

Cover Letter

To the Articles Selection Committee:

Three things happened in 2024 that make this Article urgent. The Supreme Court punted on algorithmic liability in *Moody v. NetChoice*—but Justice Barrett's concurrence opened a door, flagging that the constitutional analysis may differ entirely for algorithms that optimize for engagement rather than implementing editorial judgment. A circuit split deepened between *Anderson v. TikTok* and *M.P. v. Meta* over when algorithmic conduct becomes tortious manipulation. And MDL No. 3047 consolidated hundreds of social media harm claims that no existing tort theory can coherently resolve. Courts are asking a question that scholarship has not yet answered.

The Rage Machine: Algorithmic Behavioral Tort and the Case for Cognitive Integrity answers it. Where existing proposals target content and crash against Section 230 and the First Amendment—ABT targets *conduct*: the vulnerability detection, variable reinforcement schedules, and addiction architecture that platforms deploy independent of any particular speech. That conduct-content distinction is the Article's central doctrinal contribution, and it is the piece missing from the current litigation landscape.

What sets this Article apart from the growing body of platform accountability scholarship is that it is built to be *used*. The Manipulation-Vulnerability Matrix gives courts an administrable breach standard they can apply today. The black-letter tort elements are drafted in Restatement format. The Appendix contains a litigation playbook, legislative framework, and regulatory enforcement pathway. Judges presiding over MDL 3047 do not need another diagnosis of the problem. They need doctrine. This Article delivers it.

The Article is approximately 24,000 words with 217 footnotes. I would welcome the opportunity to discuss the submission further at your convenience.

Sincerely,

Mike Sapp
masapp@law.cwsl.edu | (760) 623-9210
J.D. Candidate, 2026
California Western School of Law

**The Rage Machine: Algorithmic Behavioral Tort
and the Case for Cognitive Integrity**

A Framework for Platform Accountability in the Digital Age

By: Mike Sapp

TABLE OF CONTENTS

ABSTRACT

INTRODUCTION: THE CRISIS OF COGNITIVE INTEGRITY

- A. The Human Cost of Intentional Design
- B. The Business Model: Why Harm Is Profitable
- C. Why Traditional Torts Fail
- D. The Causation Debate and ABT's Response

BLACK-LETTER ALGORITHMIC BEHAVIORAL TORT

- § 1. The Algorithmic Behavioral Tort
- § 2. The Manipulation-Vulnerability Matrix (Rule of Decision)

PART I: THE ALGORITHMIC BEHAVIORAL TORT FRAMEWORK

- A. Theoretical Foundations: Synthesis from Proven Principles
- B. The Protected Interest: Cognitive Integrity
- C. The Duty: The Algorithmic Fiduciary Relationship
- D. The Manipulation-Vulnerability Matrix: ABT's Central Analytical Tool

E. Temporal Dimension: Cumulative Harm as Continuing Tort

F. Evidence and Proof Map

PART II: OVERCOMING THE MAJOR LEGAL BARRIERS

A. Section 230: The Conduct-Content Distinction

B. First Amendment: Content-Neutral Regulation of Conduct

C. Anticipated Defenses

PART III: PLATFORM-SPECIFIC APPLICATIONS

A. Facebook/Meta: The Anger Amplification Engine

B. TikTok: The Rabbit-Hole Algorithm

C. Instagram: The Body Image Distortion Field

D. YouTube: The Radicalization Pipeline

PART IV: INTERNATIONAL VALIDATION

A. The European Union and United Kingdom

B. Australia: The World's First Social Media Ban for Minors

C. Converging Global Principles

CONCLUSION: COGNITIVE INTEGRITY AS LEGAL RIGHT

APPENDIX: IMPLEMENTATION PATHWAYS

A. Litigation: Playbook for Plaintiffs' Counsel

B. Legislative Framework

C. Regulatory Enforcement

ABSTRACT

Between 2011 and 2021, the share of U.S. high school students reporting persistent sadness or hopelessness rose from 28% to 42%.¹ Among female students, the share reporting suicidal ideation rose from 19% to 30%.² Platform internal research documents why: algorithms systematically amplify harmful content because it maximizes engagement. Facebook weights an “angry” reaction five times more heavily than a “like.”³ Instagram’s own studies found the platform worsens body image issues for one in three teen girls—then buried the findings.⁴ This is not a bug. It is the business model.

American tort law has been powerless to respond. Products liability founders on classification. Negligence crashes against Section 230 and algorithmic opacity. Consumer

¹See CENTERS FOR DISEASE CONTROL & PREVENTION, YOUTH RISK BEHAVIOR SURVEY (2021) (showing persistent sadness/hopelessness increase from 28% to 42%); see also Jean M. Twenge et al., *Worldwide Increases in Adolescent Loneliness*, 119 J. ADOLESCENCE 214 (2021) (documenting global increase in adolescent loneliness beginning 2012—the inflection point for smartphone adoption); Katherine M. Keyes et al., *Cohort Effects Explain Declines in Subjective Well-being Among Adolescents*, 16 J. ABNORMAL PSYCHOLOGY 398 (2019) (finding generational effects rather than age effects explain declining adolescent well-being).

²CDC, *supra* note 1 (female students reporting suicidal ideation increased from 19% to 30%); see also SALLY C. CURTIN & MELONIE HERON, DEATH RATES DUE TO SUICIDE AND HOMICIDE AMONG PERSONS AGED 10–24: UNITED STATES, 2000–2017, NCHS DATA BRIEF NO. 352 (2019) (documenting 56% increase in suicide rate among ages 10–24 from 2007 to 2017); THE LANCET, *Social Media and Adolescent Mental Health* (Editorial), 402 LANCET 1637 (2023) (calling for urgent research into causal mechanisms and regulatory responses); see, e.g., U.S. DEP’T OF HEALTH & HUMAN SERVS., OFFICE OF THE SURGEON GENERAL, SOCIAL MEDIA AND YOUTH MENTAL HEALTH: AN ADVISORY (2023) [hereinafter Surgeon General’s Advisory]; Jonathan Haidt & Jean M. Twenge, *Social Media and the Rise of Mental Illness and Suicide in Teen Girls and Boys*, 31 PREV. RESEARCHERS (Dec. 2023).

³See Jeff Horwitz, *The Facebook Files: A Wall Street Journal Investigation*, WALL ST. J. (Sept.–Oct. 2021) (series documenting Facebook’s internal research on platform harms); Karen Hao, *How Facebook Got Addicted to Spreading Misinformation*, MIT TECH. REV. (Mar. 11, 2021) (detailing Facebook’s 2017 decision to weight “angry” reactions five times more heavily than “likes” in its News Feed algorithm, and internal warnings that this weighting was “increasingly divisive”).

⁴Georgia Wells, Jeff Horwitz & Deepa Seetharaman, *Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show*, WALL ST. J. (Sept. 14, 2021), <https://www.congress.gov/117/meeting/house/114054/documents/HHRG-117-IF02-20210922-SD003.pdf> (last visited Dec. 23, 2025).

protection addresses deception, not manipulation.⁵ The doctrinal failure reflects the market failure: platforms profit from exploiting human psychological vulnerabilities, and existing law provides no remedy.

This Article proposes one. The Algorithmic Behavioral Tort (ABT) recognizes cognitive integrity—the right to mental self-determination—as a protected legal interest, completing the arc from battery (bodily integrity) through privacy (informational autonomy) to the mind itself.⁶ ABT imposes fiduciary duties on platforms exercising discretionary control over users’ information environments, and operationalizes breach through the Manipulation-Vulnerability Matrix—a practical tool courts can apply today to distinguish tortious exploitation from permissible persuasion.

The rule of decision: Liability attaches when (1) a platform owes a duty to respect cognitive integrity; (2) it breaches that duty through systematic behavioral conditioning, assessed via the Matrix; and (3) that conduct is a substantial factor in causing cognizable harm. ABT survives Section 230 by targeting platform *conduct*—vulnerability detection and addiction mechanics—rather than third-party *content*. It survives the First Amendment because Justice Barrett’s concurrence in *Moody v. NetChoice* flagged that the constitutional analysis may differ for algorithms that automatically optimize for engagement rather than implementing human editorial judgment.⁷

⁵FED. TRADE COMM’N, BRINGING DARK PATTERNS TO LIGHT: AN FTC STAFF REPORT (Sept. 2022), <https://www.ftc.gov/reports/bringing-dark-patterns-light-ftc-staff-report> (last visited Jan. 5, 2026).

⁶The concept draws from cognitive liberty philosophy; see Neil M. Richards, Intellectual Privacy, 87 TEX. L. REV. 387 (2008) [hereinafter Richards, Intellectual Privacy].

