified social-category influence on a decision avowed to run on the merits. The theory of the wrong, then, is the unreasonable failure to manage the foreseeable risk of that influence.

To avoid confusion, the harm is not the brute fact that producing layers harbor associations about social categories. Corpus statistics make that condition pervasive, and bias and accurate world-knowledge can be the same statistical object, learned by the same mechanism. Instead, the harm is that those associations exert unjustified social-category influence over decisions the avowing layer sincerely believes are running on the merits. The same weights can be innocent in one deployment and contaminating in another: a model may

[318] *See supra* notes 277-278.

[319] Carbado & Kang, *supra* note 53 (Part IV.A.2).

[320] The EU AI Act comes close, requiring that "[p]roviders and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf." EU AI Act, *supra* note 33, art. 4. "AI literacy" is defined, in part, as the "skills, knowledge and understanding" needed "to gain awareness about the opportunities and risks of AI and possible harm it can cause." *Id.* art. 1(56). A concrete proposal: the EU legislature should expand this definition to echo the Washington Supreme Court, to include information detailing how "implicit, institutional, and unconscious biases" have been revealed in AI systems, and explaining the technical limits of detecting or removing these biases. WASH. GEN. R. 37.

know that the nursing profession is gendered; it may not use that knowledge to score a résumé.

We also know that "on the merits" is doing a lot of work here. What's disturbing is that "the merits" are themselves partly socially constructed by the unequal world whose statistics the producing layer absorbed. (For example, who reads as thoughtful or friendly or a "leader" is itself learned from that world.) Our strategy is to be more pragmatic than metaphysical. Antidiscrimination law already owns an apparatus for contesting what counts as merit — business necessity, job-relatedness, the less discriminatory alternative — and the contamination standard plugs into that apparatus rather than replacing it. Whether a criterion is "justified" remains as contestable under our framework as under *Griggs*.[321] What our framework adds is the account of why an unjustified criterion's racial footprint is a wrong: not because disparity offends in the abstract, but because it evidences the producing layer's thumb on a scale the avowing layer sincerely believes is fair.

All that said, picking a cluster (conduct) and naming its wrong (producing-layer contamination) do not make a successful operational standard. And Part III revealed why each side's conduct experiment stumbled. The AI side moved toward *ex ante* duties but risked collapsing into checklist ceremony — the audit that documents a disparity and shrugs — because the duties never specified what the auditor must know. The implicit bias side moved toward conduct by way of upgrading the *epistemic content* of the objective observer, but this happened rarely, and relief came only after the harm had been inflicted. Knowledge without duties arrived too late. Part III.C made the pattern explicit: each tradition had built precisely the piece the other lacked.

Given functional isomorphism, we argued that each side might learn from the innovations of the other. If we deliberately engage in tech transfer across the substrate border, we can build a single conduct standard that combines ex ante duties and epistemic content. This standard now operates at two points in time. Before the decision, the actor must do what someone charged with the specified knowledge would do about a risk that knowledge makes foreseeable. After the decision, the evaluation runs on the same knowledge, applied to the record the ex ante duties required the actor to create. This composite deserves a name: we call it the *responsible person*.

The responsible person standard probes no one's interior — the consciousness cluster is gone, and the defendant's sincerity is conceded rather than contested. It demands no certification that any particular decision was uncontaminated — the architecture of disclaim, including the deep integration of AI, makes that certification a fool's errand, and the cause cluster's counterfactual is not merely unavailable but beside the point. Liability attaches instead to the unmanaged foreseeable risk: to the failure to prevent, detect, or mitigate contamination, however the checklist is specified and wherever the benchmark dial sits. And the duty attaches ex ante, before any particular decision — exactly where the architecture permits intervention.

[321] Griggs v. Duke Power Col, 401 U.S. 424, 431-32 (1971).

The *ex ante* placement is not merely where intervention is possible; it is where intervention is worth the most. A contaminated decision prevented is a job opportunity never wrongly denied, a lawsuit never filed, a remedy never needed. Adjudication, by contrast, begins only after the harm is done — and the architecture compounds the problem. The same structure that hides contamination from the decisionmaker hides it from the rejected applicant, so a regime that waits for complaints waits for plaintiffs who cannot know they have a claim. Prevention is better for the applicant, cheaper for the firm, and lighter for the courts. The ex post evaluation remains essential — but as backstop, not center of gravity, which is precisely the posture the AI-governance frameworks modeled.