⁷*Moody v. NetChoice, LLC*, 144 S. Ct. 2383, 2408 (2024) (Barrett, J., concurring) (“But what if a platform’s algorithm just presents automatically to each user whatever the algorithm thinks the user will like—e.g., content similar to posts with which the user previously engaged? The First Amendment implications of the laws might be different for that kind of algorithm.”).

Four contributions: First, cognitive integrity as a cognizable tort interest grounded in privacy, informed consent, and constitutional freedom of thought. Second, black-letter ABT elements adapted to algorithmic manipulation. Third, the Manipulation-Vulnerability Matrix as an administrable breach standard. Fourth, a conduct-content distinction that navigates Section 230 and First Amendment barriers.

The analysis proceeds in four Parts. Part I develops ABT’s framework. Part II addresses Section 230 and First Amendment constraints. Part III applies the Matrix to Facebook, TikTok, Instagram, and YouTube. Part IV draws validation from international regulatory convergence—the EU Digital Services Act, UK Online Safety Act, and Australia’s 2024 social media ban. An Appendix provides implementation pathways for litigation, legislation, and regulation.⁸

The human mind is not a commodity to be mined.

INTRODUCTION: THE CRISIS OF COGNITIVE INTEGRITY

A. The Human Cost of Intentional Design

The machine does not hate you. It is not trying to make you angry, depressed, or radicalized. It is optimized for one mechanical metric: *engagement*. When ranking and recommendation systems are trained and continuously tuned against that metric, they predictably prioritize whatever keeps users scrolling—including high-arousal emotions that bypass deliberation. As Facebook’s own engineers discovered, an “angry” reaction was worth five times

⁸*See, e.g.*, EUROPEAN UNION, Digital Services Act, Regulation (EU) 2022/2065 (2022) [hereinafter DSA]; Online Safety Act 2023, c. 50 (U.K.) [hereinafter OSA].

the weight of a “like” in the algorithm’s ranking logic—a design choice that systematically amplified the most divisive content because it was the most profitable.⁹

This is the “Rage Machine”—not a sentient villain, but a hyper-optimized feedback loop of behavioral conditioning. It is an engine that converts human attention into revenue by systematically exploiting our cognitive vulnerabilities.¹⁰ And for the last decade, American tort law has been powerless to stop it.

The legal system’s failure is not due to a lack of effort. Plaintiffs have flooded the courts with claims against social media giants, alleging everything from negligent design to product liability.¹¹ Yet, these cases almost invariably crash against two formidable barriers: Section 230 of the Communications Decency Act,¹² which immunizes platforms from liability for third-party content, and the First Amendment,¹³ which protects the platforms’ editorial discretion. Courts have rightly held that we cannot sue Facebook simply because someone posted a lie, nor can we sue TikTok just because its algorithm recommended a video we found offensive.¹⁴

But this analysis misses the point. The harm of the Rage Machine is not in the *content* it serves, but in the *conduct* it performs.¹⁵ It is not about the video that plays, but about the

⁹Horwitz & Hao, *supra* note 3.

¹⁰*See, e.g.*, SHOSHANA ZUBOFF, *THE AGE OF SURVEILLANCE CAPITALISM* 94 (2019) [hereinafter Zuboff, *Surveillance Capitalism*] (discussing the translation of user experience into monetizable behavioral data like “dwell time”).

¹¹*See In re Social Media Adolescent Addiction/Personal Injury Products Liability Litig.*, MDL No. 3047 (N.D. Cal. Oct. 24, 2023) (split between judges on product classification).

¹²47 U.S.C. § 230(c)(1) (“No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.”).

¹³U.S. CONST. AMEND. I (“Congress shall make no law ... abridging the freedom of speech, or of the press”).

¹⁴*Zeran v. Am. Online, Inc.*, 129 F.3d 327, 330 (4th Cir. 1997) (holding that Section 230 “creates a federal immunity to any cause of action that would make service providers liable for information originating with a third-party user of the service”).

¹⁵Courts have increasingly recognized a distinction between platform liability for third-party *content* (immunized under Section 230) and liability for the platform’s own *conduct* in designing features and systems. *See Fair Housing Council of San Fernando Valley v. Roommates.com, LLC*, 521 F.3d 1157, 1162–63 (9th Cir. 2008) (distinguishing platform’s own creation of discriminatory matching system from publication of third-party content); *Lemmon v.*

algorithmic architecture that uses behavioral telemetry to infer what reliably provokes a user and then delivers it at machine-gun velocity until the user's resistance erodes. It is about the difference between a librarian recommending a book and a casino designing a slot machine to ensure the player never leaves the chair.¹⁶

The solution lies in recognizing a new category of harm: the violation of Cognitive Integrity.¹⁷ Just as Warren and Brandeis's 1890 article demanded that law recognize a "Right to Privacy" against new photographic technology,¹⁸ the algorithmic "Rage Machine" demands the next evolutionary step. The law protects our bodies from battery and our property from trespass; it must now recognize a protected interest in the sanctity of our own mental processes—a right to be free from surreptitious, automated behavioral modification that overrides our capacity for rational choice.

To justify this right, we propose a new doctrinal framework: the Algorithmic Behavioral Tort (ABT). ABT distinguishes itself from prior theories by focusing strictly on the *interaction mechanics*—variable reward schedules, infinite scroll, notification design, and hyper-personalized targeting loops—rather than the expressive content of the feed.¹⁹ By isolating the

Snap, Inc., 995 F.3d 1085, 1091–92 (9th Cir. 2021) (platform design choices constitute conduct, not content; "Snap's alleged knowledge and design... go to its own conduct, not to monitoring or screening third-party content"); Anderson v. TikTok, Inc., 2024 WL 3879203 (3d Cir. Aug. 27, 2024) (algorithmic recommendation constitutes platform's own conduct, not publication of user-generated content).

¹⁶See Schüll, *Addiction by Design*, *infra* note 19, at 52–78 (documenting how slot machine designs exploit psychological vulnerabilities through variable reward schedules, sensory feedback loops, and "zone" states that override conscious decision-making—the same mechanics platforms now deploy digitally).

¹⁷See Richards, *Intellectual Privacy*, *supra* note 6, at 421–25; *see also* Jan Christoph Bubltz & Reinhard Merkel, *Crimes Against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination*, 8 *Crim. L. & Phil.* 51 (2014); Marcello Ienca & Roberto Andorno, *Towards New Human Rights in the Age of Neuroscience and Neurotechnology*, 13 *LIFE SCI., SOC'Y & POL'Y* 1 (2017).

¹⁸Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 *HARV. L. REV.* 193, 205 (1890) (arguing that existing legal doctrines protecting the person and property must evolve to protect "the right to be let alone" against "modern devices" like "instantaneous photographs").

¹⁹See *generally* NATASHA DOW SCHÜLL, *ADDICTION BY DESIGN: MACHINE GAMBLING IN LAS VEGAS* (2012) [hereinafter Schüll, *Addiction by Design*]; ADAM ALTER, *IRRESISTIBLE: THE RISE OF ADDICTIVE TECHNOLOGY AND THE BUSINESS OF KEEPING US HOOKED* (2017) [hereinafter Alter, *Irresistible*].

conduct of the platform’s algorithmic system (the “how”) from the *content* it delivers (the “what”), ABT threads the needle of Section 230 and the First Amendment, targeting the platform’s own behavior rather than the speech of its users.²⁰

The stakes could not be higher. As whistleblower Frances Haugen testified, these platforms know their algorithms are “toxic,” yet they refuse to change them because “they put their astronomical profits before people.”²¹ The cost of this inaction is measured in a mental health crisis of epidemic proportions,²² the erosion of our shared reality,²³ and the systematic dismantling of the cognitive autonomy required for democratic citizenship.²⁴ The scholarship on algorithmic harms has grown exponentially, yet the gap between diagnosis and remedy remains vast.²⁵

The law is not feckless in the face of new technology. It has adapted before to the steam engine, the automobile, and mass manufacturing.²⁶ It can adapt again. But to do so, we must stop

²⁰See *Fair Housing Council of San Fernando Valley v. Roommates.com, LLC*, 521 F.3d 1157, 1162–63 (9th Cir. 2008); *Lemmon v. Snap, Inc.*, 995 F.3d 1085, 1091–92 (9th Cir. 2021); *Anderson v. TikTok, Inc.*, 2024 WL 3879203 (3d Cir. Aug. 27, 2024).

²¹*Protecting Kids Online: Testimony from a Facebook Whistleblower: Hearing Before the Subcomm. on Consumer Prot., Prod. Safety, and Data Sec. of the S. Comm. on Com., Sci., and Transp.*, 117th Cong. (2021) (statement of Frances Haugen).

²²*Supra* note 1 (documenting adolescent mental health crisis).