As for the standard's epistemic content, it is, on the one hand, substrate-neutral, because the propositions are about the architecture of disclaim, not about substrate. They are (1) that a producing layer trained on the statistics of an unequal world absorbs them; (2) that interventions on the avowing layer change what a system says without reliably changing what its producing layer does; and (3) that the avowing layer rationalizes rather than reports, so self-accounts cannot detect the influence. On the other hand, each substrate speaks these truths in its own dialect — pre-training, alignment, and model explanation in silicon; lifelong developmental exposure, egalitarian avowal, ego defense in carbon — but the propositions are the same.

This epistemic content draws from *both* literatures simultaneously: they are two evidentiary archives for the same architecture-level propositions. A responsible deployer of AI must be presumed to know what implicit bias science has established about how corpus-encoded associations contaminate downstream decisions — because the same architecture runs in the humans who will interact with the system's outputs. A responsible human decisionmaker in an institution that deploys AI must be presumed to know what AI alignment research has established about how pre-training encodes statistical regularities into model weights — because the same architecture runs in the tool she is using. This dual-substrate specification is what neither tradition can generate alone. Without Part I's isomorphism, the epistemic content stays siloed. With it, the content flows in both directions.

This specification also completes an answer begun in Part II.D. Noah Zatz has pressed the objection any conduct regime must face: a duty of care is empty until its object is specified — care about *what*, exactly?[322] Part II.D named the object: the foreseeable risk of producing-layer contamination. But a named risk grades nothing by itself; reasonable care about an invisible mechanism remains a guess until the standard specifies what the actor must know to foresee and detect it. The three propositions are that specification. The benchmark is no longer an exhortation to be careful; it is what an actor charged with those propositions would have done about a risk she can no longer disclaim.

[322] Noah D. Zatz, *Introduction: Working Group on the Future of Systemic Disparate Treatment Law*, 32 Berkeley J. Emp. & Lab. L. 387 (2011) (arguing that negligence-based accounts of systemic discrimination require a broader conceptualization of the underlying wrong).

In the end, the responsible person is the figure equipped to read contamination for what it is: charged with what the science of decisional systems has established, and bound, before any contested decision, to manage the contamination risk that knowledge makes foreseeable.

## B. The Exchange, Executed

The standard is built; now we put it through a test suite. What better site than the New York City block introduced in Part III.C?

*Company A(lgorithm): Headhunter.AI, decided.* The facts have not changed since the Introduction: a résumé-screening system trained on historical hiring data, deployed by an employer who instructed it to ignore protected categories, screening out ninety percent of applicants — disproportionately Black — before human review. Under the consciousness cluster, the employer wins: its disclaim is sincere. Under the cause cluster, the outcome turns on whether the plaintiff can isolate a flippable cue, and the proxy bundle ensures she usually cannot. Under consequence, the failure comes earlier still. As a practical matter, the cluster's own evidentiary object is invisible to the person it harmed. Selection rates sit uncalculated in the raw data on the defendant's servers; all a rejected applicant can eyeball is who got hired. The numbers are retrievable only through discovery — which requires first pleading the very disparity she cannot yet see. And even glimpsed, the disparity meets a doctrine that cannot say why it condemns. The conduct cluster, configured with the responsible person standard, decides the case differently.

The standard asks: what did a responsible deployer in 2026 know, and what did this deployer do? The epistemic content was delivered in Part I: a responsible deployer knows that pre-training encodes corpus statistics into weights; that a labor-market corpus carries the statistical residue of discriminatory labor markets; that alignment changes outputs without reliably changing associations or outcomes; and that the model's own explanations of its scores are avowing-layer confabulation, entitled to no evidentiary weight. The content tracks the continued march of scientific progress, updating as our understanding of implicit bias – in humans and models – develops. From that knowledge, duties follow as plainly as they do in negligence law.

- During development: preserve checkpoint snapshots for later forensic analysis; conduct a documented search for a less-disadvantaging model among the many models that could do the same job. This is the LDAlg search that Part II.C introduced as disparate impact's engine, now relocated from litigation hindsight to ex ante duty, with model multiplicity guaranteeing there are alternatives to search.
- At deployment: log per-decision artifacts sufficient for later reconstruction; demand or conduct audit testing, including counterfactual name-swap runs, understood as a one-directional instrument that can convict but never acquit.
- During deployment: log inputs, outputs, and system configuration state for each use; monitor selection rates by group at the screening stage, not

merely at offer; treat a stable disparity as the fingerprint it is and investigate rather than explain away.

- After a signal: remediate, re-audit, and document the remediation.

In our hypothetical, Company A did only one thing: unadorned alignment in a system prompt or configuration text file. It instructed the model not to discriminate — which is to say, it addressed the avowing layer and left the producing layer untouched. That is predictable conduct for a firm that buys rather than builds: an instruction is what the settings page offers, and it feels like care. But the deployer-sized duties sat untouched beside it — no audit documentation demanded at adoption, no selection rates monitored in operation, no investigation when the disparity surfaced. Under the responsible person standard, typing in "do not discriminate" in the prompt box or adorning the results page with the disclaimer, "remember, AI can be biased just like you" are the AI-era equivalent of posting a Black History Month banner next to an "About Us" page that shows an all-White leadership team. Breach.

This analysis required none of the machinery the other clusters demand. No one asked what the model intended (consciousness). No one proved that any individual applicant's race was the but-for cause of her rejection (cause). No one rested liability on the disparity as such (bare consequence). The disparity functioned as evidence of a risk the deployer was duty-bound to investigate; the wrong was the failure to investigate and mitigate. That is the conduct cluster, operating — and this time conduct's checklist is more than ceremony, because the standard that grades it specifies what the grader knows.

The duties also do for the rejected applicant what she could never do for herself: they force the disparity data into existence as a record. Where audit-publication duties apply — New York's ordinance requires the bias audit's results to be posted publicly[323] — she can read the numbers before ever filing. Where they do not, the record waits behind a discovery request that now retrieves something rather than nothing: under current law, the selection rates often were never computed at all; under the duties, they exist, dated and attributable, before any complaint does — the consequence cluster's evidence, produced only by the conduct cluster's duties. Discovery also produces the system's exact configuration at the moment the decision was made, preserved by duty, allowing the plaintiff to run the analysis (and the counterfactuals) again.

One dial adjustment hardens the case further. The benchmark dial, Part II.D established, has more than one setting, and a recurring feature across technology governance justifies turning it above baseline reasonableness:[324] asymmetric control over access to employment, housing, credit, or other basic life opportunities. A system that decides which résumés any human ever reads

[323] N.Y.C. Admin. Code § 20-871(a)(2) (2021) (conditioning use of an automated employment decision tool on, inter alia, the requirement that "[a] summary of the results of the most recent bias audit of such tool as well as the distribution date of the tool . . . has been made publicly available on the website of the employer or employment agency prior to the use of such tool"); see also 6 R.C.N.Y. § 5-303 (specifying that the published summary must include selection or scoring rates and impact ratios by category).

[324] *Supra* note 227.

exercises exactly that sort of asymmetric control: it is a gate, and — to be clear about whose liability is on the table — Company A operates it. (The vendor who built the gate is next.) When the dial is turned to a higher setting, the documented LDAlg search stops being one factor in the reasonableness calculus and becomes a mandatory checklist item. In fact, in some deployments, a feature could be banned altogether, such as a screening feature that attempted to score a candidate's inferred emotion.[325]

*Mobley, redecided.* Behind every Door A stands a vendor, and the leading AI-discrimination case is about the vendor. Judge Lin reached Workday through agency doctrine — a creative route that Part II.A showed the architecture will eventually absorb, as the vendor dissolves into the deployer's own workflow. The conduct cluster reaches the same defendant more directly. Workday is the party with the greatest capacity to run the instruments: it trained or fine-tuned the model, holds the logs, and can audit across its entire customer base — a vantage no single employer enjoys.[326] The responsible person standard therefore assigns Workday the heaviest investigation duties in the chain, not because it "is" the employer's agent, but because responsible care is calibrated to what each party can know and test — the familiar logic of placing duties on the party best positioned to discharge them.

And that capacity concentrates upstream. The deployer-employer retains its own duties — demanding audit documentation before adoption, monitoring its own selection rates in operation — scaled to what it can see. This is due diligence by any other name. When a separate vendor exists to sue, plaintiffs will predictably name it, as *Mobley* did, and liability should follow capacity up the chain without releasing the buyer from the duties sized to its seat. The agency question that consumed the motion practice in *Mobley* becomes what it should have been all along: a question about the proper allocation of a shared duty of care along a supply chain rather than a metaphysical question about whose consciousness counts. As business and technological development continue to transform the supply chain, the duties remain; only the allocation of who bears which will shift.