²³See, e.g., DAN M. KAHAN et al., *The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks*, 36 NAT. CLIMATE CHANGE 322 (2017).

²⁴See, e.g., SANDER VAN DER LINDEN et al., *Inoculating Against Misinformation*, 377 Science 6602 (2022); DAN M. KAHAN, *The Expressive Overdetermination of Gun-Control*, 1 J.L. & PUB. POL’Y 17 (2018) (discussing identity-protective cognition and moral-emotional reasoning).

²⁵See SAFIYA UMOJA NOBLE, *ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM* 1–14 (2018) (documenting discriminatory effects of algorithmic systems); VIRGINIA EUBANKS, *AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR* 3–11 (2018) (analyzing algorithmic harms to vulnerable populations).

²⁶See, e.g., *MacPherson v. Buick Motor Co.*, 111 N.E. 1050 (N.Y. 1916) (abandoning privity requirement for product liability); see also G. EDWARD WHITE, *TORT LAW IN AMERICA: AN INTELLECTUAL HISTORY* 63–102 (expanded ed. 2003) (documenting tort law’s adaptation to industrial-age technologies).

treating algorithmic curation systems as passive publishers and start treating them as what they are: products designed to alter human behavior.²⁷ It is time to break the machine.

To be clear at the outset: ABT does not target viewpoint or content. A platform may curate any ideological mix using non-exploitative methods. ABT targets only the manipulation mechanics—the behavioral conditioning architecture that operates independently of the ideas conveyed.

B. The Business Model: Why Harm Is Profitable

The platform’s system does not malfunction when it harms children. It functions exactly as designed and tuned. Harm *is* the business model.²⁸ Understanding these incentives is essential to diagnosing both why existing law fails and why new legal tools are necessary.

1. The Outrage Economy

Social media business models sell something deceptively simple: user attention. More precisely, they sell access to user neurology. In this “attention economy,”²⁹ platforms maximize engagement through a discovered truth: outrage drives engagement more effectively than almost anything else.³⁰

²⁷Haugen, *supra* note 21.

²⁸See, e.g., ALLAN M. BRANDT, *THE CIGARETTE CENTURY: THE RISE, FALL, AND DEADLY PERSISTENCE OF THE PRODUCT THAT DEFINED AMERICA* (2007) (providing a comprehensive history of the tobacco industry’s business model of privatizing profit while socializing harm).

²⁹The foundational account of the attention economy is TIM WU, *THE ATTENTION MERCHANTS: THE EPIC SCRAMBLE TO GET INSIDE OUR HEADS* (2016). See also JAMES WILLIAMS, *STAND OUT OF OUR LIGHT: FREEDOM AND RESISTANCE IN THE ATTENTION ECONOMY* (2018) (arguing that the attention economy constitutes a systemic threat to human autonomy); Tristan Harris, *How Technology is Hijacking Your Mind*, MEDIUM (May 18, 2016) (former Google design ethicist describing persuasive technology techniques).

³⁰See William J. Brady et al., *Emotion Shapes the Diffusion of Moralized Content in Social Networks*, 114 PROC. NAT’L ACAD. SCI. 7313 (2017) (documenting that moral-emotional language increases message diffusion on social media by 20% per moral-emotional word); Molly J. Crockett, *Moral Outrage in the Digital Age*, 1 NATURE HUM. BEHAV. 769 (2017) (analyzing how digital platforms amplify outrage expression).

The mechanism exploits innate psychological biases. Engagement-optimized ranking systems amplify content triggering the PRIME framework—information that is Prestigious, from an Ingroup source, and Moral-Emotional.³¹ Regardless of accuracy, such content generates extreme, polarizing responses.

The 5× anger weighting described above was no accident—it was a deliberate 2017 redesign.³² As whistleblower Frances Haugen testified: “Anger and hate is the easiest way to grow on Facebook.”³³ The rage machine is literal: an engagement-optimized ranking system tuned to amplify fury because fury sells.

Designs marketed as “delight” to users are “dwell time” on earnings calls.³⁴

2. Addiction by Design: The Neurological Engine

Platform mechanics deliberately foster addiction using casino slot machine principles: endless scroll (no stopping point), instant feedback (likes reinforce posting), variable reinforcement (unpredictable notifications drive compulsion), and personalized algorithms (tailored to maximize individual engagement).³⁵ The still-developing adolescent brain is particularly vulnerable—the prefrontal cortex responsible for impulse control does not fully mature until the mid-twenties.³⁶ A 2017 leaked Facebook document revealed the company

³¹van der Linden et al., *supra* note 24; Kahan, *supra* note 24.

³²Horwitz & Hao, *supra* note 3 (detailing Facebook’s 5× weighting of “angry” reactions).

³³Haugen, *supra* note 21 (testifying that “anger and hate is the easiest way to grow on Facebook”).

³⁴Zuboff, *supra* note 10, at 94 (discussing translation of user experience into monetizable “dwell time”).

³⁵Schüll, *supra* note 19; Alter, *supra* note 19.

³⁶See Candice L. Odgers & Michaeline R. Jensen, *Annual Research Review: Adolescent Mental Health in the Digital Age*, 26 J. CHILD PSYCHOL. & PSYCHIATRY 336 (2020); see Laurence Steinberg, *A Social Neuroscience Perspective on Adolescent Risk-Taking*, 28 DEVELOPMENTAL REV. 78 (2008) (documenting delayed prefrontal cortex maturation relative to reward-system development); B.J. Casey et al., *The Adolescent Brain*, 28 DEVELOPMENTAL SCI. 111 (2008) (explaining neurobiological basis for adolescent vulnerability to reward-based manipulation); SARAH-JAYNE BLAKEMORE, *INVENTING OURSELVES: THE SECRET LIFE OF THE TEENAGE BRAIN*

pitched advertisers on its ability to identify in real-time when teenage users felt “insecure,” “worthless,” or “stressed.”³⁷

3. The Profit-Harm Alignment Problem

This economic structure creates profound misalignment between platform and user interests. The business model mirrors the twentieth-century tobacco industry: privatizing profit while socializing harm.³⁸ Social costs—therapy bills, school disruption, social fabric degradation—fall on families and taxpayers. Financial benefits accrue to platforms.

The dynamic parallels the notorious *Grimshaw v. Ford Motor Co.* case, where internal memos revealed Ford’s cost-benefit analysis valuing human lives less than recall costs—concluding it was cheaper to pay damages than fix a deadly fuel tank defect.³⁹ Platform internal research mirrors this calculus: when studies show safer algorithms would reduce engagement (and revenue), companies engage in the same grim calculation, accepting greater harm for greater profit. As Haugen testified: “If they change the algorithm to be safer, people will spend less time on the site, they’ll click on less ads, and Facebook will make less money.”⁴⁰

(2018) (accessible synthesis of adolescent neuroscience research); see JAAK PANKSEPP, *AFFECTIVE NEUROSCIENCE: THE FOUNDATIONS OF HUMAN AND ANIMAL EMOTIONS* (2005).

³⁷Sam Levin, *Facebook told advertisers it can identify teens feeling ‘insecure’ and ‘worthless’*, THE GUARDIAN (May 1, 2017), <https://www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens> (last visited Dec. 23, 2025); see also Darren Davidson, *Facebook targets ‘insecure’ young people*, THE AUSTRALIAN (May 1, 2017), <https://www.theaustralian.com.au/business/media/facebook-targets-insecure-young-people-to-sell-ads/news-story/a89949ad016eee7d7a61c3c30c909fa6> (last visited Dec. 23, 2025).

³⁸Brandt, *supra* note 28.

³⁹*Grimshaw v. Ford Motor Co.*, 119 Cal. App. 3d 757, 774–75 (1981) (describing Ford’s internal memo valuing human lives less than recall costs in the Pinto fuel tank litigation).

⁴⁰*Testimony of Frances Haugen before the U.S. Senate Commerce Subcomm. on Consumer Protection, Product Safety, and Data Security*, 117th Cong. (Oct. 5, 2021) (statement of Frances Haugen); see also SHEERA FRENKEL & CECILIA KANG, *AN UGLY TRUTH: INSIDE FACEBOOK’S BATTLE FOR DOMINATION* 302–05 (2021) (documenting Haugen’s disclosures).