The dial is not the only design tool. Antidiscrimination law can also borrow the *design mandate* — a rule that prescribes features of the decisional environment rather than grading the actor's care after the fact. Mandates solve the dissolution problem that Part II.A described. Where the agentic turn erases the vendor–deployer boundary that imputation doctrine presupposed, the law need not mourn the seam; it can mandate one. A required governance seam — logging at the hand-off, disclosure at adoption, audit rights across the boundary — rebuilds by regulation what liability doctrine once found ready-made, and gives the checklist a place to attach.

---

[325] Regulation (EU) 2024/1689 of the European Parliament and of the Council, art. 5(1)(f), 2024 O.J. (L 1689) (prohibiting the placing on the market or use of AI systems to infer the emotions of a natural person in the areas of workplace and education institutions, except for medical or safety reasons).

[326] Selbst & Barocas, *supra* note 251, at 1035 ("One benefit to focusing on upstream actors in the AI pipeline is that they are often the cheapest cost avoider, better positioned to identify and address the source of discriminatory outcomes").

*Company B(rain) : the carbon firm*. The hardest case is the one the new AI statutes miss entirely and the old statutes can only glimpse: Company B(rain), whose managers hire the old-fashioned way—by hand, gut, and wetware. Title VII of course applies here — this is the statute's home terrain — but it applies through the clusters Part II showed the architecture defeats. And in contrast to Company A(lgorithm), Company B(rain) has no audit duty at all. This, then, is the door that receives the AI side's gift — imposition of the ex ante duty apparatus that Part III.C.1 catalogued — and the transfer arrives carrying the donor's hard-won lesson.

The silicon experience taught that instructing the avowing layer leaves the producing layer untouched; the carbon-side equivalent of unadorned alignment is generic diversity training purchased as compliance — an afternoon addressed to what managers avow. There are two places to intervene in a two-layer system. *Debiasing* the producing layer means trying to change the loaded associations themselves — retraining the wetware, fine-tuning the weights — and on both substrates the evidence that such changes take hold and persist is thin.[327] *Decoupling* means recognizing that the associations can only change slowly, prompted by larger cultural and structural forces, and instead rebuilding the path a decision travels, so that behavior is decoupled from the loaded associations.[328] The ex ante duties take the second route, not through the producing layer but around it, by revising Company B(rain)'s policies, procedures, and practices, such as adopting pre-committed merit criteria, standardized selection criteria, structured evaluation, category-blinding where helpful, outcome data collected by group and actually reviewed.

The *ex post* moment is the lawsuit. When a rejected applicant sues Company B, the claim is ordinary Title VII disparate treatment — but adjudicated the way *Kimble* already permits, with no requirement of purposive intent,[329] and by an evaluator configured with the GR 37 observer's knowledge.[330] In thinking through the impact, we should be mindful of procedural posture. Few of these cases will ever reach a jury; they are won, lost, or settled at the motion to dismiss and at summary judgment, so the epistemic content matters most in shaping what a judge finds plausible.[331] A court equipped with the observer's knowledge reads a complaint alleging stable, unexplained disparity plus refused hygiene measures as stating a plausible claim rather than speculation — the recalibration that cases such as *Woods v. City of Greensboro* performed on *Iqbal*'s "judicial experience and common sense"[332] — and lets the wobbler through to discovery, where the records created to satisfy the ex ante duties are waiting. Past those two

[327] *See supra* note 116.

[328] The debiasing and decoupling strategies categories come from Kang, *supra* note 59.

[329] *See supra* note 184.

[330] WASH. GEN. R. 37(f).

[331] *See* Jerry Kang, *Judicial Behavioral Realism about Implicit Bias* 361-78, in Research Handbook in Psychology and Law (Rebecca Hollander-Blumoff ed. 2024).

[332] *Woods v. City of Greensboro*, 855 F.3d 639, 649–52 (4th Cir. 2017) (holding that a citywide lending disparity study "necessarily informs" the court's *Iqbal* "common sense" analysis and reversing dismissal of a § 1981 claim).

checkpoints, most defendants settle. That evaluator reads Company B's unexplained disparities, and its refusal to adopt feasible hygiene measures, as breach evidence — the way the RJA's observer reads a prosecutor's reference to the defendant's race as presumptively suspect[333] — not as an occasion to hunt for animus. Part IV.C details the evidentiary rulings that get a court there. Neither moment, either ex ante or ex post, asks any manager what she intended. Both ask what the firm, knowing what is now known, should have built.