This market failure cannot self-correct. Publicly traded companies often read fiduciary duties as pressure to maximize shareholder value, even at the cost of safer design—punishing any platform that unilaterally incurs safety costs.⁴¹

C. Why Traditional Torts Fail (The Doctrinal Crisis)

Despite emerging litigation and legislative momentum, the doctrinal foundation for algorithmic accountability in the United States remains fragile. The pattern reflects not minor inadequacies but fundamental structural mismatch:

Products Liability is mired in a classification crisis, with sophisticated courts unable to agree whether a social media platform is a “product.” Some judges dismiss strict liability claims, reasoning platforms are intangible services; others adopt feature-by-feature analysis.⁴² The struggle to classify information-based services has plagued courts for decades, with no coherent doctrinal resolution.⁴³

Negligence is paralyzed by five interlocking barriers: (1) Section 230 immunity, (2) algorithmic opacity (“black box” problem), (3) causation impossibility (but-for test collapses for multi-causal mental health harms), (4) the foreseeability paradox (platforms tout omniscience to

⁴¹See, e.g., *Dodge v. Ford Motor Co.*, 170 N.W. 668, 684 (Mich. 1919) (famously stating that a business corporation is organized primarily for the profit of the stockholders); see also Leo E. Strine, Jr., *The Dangers of Denial: The Need for a Clear-Eyed Understanding of the Power and Accountability Structure of the American Corporation*, 60 WAKE FOREST L. REV. 761 (2015) (arguing that corporate law requires directors to prioritize stockholder welfare).

⁴²See *In re Social Media Adolescent Addiction/Personal Injury Products Liability Litig.*, MDL No. 3047 (N.D. Cal. Oct. 24, 2023) (split between judges on product classification).

⁴³See *Winter v. G.P. Putnam’s Sons*, 938 F.2d 1033, 1034–35 (9th Cir. 1991) (refusing to apply strict products liability to book containing erroneous mushroom identification information); RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 19 cmt. d (1998) (noting difficulty of classifying information products). The product-service distinction has proven increasingly incoherent as digital services incorporate both informational and physical elements.

advertisers, claim helplessness in court), and (5) the standard of care vacuum (harmful industry custom is the only guide).⁴⁴

Public Nuisance faces justified resistance to expansion into what courts view as a boundless “tort of everything.”⁴⁵

Consumer Protection Law, while potent for regulators, targets deception rather than manipulation. As the FTC recognized, user interfaces can be designed to “obscure, subvert, or impair consumer choice” without any false statement.⁴⁶ A user is not deceived into developing an eating disorder; she is manipulated.

The doctrinal message is clear: the solution is not to stretch old categories, but to build a new one.

D. The Causation Debate and ABT’s Response

Critics demand scientific certainty before legal action. This demand would have paralyzed tobacco litigation in 1960. The law does not wait for certainty.

The skeptics’ case: Effect sizes appear small (under one percent of well-being variance);⁴⁷ historical moral panics counsel caution;⁴⁸ most studies are correlational.⁴⁹ But contemporary research identifies ABT’s actual target: platforms disproportionately amplify

⁴⁴See, e.g., 47 U.S.C. § 230(c)(1) (Section 230); Kate Crawford & Brent Mittelstadt, *The Malicious Use of Artificial Intelligence*, 3 J. ARTIF. INTELL. RES. 1 (2020) (black box problem).

⁴⁵See, e.g., *School District of Philadelphia v. Meta Platforms, Inc.*, 2023 WL 6821906 (E.D. Pa. Oct. 18, 2023).

⁴⁶FED. TRADE COMM’N, *BRINGING DARK PATTERNS TO LIGHT: AN FTC STAFF REPORT* (Sept. 2022), <https://www.ftc.gov/reports/bringing-dark-patterns-light-ftc-staff-report> (last visited Jan. 31, 2026).

⁴⁷Amy Orben & ANDREW K. PRZYBYLSKI, *The Association Between Adolescent Well-Being and Digital Technology Use*, 3 NAT. HUM. BEHAV. 173–182 (2019).

⁴⁸Odgers & Jensen, *supra* note 36.

⁴⁹Brown et al., *Rethinking Social Media and Mental Health*, 174 COMPUTERS IN HUMAN BEHAVIOR 108815 (2026), <https://doi.org/10.1016/j.chb.2025.108815> (last visited Jan. 31, 2026).

harms for *vulnerable* users—those with anxiety, depression, or developmental susceptibility.⁵⁰

The vulnerability-mediated model explains why population-wide averages miss the point.

Three frameworks justify action despite uncertainty:

The Tobacco Precedent. In the 1950s-1960s, tobacco exploited uncertainty—correlation was clear, mechanism contested. Courts did not wait. Bradford Hill criteria guided analysis; burden shifted once plaintiffs showed association plus exposure.⁵¹ Platform internal research provides stronger evidence than tobacco had: Facebook’s “emotional contagion” experiment proved manipulation of mental states;⁵² Instagram research directly quantified “1 in 3 teen girls” harm. These are controlled tests, not correlations.

The Precautionary Principle. Environmental law holds that threats of serious or irreversible damage justify action despite uncertainty.⁵³ Youth mental health satisfies this standard: serious developmental harm, vulnerable population, cost-effective remedies available.

The Risk-Factor Model. Smoking does not cause cancer in all smokers, yet justifies regulation. Modest risk increases (20-30%) across ninety-five percent of teens means millions affected.⁵⁴

The evidence bundle: Temporal correspondence (including the rise in reported persistent sadness/hopelessness from 28% to 42% (2011–2021) and the rise in reported suicidal ideation

⁵⁰Brown et al., *supra* note 49 (demonstrating vulnerability-mediated harm model).

⁵¹See, e.g., TROYEN A. BRENNAN, *Causal Chains and Statistical Links: The Role of Scientific Uncertainty in Hazardous-Substance Litigation*, 73 CORNELL L. REV. 469 (1988); see Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 PROC. ROYAL SOC’Y MED. 295 (1965).

⁵²Adam D. I. Kramer, Jamie E. Guillory & Jeffrey T. Hancock, *Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks*, 111 PROC. NAT’L ACAD. SCI. U.S.A. 8788 (2014).

⁵³Rio Declaration on Environment and Development, Principle 15 (1992).

⁵⁴See, e.g., P.N. Lee et al., *The Proportion of Never-Smokers Among Patients with Lung Cancer*, 129 BRIT. J. CANCER 1725 (2023) (finding that a significant proportion of lung cancer patients are never-smokers, illustrating the probabilistic link); see also CENTERS FOR DISEASE CONTROL & PREVENTION, *Health Effects of Cigarette Smoking* (last reviewed Nov. 29, 2022).

among female students from 19% to 30%, with cross-national replication);⁵⁵ dose-response (heavier use correlates with worse outcomes);⁵⁶ quasi-experimental evidence (Facebook rollout increased treatment utilization; outages improved well-being; detox trials show improvement).⁵⁷

Critics demanding scientific certainty before legal action misunderstand both the science and the law. The effect-size critique—that social media explains under one percent of well-being variance—commits a fundamental error: it measures population means when ABT targets population tails.⁵⁸ Contemporary research demonstrates that platforms disproportionately harm the most vulnerable users—those with pre-existing anxiety, depression, or developmental susceptibility.⁵⁹ For a thirteen-year-old girl with emerging body image concerns, Instagram’s recommendation system doesn’t explain one percent of her distress; the algorithm identified her

⁵⁵See JONATHAN HAIDT & JEAN M. TWENGE, *Social Media and the Rise of Mental Illness and Suicide in Teen Girls and Boys*, 31 PREV. RESEARCHERS (Dec. 2023) (documenting similar trends in mental health and social media use across multiple English-speaking countries); see JONATHAN HAIDT, *THE ANXIOUS GENERATION: HOW THE GREAT REWIRING OF CHILDHOOD IS CAUSING AN EPIDEMIC OF MENTAL ILLNESS* 24–35 (2024) (systematically analyzing and dismissing alternative explanations for the post-2010 decline in adolescent mental health, including economic cycles, academic pressure, and changes in diagnostic practices).

⁵⁶Ethan Kross et al., *Facebook Use Predicts Declines in Subjective Well-Being in Young Adults*, 8 PLoS ONE e69841 (2013); see Austin Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 PROC. ROYAL SOC’Y MED. 295 (1965) (articulating nine criteria for inferring causation from epidemiological data). Hill’s criteria—strength, consistency, specificity, temporality, biological gradient (dose-response), plausibility, coherence, experiment, and analogy—remain the gold standard for causal inference in public health and toxic tort contexts. Social media evidence increasingly satisfies multiple Hill criteria.

⁵⁷Luca Braghieri, Ro’ee Levy & Alexey Makarin, *Social Media and Mental Health*, 112 AM. ECON. REV. 3660 (2022); see, e.g., Dhruv Tyagi, *Facebook’s Great Outage of 2021 Was Great for Our Mental Health*, FORBES (Oct. 11, 2021), <https://www.forbes.com/sites/dhruvtyagi/2021/10/11/facebooks-great-outage-of-2021-was-great-for-our-mental-health/> (last visited Dec. 23, 2025) (discussing self-reported improvements in well-being during the outage); Hunt Allcott, Luca Braghieri, Sarah Eichmeyer & Matthew Gentzkow, *The Welfare Effects of Social Media*, 110 AM. ECON. REV. 629–676 (2020); see Kramer et al., *supra* note 52.