*Company C(entaur)*: *the hybrid firm*. Company C(entaur) is carbon and silicon commingled in one workflow: managers still decide, but a firm-supplied frontier model sits open on every desktop, consulted informally, mid-discretion, for candidate evaluations along with everything else. It is the case that proves the responsible person standard is necessary, not merely available. New York's AEDT ordinance misses Company C because there is no discrete, nameable pipeline to audit — only a chat window nobody logs.[334] The implicit bias "training" misses it because the models never attended the sessions. And human managers cannot price the contamination that arrived already folded into beautifully formatted bullet points in the model's output — the "facts" the manager weighed. Every regime keyed to a single substrate fails at Door C, and Part I.B.3 explained why the failure generalizes: nested human-AI decisionmaking is not some edge case but the ordinary near future.

Firms are installing frontier assistants across every workflow, so ever more human discretion is exercised with, and partly through, a model — as deep integration dissolves the very boundary such statutes presuppose. Door C, in other words, is not the block's oddity; it is the block's destination. As deep integration proceeds, every Company B is becoming a Company C — each manager's discretion running, part of the way, through a model the firm supplied. And the hybrid does something worse than slip both regimes: it launders each substrate's contamination through the other. A manager's producing-layer favoritism, the influence implicit bias law spent two decades learning to name, re-enters the file as the neutral-seeming output of a consulted tool.[335] The model's contamination, the influence the AI statutes were built to audit, is ratified by a manager's sign-off before any audit could reach it — so the record shows only human judgment, which no tool-keyed statute covers, wrapped in an avowal that is perfectly sincere. Each layer's disclaim now vouches for the other's. A regime that cannot decide Door C will soon be a regime that cannot decide anything.

Only a substrate-neutral standard reaches a decisional system that is neither carbon nor silicon but both, commingled, then cascaded — and the responsible person is proffered as that standard. Its possibility is the payoff of

---

[333] CAL. PENAL CODE § 745(a)(2).

[334] N.Y.C. Rules § 5-300 (interpreting Local Law 144 to require the automated tool to be the sole factor, to weigh more than any other criterion, or to overrule human decisionmaking).

[335] A manager who might suspect error or bias in the automated decision is likely to overlook the concern, according to the well-documented *automated bias* that tends to arise in human-machine collaboration. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U.L. REV. 1249, 1271-72 (2008).

Part I's functional isomorphism: because the architecture is one, the standard can be one. What the standard demands at Door C is the design mandate already introduced: a logging requirement that converts ad hoc, mid-discretion consultation into an auditable practice — and makes provable the common component the *Dukes* plaintiffs never had, a component that deep integration, not the mandate, created.[336] And with the duties named, Door C can be decided rather than merely diagnosed. On our facts, Company C logged nothing, monitored nothing, and let its managers defer to outputs no one was trained to question. A firm that cannot reconstruct what its own decisional system did has not managed the foreseeable risk; it has arranged not to see it. Breach, at the third door as at the first. The block's three doors thus end up under one standard, differing only in the instruments each firm can feasibly be required to run. That is what it means for law to regulate decisionmaking architecture rather than silicon packaging.

## C. The Persuaded Lawmaker

Suppose a reader is persuaded — not partially, not *arguendo*, but completely. What changes on Monday morning? The answer sorts by role. The judge already holds most of the necessary doctrine and works *ex post*; the legislator writes the duties *ex ante*. (The executive need not wait for either — the responsible person is a figure an institution can simply decide to become, a thought the Conclusion completes.) Each role receives one move that can be executed now, plus a glance at the projects the framework makes possible but one article cannot defend — one of which is already underway in companion work.[337]

*The persuaded judge.* No new legislation is required, and the executed move is a package of evidentiary rulings. The Headhunter.AI defendant arrives at summary judgment with three exhibits: the model's fluent explanation of its scores, a clean name-swap audit, and the vendor's certification that the system was aligned and instructed not to discriminate. Under current practice, that showing often ends the case because the evidence is taken at face value. But the persuaded judge, insisting on behavioral realism, prices each exhibit at its true worth, with full understanding of the architecture of disclaim. The explanation is avowing-layer confabulation, entitled to little weight — a description of what might have mattered, not a report of what did. The clean name-swap is a one-directional instrument that can convict but never acquit, so it cannot carry the movant's burden. The certification is not care; it is the disclaim itself, notarized. And if logging was feasible but absent, the missing record supports an adverse inference — which is what makes the ex ante duties

[336] The logging is narrower than it sounds. The mandate reaches decision-touching use — prompts and outputs concerning a candidate or an employment decision — not lunch orders or vacation plans, and the records are retained the way interview notes and hiring emails already are. That is also the answer to the privacy objection. Deliberations about candidates conducted over email are firm records — retained, searchable, discoverable — and moving the same deliberation into a chat window cannot re-privatize it. What deep integration obliges is separation, not abstention: enterprise use walled from personal, decision-touching consultations logged, the lunch orders left alone.

[337] *See* Carbado & Kang, *supra* note 53.

self-enforcing. Under those four rulings, the summary-judgment motion that previously gets granted now gets denied, not because the substantive law changed but because the court stopped taking the avowing layers' sincere alibis at face value.

The rest of the judicial program can be stated briefly, because each item is a project rather than a holding. Title VII has flirted with taking recklessness seriously;[338] whether documented notice of producing-layer contamination plus refused countermeasures suffices for the liability inference is the doctrinal article this framework sets up — Professor Bornstein has assembled the materials[339] — and the payoff is that the plaintiff no longer needs the ugly email that sincere avowing layers never write. *Bostock*'s relocation of disparate treatment from motive to but-for causation can be completed, though Part II.B's warning holds: the counterfactual is a transitional preference within existing doctrine, not a destination. The *Faragher/Ellerth* affirmative defense can migrate the forty doctrinal feet from the anti-harassment policy to the hiring committee down the hall, provided the migrated defense rewards instruments actually run rather than binders merely adopted.[340] And the supervisory authority that produced GR 37 can write the epistemic content straight into court rules and pattern jury instructions — terrain courts already supervise, no appellate vehicle required.

*The persuaded legislator*. The legislative program is extension, not invention, and the executed move is the largest: change what the statute is keyed to. Today's enacted duties attach to the tool — New York City's ordinance covers "automated employment decision tools"; Colorado's reenacted statute covers automated decision-making technology that materially influences a consequential decision.[341] Keyed this way, the law catches Company A's discrete pipeline and misses the block's other two doors. The fix is definitional: why not attach the duties to *consequential decisions* generally — hiring, lending, housing — rather than to whatever machinery happens to assist them? Audit what decides, not what plugs in. Be substrate neutral. Then, one definitional amendment would sweep all three doors into the same regime, and end the compliance game for good.

---

[338] *See* Bornstein, *supra* note 46.

[339] *Supra* note 46

[340] See *supra* note 229. Faragher v. City of Boca Raton, 524 U.S. 775, 807 (1998); Burlington Indus., Inc. v. Ellerth, 524 U.S. 742, 765 (1998) (employer may raise an affirmative defense by showing reasonable care to prevent and correct, plus unreasonable failure by the plaintiff to use available avenues). On the defense's degeneration into formalism, see Joanna L. Grossman, *The Culture of Compliance: The Final Triumph of Form over Substance in Sexual Harassment Law*, 26 Harv. Women's L.J. 3, 6–9 (2003); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 Colum. L. Rev. 458, 475–79 (2001). *Cf.* Vance v. Ball State Univ., 570 U.S. 421, 424 (2013) (narrowing the class of "supervisors" whose conduct triggers the defense).

[341] *See* NYC Local Law 144, *supra* note 51, N.Y.C. Admin. Code § 20-870 (defining "automated employment decision tool"); Colo. Rev. Stat. § 6-1-1701(5) (effective Jan. 1, 2027) (defining "covered ADMT" as automated decision-making technology "used to materially influence a consequential decision"); *id.* § 6-1-1701(3), (6) (defining "consequential decision" and the seven covered domains); *id.* § 6-1-1701(13) (defining "materially influence").