⁵⁸*Compare* Orben & Przybylski, *supra* note 47 (finding small effect sizes in correlational data), *with* Jonathan Haidt & Jean M. Twenge, *Social Media and Mental Health: A Collaborative Review* (working paper 2024), <https://jonathanhaidt.com/social-media> (last visited Jan. 7, 2026) (responding that Orben-Przybylski studies (1) use controls that are themselves affected by social media, (2) fail to distinguish passive consumption from active use, and (3) measure population averages rather than vulnerable subpopulations). The vulnerability-mediated model explains the discrepancy: most users experience minimal harm, but algorithms concentrate severe harm on the most susceptible—precisely the population ABT protects.

⁵⁹Brown et al., *supra* note 49 (demonstrating vulnerability-mediated harm model).

vulnerability and accelerated her into pro-anorexia networks within days. Population averages obscure individual catastrophes.

The law does not wait for scientific certainty. Tobacco litigation succeeded in the 1990s when the biological mechanism linking smoking to cancer remained contested.⁶⁰ Courts applied the substantial factor test: when defendant's conduct materially increases the risk of harm that materializes, causation is established—even if other factors contributed. ABT adopts this framework precisely because mental health is multifactorial. The question is not whether social media is the sole cause of adolescent depression, but whether algorithmic manipulation is a substantial factor in the harms plaintiffs suffer. Platform internal research—controlled experiments proving emotional manipulation,⁶¹ not mere correlations—establishes that it is.

BLACK-LETTER ALGORITHMIC BEHAVIORAL TORT

Restatement (Third) of Torts: Liability for Physical and Emotional Harm (Author-Proposed Supplement; Not an ALI Restatement)

§ 1. The Algorithmic Behavioral Tort

- (a) General Rule. An actor who designs, deploys, or operates an algorithmic curation system is subject to liability for physical or emotional harm to a user if: (1) the actor owes a duty of care to the user to respect their cognitive integrity; (2) the actor breaches that duty by engaging in Actionable Behavioral Conditioning; (3) the breach is a factual and

⁶⁰See Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 PROC. ROYAL SOC'Y MED. 295 (1965).

⁶¹Adam D. I. Kramer, Jamie E. Guillory & Jeffrey T. Hancock, *Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks*, 111 PROC. NAT'L ACAD. SCI. U.S.A. 8788 (2014).

proximate cause of the user’s harm; and (4) the user suffers legally cognizable physical or emotional injury.

- (b) Actionable Behavioral Conditioning. “Actionable Behavioral Conditioning” means the systematic use of algorithmic feedback loops to: (1) identify a user’s specific cognitive vulnerabilities; and (2) deliver high-velocity, variable-reward stimuli designed to override the user’s rational decision-making and compel engagement; (3) in a manner that exceeds the bounds of legitimate persuasion or editorial discretion.
- (c) Cognitive Integrity. “Cognitive Integrity” is the interest in the freedom of one’s mental processes from surreptitious, automated manipulation that substantially impairs the capacity for autonomous choice.
- (d) Algorithmic Curation System. An “algorithmic curation system” includes automated ranking, recommendation, and sequencing systems—including machine-learning models—and associated engagement features (e.g., autoplay, notifications, infinite scroll) that create feedback loops between user behavior and future curation outputs.
- (e) Scope of Liability. Liability under this Section is limited to the harm caused by the *conduct* of the algorithmic system itself, distinct from any harm caused solely by the expressive content of the information delivered.

Comment: *a. Rationale.* This Section recognizes that modern algorithmic systems—including machine-learning ranking and recommendation models and associated engagement mechanics—do not merely organize information but actively shape user behavior through variable reinforcement schedules. The harm addressed is not the offensiveness of the content, but the addictiveness of the delivery mechanism. *b. Cognitive Integrity.* This interest parallels the right to bodily integrity. Just as a surgeon cannot operate without informed consent, a platform cannot

covertly use an algorithmic system to condition a user's neural pathways without violating their cognitive autonomy.

§ 2. *The Manipulation-Vulnerability Matrix (Rule of Decision)*

- (a) To determine whether an actor's conduct constitutes Actionable Behavioral Conditioning under § 1(b), the trier of fact shall apply the Manipulation-Vulnerability Matrix, balancing: (1) The Degree of Algorithmic Manipulation: The intensity, velocity, and clandestine nature of the behavioral modification techniques employed (e.g., variable ratio reinforcement schedules, dark patterns, subliminal priming); against (2) The Degree of User Vulnerability: The specific sensitivity of the user, known or constructively known to the actor (e.g., age, addiction history, emotional state).
- (b) Presumption of Breach. A rebuttable presumption that the actor engaged in Actionable Behavioral Conditioning arises where the actor deploys high-intensity manipulation techniques against a user with high vulnerability (the "Red Zone").
- (c) Safe Harbor. No liability attaches where the actor employs only standard persuasive techniques (e.g., chronological feeds, transparent recommendations) against a user of ordinary susceptibility (the "Green Zone").

Comment: *a. Limiting Principles.* This Section regulates conduct, not ideas: it targets only content-agnostic behavioral conditioning mechanics and vulnerability targeting, not the viewpoints conveyed in third-party content. Section 230 need not bar claims where liability turns on the design and operation of those mechanics rather than the publication of third-party speech. The First Amendment protects editorial discretion about ideas; it does not extend to content-

neutral conditioning mechanics that operate independently of content selection. No ABT claim lies in the Green Zone; the clearest cases involve minors subjected to Red Zone conduct.

PART I: THE ALGORITHMIC BEHAVIORAL TORT FRAMEWORK

Existing tort doctrines fail systematically against algorithmic harms. ABT provides a coherent, implementable doctrine built from first principles. The task is not to invent, but to synthesize—to recognize that the common law already possesses the conceptual tools necessary to address algorithmic manipulation, if courts will adapt them to modern threats as they have for centuries.

This Part develops ABT through seven components: theoretical foundations, the protected interest (cognitive integrity), the source of duty (algorithmic fiduciary relationship), formal elements, an analytical tool (the Manipulation-Vulnerability Matrix), damages and apportionment, and an evidence proof map. The emphasis throughout is practical applicability—ABT is workable doctrine courts can apply today, without legislative action, without waiting for academic consensus, and without abandoning centuries of tort law wisdom.

A. Theoretical Foundations: Synthesis from Proven Principles

ABT does not create new legal principles; it adapts existing doctrine to new threats.⁶² The common law has confronted similar challenges before—complexity, power asymmetry, diffuse

⁶²The emerging field of algorithmic accountability draws on foundational work across multiple domains. *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008) (applying due process principles to automated decision-making); Frank Pasquale, *THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION* (2015) (documenting algorithmic opacity and proposing transparency remedies); Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477 (2006) (providing analytical framework for privacy harms adaptable to algorithmic contexts); Julie E. Cohen, *BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM* (2019) (analyzing how platform power shapes legal categories).

harm—and developed workable solutions.⁶³ ABT draws on four domains that successfully addressed these challenges.

1. Environmental Law: Cumulative and Probabilistic Harm

Toxic tort litigation faced challenges analogous to algorithmic harms: long latency, multiple causes, cumulative exposure. Courts adapted. They abandoned rigid “but-for” causation and embraced the “substantial factor” test—recognizing that when multiple sources contribute to injury, plaintiffs need not isolate a single culprit.⁶⁴ The doctrinal evolution from *Anderson v. Minneapolis, St. P. & S. Ste. M. Ry. Co.* through the Restatement framework demonstrates courts’ capacity to adapt causation standards to complex, multi-source harms.⁶⁵ Some jurisdictions developed market-share liability, most notably in *Sindell v. Abbott Laboratories*, where the California Supreme Court addressed the impossible identification problem for DES manufacturers.⁶⁶ Courts also accepted epidemiological evidence to bridge evidentiary gaps where individualized proof was impossible.⁶⁷

These principles translate directly to algorithms. Environmental law treats prolonged exposure as a “continuing tort”—a single ongoing wrong.⁶⁸ ABT adopts this approach: sustained algorithmic manipulation is a continuing tort. The logic is straightforward. If the law holds

⁶³See GUIDO CALABRESI, A COMMON LAW FOR THE AGE OF STATUTES 1–7 (1982) (analyzing common law’s adaptive capacity); MORTON J. HORWITZ, THE TRANSFORMATION OF AMERICAN LAW, 1870–1960, at 51–63 (1992) (documenting tort law’s evolution in response to industrial revolution).

⁶⁴See, e.g., TROYEN A. BRENNAN, *Causal Chains and Statistical Links*, 73 CORNELL L. REV. 469 (1988).

⁶⁵*Anderson v. Minneapolis, St. P. & S. Ste. M. Ry. Co.*, 179 N.W. 45, 46–47 (Minn. 1920) (establishing substantial factor test for concurrent causation). See also RESTATEMENT (SECOND) OF TORTS § 431 (1965) (adopting substantial factor terminology); RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 27 (2010) (refining framework for multiple sufficient causes).

⁶⁶See *Sindell v. Abbott Labs.*, 607 P.2d 924, 937 (Cal. 1980).

⁶⁷See *Sindell*, 607 P.2d at 937.

⁶⁸See, e.g., *Hoery v. United States*, 64 P.3d 214, 217–18 (Colo. 2003).

companies accountable for asbestos causing cancer decades later, it can address algorithms cultivating eating disorders over months.

2. Public Health Law: Population-Level Protection

The youth mental health crisis is a population-level phenomenon. Public health law addresses precisely this type of widespread, low-probability but high-impact risk, identifying risk factors like smoking or lead exposure and intervening at population scale with special focus on vulnerable groups.

ABT operationalizes this approach within tort law. Rather than requiring individualized proof of sole causation, ABT treats manipulative algorithmic design as a significant risk factor platforms must mitigate. This aligns with the precautionary principle: where threats of serious irreversible damage exist, lack of full scientific certainty does not justify postponing cost-effective preventive measures.⁶⁹ By incorporating population-level data—“legal epidemiology”—to inform duty and causation standards, courts can recognize that when platform design choices demonstrably increase population-wide incidence of depression or self-harm, a legally cognizable wrong has occurred.

3. Financial Regulation: Fiduciary Duties for Algorithms

Financial regulation provides one of ABT’s most powerful precedents. Investment advisers have long faced fiduciary standards. The Securities and Exchange Commission has clarified that this duty applies to the entire advisory relationship, including advice provided by “robo-advisers.”⁷⁰ This principle has been reinforced by DOL fiduciary rules and case law

⁶⁹See Rio Declaration on Environment and Development, *supra* note 53.

⁷⁰SEC, COMMISSION INTERPRETATION REGARDING STANDARD OF CONDUCT FOR INVESTMENT ADVISERS, 84 FED. REG. 33,669 (July 12, 2019).

holding that algorithmic recommendations do not shield advisers from fiduciary obligations.⁷¹

The firm remains responsible for ensuring algorithms serve client interests, not exploit them for the firm's benefit.

This principle—that automation does not extinguish human responsibility—is ABT's cornerstone. Like financial advisers, platforms exercise discretionary control over crucial aspects of users' lives: their information environment. They operate with vast information asymmetry, cultivate trust and dependence. If an algorithm recommending stock portfolios owes a duty of loyalty, then one recommending worldviews to vulnerable adolescents should too.

4. Consumer Protection: Recognizing Manipulation

Finally, ABT builds on consumer protection law's evolution to recognize manipulation as harm distinct from deception. Traditional fraud requires a false statement. The FTC's 2022 "dark patterns" report detailed how user interfaces can be designed to "obscure, subvert, or impair consumer choice," manipulating users without any false statement.⁷² Empirical research has documented the prevalence and effectiveness of dark patterns across digital platforms, demonstrating that manipulative design is systematic rather than incidental.⁷³ ABT completes

⁷¹See Definition of the Term "Fiduciary"; Conflict of Interest Rule—Retirement Investment Advice, 81 Fed. Reg. 20,946 (Apr. 8, 2016) (DOL rule extending fiduciary standard to rollover advice); SEC, Commission Interpretation Regarding Standard of Conduct for Investment Advisers, Investment Advisers Act Release No. 5248 (June 5, 2019) (confirming fiduciary duty applies regardless of whether advice is delivered by algorithm). See also Arthur B. Laby, *SEC v. Capital Gains Research Bureau and the Investment Advisers Act of 1940*, 91 B.U. L. REV. 609 (2011) (analyzing scope of adviser fiduciary duties).

⁷²See FEDERAL TRADE COMMISSION, *supra* note 5.

⁷³See Arunesh Mathur et al., *Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites*, 3 PROC. ACM ON HUM.-COMPUTER INTERACTION 1 (2019) (documenting prevalence of manipulative design across e-commerce); Jamie Luguri & Lior Jacob Strahilevitz, *Shining a Light on Dark Patterns*, 13 J. LEGAL ANALYSIS 43 (2021) (empirical study demonstrating dark patterns' effectiveness and proposing regulatory responses); Harry Brignull, *DARK PATTERNS: DECEPTION VS. HONESTY IN UI DESIGN*, A LIST APART (Nov. 1, 2011), <https://alistapart.com/article/dark-patterns-deception-vs-honesty-in-ui-design/> (last visited Dec. 23, 2025) (coining the term "dark patterns").

this doctrinal evolution, making covert psychological manipulation itself actionable when it breaches a duty of care and results in foreseeable cognitive harm.

B. The Protected Interest: Cognitive Integrity

Every tort protects an interest: battery protects bodily integrity; defamation, reputation. ABT protects cognitive integrity—an individual’s right to mental self-determination free from covert, non-consensual manipulation. While seemingly novel, it is a necessary extension of deeply rooted legal principles of autonomy into an era where technology has unprecedented capacity to influence the human mind. It is the fundamental digital-age right.

1. Defining Cognitive Integrity

Cognitive integrity—grounded in the philosophical concept of cognitive liberty⁷⁴—is the right to have thoughts, preferences, and decisions be products of authentic, autonomous mental processes—the right to think for oneself.⁷⁵ It comprises three essential components:

Transparency of Influence: The right to know when and how one is being systematically targeted for influence.

Meaningful Choice: The right to accept, reject, or disengage from such influence without undue friction or penalty.

⁷⁴Jan-Christoph Bublitz, *My Mind Is Mine!? Cognitive Liberty as a Legal Concept*, in NEURORIGHTS AND AUTONOMOUS VEHICLE DESIGN (2013).

⁷⁵See Jan Christoph Bublitz & Reinhard Merkel, *Crimes Against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination*, 8 CRIM. L. & PHIL. 51 (2014) (foundational articulation of cognitive liberty as protected interest); Marcello Ienca & Roberto Andorno, *Towards New Human Rights in the Age of Neuroscience and Neurotechnology*, 13 LIFE SCI., SOC’Y & POL’Y 1 (2017) (proposing cognitive liberty as emerging human right); see also NITA A. FARAHANY, *THE BATTLE FOR YOUR BRAIN: DEFENDING THE RIGHT TO THINK FREELY IN THE AGE OF NEUROTECHNOLOGY* (2023) (comprehensive treatment of cognitive liberty threats and protections).

Protection from Exploitation: The right not to have identified psychological vulnerabilities systematically targeted for another's commercial or political gain.

Cognitive integrity is not a right to be free from persuasion or advertising. It protects the boundary between legitimate persuasion (which appeals to conscious deliberative capacities) and illegitimate manipulation (which bypasses or subverts them). To be clear: ABT does not reach ordinary advertising, political advocacy, or ideological content curation. It reaches only the combination of (1) covert behavioral conditioning techniques, (2) systematic targeting of detected psychological vulnerabilities, and (3) conduct causing substantial documented harm.

Practical test: (1) Would a reasonable user, fully informed of the psychological technique, consent to its application? (2) Does the technique primarily work by covertly exploiting nonconscious vulnerabilities rather than engaging deliberative choice? If effectiveness depends on covert or non-consensual application, it presumptively violates cognitive integrity.

The empirical foundation: when platforms are transparent, users resist manipulation.⁷⁶ But platforms deliberately obscure their systems—opacity is the essential precondition for effective manipulation. This distinction echoes behavioral economics' foundational concept of “choice architecture,” differentiating transparent design from manipulative design that exploits nonconscious vulnerabilities.⁷⁷

The Evolutionary Arc: Battery → Privacy → Cognitive Integrity

⁷⁶Corrigan & Forzani, *I Would Have Gone to the Original Source*, 174 COMPUTERS IN HUMAN BEHAVIOR 108818 (2026), <https://doi.org/10.1016/j.chb.2025.108818> (last visited Dec. 23, 2025).

⁷⁷Richard H. Thaler, Cass R. Sunstein & John P. Balz, *Choice Architecture*, SSRN (2008), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1583509 (last visited Dec. 23, 2025); RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 3–6 (2008) (defining choice architecture and distinguishing between libertarian paternalism and manipulative design); Susanne K. Smaak & Bettina Wiskemann, *Algorithmic Transparency and Manipulation*, 41 PHIL. & TECH. 18 (2023), <https://doi.org/10.1007/s13347-023-00678-9> (last visited Dec. 23, 2025).