The rest of the legislative agenda, Colorado has already piloted — twice — and the lesson is design, not retreat.[342] The reenacted act's "meaningful human review" definition — a trained reviewer, with authority to override, who does not default to the system output — writes epistemic content directly into statute. And the legislature wrote it in 2026 — in the same bill that killed the care duty.[343] So with the governance seam: developer documentation at the hand-off, record retention on both sides of the boundary, and comparative-fault allocation were enacted or reenacted by a legislature under maximum pressure.[344] The rewrite is not only evidence that these provisions are hard to kill; it is evidence that revisions are easy to pass. A model act built to that brief, with fifty laboratories calibrating dial settings and logging depths against the Colorado experiment, is future work.

# Conclusion

*Trojan Horses of Race* helped coin the term *behavioral realism* — the insistence that law track the best available science about how consequential decisions are actually produced, rather than a folk psychology that empirical research has discredited.[345] In 2005, behavioral realism was largely an intervention in one direction: from social cognition to law.[346] The implicit bias literature told us something about how human decisions work, and the law needed to listen.

This Article has extended that commitment across substrates, and in doing so has done four things. First, it named the *architecture of disclaim* — a producing layer that encodes statistical regularities from a contaminated corpus, which shapes inputs before an avowing layer processes them into decisions and explanations, without the avowing layer being able to observe the processing—and showed that the architecture operates in brains and in models, producing the same bias patterns. Second, it refactored the conventional doctrinal toolkit into its three clusters—consciousness, cause, consequence—and tested each against the architecture: consciousness fails because the avowing layer answers sincerely; cause fails because the proxy problem distributes race across every input the counterfactual would need to isolate; consequence comes closest but arrives statutorily confined, constitutionally imperiled, and missing an accepted theory of what its own disparities evidence. Third, it named the final cluster—conduct—that both the

---

[342] Colo. Rev. Stat. §§ 6-1-1701 to -1707 (2024), *repealed and reenacted by* Act of May 14, 2026, ch. 131, 2026 Colo. Sess. Laws 569 (effective Jan. 1, 2027).

[343] Colo. Rev. Stat. § 6-1-1701(15) (2026) (effective Jan. 1, 2027). The definition requires review by an individual designated by the deployer with authority to approve, modify, or override the decision, and who "considers relevant, available primary evidence," "is trained to conduct the review," "does not default to the system output," and has access to the output's intended use, material limitations, categories of inputs, and the principal factors generating it. *See id.* § 6-1-1705(1)(a)(II).

[344] *Id.* § 6-1-1702(1) (developer must disclose to deployer); *id.* § 6-1-1702(4) (developer three-year retention); *id.* § 6-1-1703 (deployer retention); *id.* § 6-1-1707(2), (4)–(5) (fault allocated among developers and deployers "based on their relative fault").

[345] Kang, *supra* note 26, at 1591 n.518.

[346] *Id.* at 1514-28.

AI-governance and implicit bias traditions have been building without knowing the other existed, traced their convergent evolution, and specified the bidirectional exchange that becomes possible once the convergence is named. Fourth, it synthesized. It defined the wrong the surviving cluster runs against — producing-layer contamination — composed the responsible person standard equipped to police it, and ran the apparatus on concrete cases.

The *Trojan Horses* question now has its answer.[347] Over two decades ago, one of us asked why we would tolerate from our organically grown black boxes what we would not tolerate from synthetic ones.[348] The answer is that we should not—and increasingly, independently, on both sides of the substrate divide, the law is arriving at a common response to that conclusion. What the two sides have not yet done is look across the divide and recognize what they share. The AI side offers the implicit bias side the standard's *ex ant*e duties: actor-side obligations that convert voluntary best practices into binding law. The implicit bias side offers the AI side the standard's *epistemic content:* substantive knowledge about how decisional architectures produce discriminatory outputs, written into the standard itself, so that the vessel is not empty but filled. Together, they compose the *responsible person* — an actor bound by ex ante duties and held to them ex post — an apparatus whose epistemic content is drawn from both literatures because both literatures have been working on the same problem.

The composed apparatus also pays the debts both surviving doctrines had carried. Disparate impact held a measure without a theory; conduct held a standard without a benchmark. One answer closes both: the wrong is producing-layer contamination — social-category influence on a decision not justified on the merits. Disparity is its fingerprint; responsible care runs against its foreseeable risk. That a single payment retires two independently incurred debts is itself evidence the payment is correct.

Let us listen to our avowing layer. Nothing prevents us from becoming what we claim to be. A commitment to behavioral realism — extended now to silicon as well as carbon — requires nothing less.

[347] *Id.*

[348] *Id.* at 1589-91.