This legal recognition of cognitive integrity follows a well-established pattern of common-law evolution. Each major technological shift that threatened a sphere of human autonomy has prompted tort law to extend protection to a previously undefended interest. The progression is logical and inevitable.

In the nineteenth century, the common law recognized battery to protect bodily autonomy from unwanted physical contact.⁷⁸ The principle was revolutionary for its time: your body is inviolate; another's touch without consent is a legal wrong, regardless of whether it causes physical injury. Battery established that autonomy itself—not just freedom from harm—is a protected interest.

In 1890, Warren and Brandeis argued that new technology—the instant camera and the tabloid press—required recognition of a “Right to Privacy” to protect the “inviolable personality” from unwanted publication.⁷⁹ Photography had made it possible to capture and disseminate a person's image without consent, threatening a new dimension of autonomy. Their argument succeeded because it was conservative, not radical: they showed that privacy was the logical extension of existing principles (battery, property, confidence) to a new threat.⁸⁰

Today, the algorithmic “Rage Machine” demands the next evolutionary step: cognitive integrity, protecting mental processes from unwanted manipulation. The pattern repeats. Just as battery does not prohibit all touching (a tap on the shoulder to get attention remains lawful), and

⁷⁸See RESTATEMENT (SECOND) OF TORTS § 18 (Am. L. Inst. 1965) (“An actor is subject to liability to another for battery if (a) he acts intending to cause a harmful or offensive contact with the person of the other or a third person, or an imminent apprehension of such a contact, and (b) a harmful [or offensive] contact with the person of the other directly or indirectly results.”). The key innovation was recognizing that *offensive* contact—not just physically harmful contact—constitutes a tort. See DAN B. DOBBS, THE LAW OF TORTS § 29 (2000).

⁷⁹Warren & Brandeis, *supra* note 18.

⁸⁰*Id.*, at 213 (“The common law has always recognized a man's house as his castle, impregnable, often, even to its own officers engaged in the execution of its commands. Shall the courts thus close the front entrance to constituted authority, and open wide the back door to idle or prurient curiosity?”). Warren and Brandeis framed their argument as *extending* existing principles, not inventing new ones—a method ABT consciously emulates.

privacy does not prohibit all publication (newsworthy facts remain publishable), cognitive integrity does not prohibit all persuasion. It prohibits only the surreptitious, systematic exploitation of detected psychological vulnerabilities to override autonomous choice—the mental equivalent of the nonconsensual touching that battery forbids.

The doctrinal arc is clear: Battery (1800s) → Privacy (1890) → Cognitive Integrity (today). Each protects a sphere of self-determination against the technology that threatens it. ABT is not innovation; it is synthesis.

2. Legal and Constitutional Foundations

The Supreme Court’s “fixed star in our constitutional constellation,” from *West Virginia State Board of Education v. Barnette*, declares that “no official, high or petty, can prescribe what shall be orthodox in politics, nationalism, religion, or other matters of opinion.”⁸¹ The mind is a sanctuary. If government is forbidden from coercing thought, commercial regimes enabling private corporations to systematically manipulate it for profit warrant legal skepticism.

The Due Process Clause provides another foundation. *Planned Parenthood v. Casey* states: “[a]t the heart of liberty is the right to define one’s own concept of existence, of meaning, of the universe, and of the mystery of human life.”⁸² Mental autonomy is prerequisite to such self-definition. Cognitive integrity is the internal analogue to external liberties long protected.

The common law tort of intrusion upon seclusion offers a direct antecedent. If law protects physical solitude from invasive photography, it must protect “mental solitude” from invasive algorithms mapping and exploiting psychological vulnerabilities.⁸³ The emerging field

⁸¹West Virginia State Bd. of Educ. v. Barnette, 319 U.S. 624, 642 (1943).

⁸²Planned Parenthood of Se. Pa. v. Casey, 505 U.S. 833, 851 (1992).

⁸³See, e.g., RESTATEMENT (SECOND) OF TORTS § 652B (1977).

of design regulation provides additional conceptual support: platforms that design systems to exploit psychological vulnerabilities engage in what scholars term “abusive design”—manipulative architecture that undermines user autonomy through the structure of choice environments.⁸⁴

These constitutional and common law principles are gaining contemporary articulation through neurorights scholarship, which recognizes that cognitive liberty must be protected as an emerging fundamental right in the digital age.⁸⁵ International human rights frameworks increasingly establish protection of mental autonomy as a foundational human right in the age of neuroscience and neurotechnology.⁸⁶

3. The Internal Logic of Autonomy Protection

Cognitive integrity completes tort law’s architecture of personal autonomy. Battery protects bodily integrity.⁸⁷ Defamation protects reputational integrity.⁸⁸ Privacy torts protect informational autonomy.⁸⁹ IIED protects psychological integrity.⁹⁰ Cognitive integrity protects

⁸⁴See WOODROW HARTZOG, *PRIVACY’S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES* 7–12, 165–89 (2018) (developing framework for design regulation and identifying “abusive design” as cognizable harm); see also Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995 (2014) (analyzing how digital interfaces enable new forms of consumer manipulation); BRETT FRISCHMANN & EVAN SELINGER, *RE-ENGINEERING HUMANITY* 5–8 (2018) (documenting how technological design shapes human agency).

⁸⁵Nita A. Farahany, *The Overdue Arrival of Neurorights*, 50 CONN. L. REV. 1 (2020).

⁸⁶Eloïse S. Pogrebin, *Towards New Human Rights in the Age of Neuroscience and Neurotechnology*, 10 LIFE SCI., SOC’Y & POL’Y 2 (2014).

⁸⁷See RESTATEMENT (SECOND) OF TORTS § 13 (1965) (defining battery as harmful or offensive contact with another); see, e.g., *Cole v. Turner*, 90 Eng. Rep. 958, 958 (K.B. 1704) (“the least touching of another in anger is a battery”); RESTATEMENT (SECOND) OF TORTS § 18 cmt. c (1965).

⁸⁸See RESTATEMENT (SECOND) OF TORTS §§ 558–559, 623A (1977) (defamation protects reputation as external manifestation of personality and character).

⁸⁹See RESTATEMENT (SECOND) OF TORTS § 652B (1977) (intrusion upon seclusion); see RESTATEMENT (SECOND) OF TORTS § 652D (1977) (public disclosure of private facts).

⁹⁰See RESTATEMENT (SECOND) OF TORTS § 46 (1965) (intentional infliction of emotional distress requires extreme and outrageous conduct).

the internal processes that make all other autonomy meaningful—what good is bodily autonomy if decisions are products of covert manipulation? Battery says you own your body. Privacy says you control your information. Cognitive integrity says you own your mind.

4. Cognitive Integrity and the Cartesian Self

These legal protections for mental autonomy are not mere constitutional artifacts or doctrinal conveniences. They are the legal expression of a philosophical cornerstone of Western thought: the sanctity of the thinking, autonomous self. The foundation of modern individualism is René Descartes's declaration in his *Meditations on First Philosophy*: "*Cogito, ergo sum*"—"I think, therefore I am."⁹¹ Yet the enduring power of this statement lies not in its conclusion, but in the radical doubt from which it emerged.

To arrive at certainty, Descartes imagined a "malicious demon" (*malin génie*) who could fabricate every perception.⁹² His answer: Even if deceived about everything else, the very act of thinking proved the existence of a thinking self.

Today, Descartes's demon is no longer hypothetical—it is an architectural reality. The engagement-maximizing algorithm is the modern *malin génie*: a powerful, opaque system curating individualized reality for each user.⁹³ Like the demon, it operates with "utmost cunning." Like the demon, it works covertly, making it nearly impossible to distinguish authentic experience from manufactured influence.

⁹¹RENÉ DESCARTES, *MEDITATIONS ON FIRST PHILOSOPHY* (1641), Second Meditation (John Cottingham trans., Cambridge Univ. Press 1996) (originally published as *Meditationes de Prima Philosophia*).

⁹²*Id.*, First Meditation. Descartes writes: "I will suppose therefore that not God, who is supremely good and the source of truth, but rather some malicious demon of the utmost power and cunning has employed all his energies in order to deceive me."

⁹³Kate Crawford, *Enchanted Determinism: Power without Responsibility in Artificial Intelligence*, 2 *ENGAGING SCI., TECH., & SOC'Y* 1 (2016).

Where Descartes's demon threatened the *external* world, the algorithmic demon threatens the *internal*. It identifies when an adolescent feels “worthless” or “insecure” and exploits that vulnerability—not to support well-being, but to maximize engagement.⁹⁴ Users still think, but increasingly think *what they have been shaped to think*. The self becomes partly an algorithmic artifact.⁹⁵

ABT is the modern shield against the algorithmic demon—protecting the *process* of thought, not merely content. Cognitive integrity is the ancient right of mental self-determination, translated into law to meet the most sophisticated threat to autonomous thought: algorithmic systems operating at planetary scale.⁹⁶

Descartes imagined an all-powerful demon controlling his perceptions. Today, the demon has a Terms of Service.

C. The Duty: The Algorithmic Fiduciary Relationship

Identifying cognitive integrity as the protected interest establishes what deserves protection. Now we must establish when platforms owe a legal duty. The answer is simple: when a platform exercises discretionary, opaque algorithmic control over a user's information environment, it owes that user a duty.

⁹⁴See *supra* note 37 (Facebook document describing advertisers' ability to identify and target teen emotional states); VIRGINIA DIGNUM, RESPONSIBLE ARTIFICIAL INTELLIGENCE: HOW TO DEVELOP AND USE AI IN A RESPONSIBLE WAY (Springer 2019); see also Kramer et al., *supra* note 52 (Facebook's emotional contagion experiment demonstrating platform's ability to manipulate user emotional states through algorithmic curation).

⁹⁵SAFIYA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (New York Univ. Press 2018).

⁹⁶KATE CRAWFORD, ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE (Yale Univ. Press 2021).

ABT proposes: Duty arises from an algorithmic fiduciary relationship—when a platform exercises significant, discretionary, opaque algorithmic control over a user’s information environment, it enters a trust relationship analogous to traditional fiduciary relationships, thereby incurring fiduciary-like duties of loyalty, care, and candor. Power creates duty. The duty must match the power.

This fiduciary relationship rests on a deeper truth: algorithms become integrated into users’ cognitive processes, forming what philosophers term the “extended mind.”⁹⁷ Users increasingly internalize algorithmic rankings as their own preferences. Just as we prohibit therapists and financial advisors from exploiting their deep knowledge of individual psychology, we cannot permit algorithms that have become part of users’ cognitive functioning to do so.

1. Foundation: Information and Power Asymmetry

Fiduciary duties address inherent inequality, arising where one party must trust another possessing discretionary power. The platform-user relationship exhibits fiduciary hallmarks:

Knowledge asymmetry is profound. Platforms possess comprehensive visibility into users, collecting behavioral data to build psychological profiles predicting emotional states, personality traits, and vulnerabilities. Users know almost nothing about governing algorithms.

This creates stark power asymmetry. Platforms exercise near-total control over users’ information environment. User control is largely illusory. While theoretically they can leave, network effects and design-induced addiction make exit extraordinarily difficult.

⁹⁷Andy Clark & David Chalmers, *The Extended Mind*, 58 ANALYSIS 7 (1998) (foundational articulation of extended cognition thesis). Clark and Chalmers argued that cognitive processes can extend beyond the brain to include external tools and technologies when those tools are reliably available, automatically endorsed, and easily accessible. Algorithmic recommendation systems increasingly satisfy these criteria for heavy users, suggesting platforms have become integrated into users’ cognitive apparatus; Rapp & Boldi, *The Quantification of the Gaming Experience*, 174 COMPUTERS IN HUMAN BEHAVIOR 108826 (2026), <https://doi.org/10.1016/j.chb.2025.108826> (last visited Dec. 23, 2025).

Sidebar: Why Exit Isn't Easy

- Network effects: Friends, school groups, and social identity reside on the platform; leaving carries social cost⁹⁸
- Compulsion design: Variable rewards and frictionless scroll keep attention captured⁹⁹
- Asymmetry: Platforms observe everything and optimize; users see outputs, not mechanics. The system can infer your vulnerabilities faster than you can identify them¹⁰⁰

This is a relationship of dependence. For millions, especially adolescents, platforms are primary portals for social connection, information, and identity formation. This confluence is functionally identical to traditional fiduciary relationships: the platform knows more, controls more, holds users in dependence.¹⁰¹ Critics have questioned whether information fiduciary frameworks can adequately constrain platforms without more specific substantive obligations;¹⁰² ABT addresses this concern by operationalizing fiduciary principles through the Manipulation-

⁹⁸See, e.g., Michael L. Katz & Carl Shapiro, *Network Externalities, Competition, and Compatibility*, 75 *Am. Econ. Rev.* 424 (1985).

⁹⁹See Schüll, *Addiction by Design*, *supra* note 19.

¹⁰⁰See Zuboff, *Surveillance Capitalism*, *supra* note 10, at 93–97 (documenting platforms' unprecedented informational advantages through behavioral surplus extraction).

¹⁰¹See Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 *U.C. DAVIS L. REV.* 1183 (2016) [hereinafter Balkin, *Information Fiduciaries*].

¹⁰²See Lina M. Khan & David E. Pozen, *A Skeptical View of Information Fiduciaries*, 133 *HARV. L. REV.* 497, 520–41 (2019) (arguing information fiduciary frameworks lack determinate content and effective enforcement mechanisms). ABT addresses Khan and Pozen's administrability concerns through two mechanisms. First, the Manipulation-Vulnerability Matrix provides the concrete, case-specific standards they found lacking—courts assess manipulation intensity and user vulnerability on defined scales with operational thresholds, not vague “act in user interest” mandates. *See infra* Part I.D (providing threshold tables). Second, ABT's alternative pathway through the special relationship doctrine, *see infra* Part I.C.4, reaches identical results using established precedent (*Tarasoff*, *Rowland*) rather than novel fiduciary classifications. Courts skeptical of extending fiduciary obligations to platforms can impose identical duties through the special relationship framework without adopting the “information fiduciary” label.

Vulnerability Matrix, providing courts with concrete, administrable standards rather than abstract duties.¹⁰³

2. Three Fiduciary Duties Adapted for Algorithms

Recognition of algorithmic fiduciary relationship gives content to the platform's duty of care through three traditional fiduciary obligations:

Duty of Loyalty: The core fiduciary obligation requiring action in the beneficiary's interest without self-dealing. For platforms, this means not prioritizing financial interest in engagement over known cognitive well-being. A platform breaches when knowingly designing algorithms to amplify harmful content simply because it drives more engagement. This duty does not prohibit profiting; it prohibits profiting through exploitation.

Duty of Care: This requires acting with skill and prudence a reasonably careful person would exercise in a like position. For platforms, this means exercising reasonable care in designing, testing, and monitoring algorithms with foreseeable impacts on cognitive processes. Breach occurs when platforms fail to conduct risk assessments, ignore internal research documenting harms, or deploy known-addictive features without adequate warnings or controls. This duty is heightened for vulnerable users, such as children.

Duty of Candor: This requires disclosing all material information, including conflicts of interest. For platforms, this requires meaningful, understandable explanation of recommendation systems, disclosure of conflicts (engagement maximization vs. user well-being), and warnings of known risks.

¹⁰³See *infra* Part I.D (developing Matrix as operational tool translating fiduciary principles into administrable standards).

3. Defining Scope: When the Duty Arises

This duty is not absolute nor does it apply to the entire internet. The algorithmic fiduciary duty attaches only when platforms cross a specific threshold: using personalized algorithmic curation as the primary means of controlling users' information environment.

This duty applies to: Social media news feeds, “For You” pages, auto-play systems.

This duty does not apply to: Chronologically sorted feeds, basic search engines, passive websites.

A “covered platform” exhibits several factors:

Primary use of personalized recommender systems to curate individual information environments

Significant opacity of ranking/selection criteria

Collection and use of behavioral telemetry to infer and target psychological states

Material user dependence for social connection (especially minors)

Demonstrated ability to materially alter user behavior through design changes

The duty's intensity is proportional, scaling based on: (1) degree of algorithmic control; (2) vulnerability of user population; (3) foreseeability of severe harm; and (4) availability of feasible safer alternatives. Courts uncomfortable with “fiduciary” language may reach identical results through the special relationship doctrine.¹⁰⁴

¹⁰⁴ALTERNATIVE PATHWAY: SPECIAL RELATIONSHIP DOCTRINE. While the fiduciary framework provides the strongest doctrinal foundation for ABT, courts hesitant to adopt “information fiduciary” terminology can reach identical results through the established concept of a “special relationship” giving rise to heightened duties. The platform-user relationship exhibits classical hallmarks: (1) profound information and power asymmetry—platforms possess comprehensive behavioral surveillance while users remain blind to algorithmic mechanics; (2) discretionary control over critical resources—platforms individually tailor users' information environments without meaningful user oversight; (3) reasonable reliance and induced dependence—users depend on platforms for essential social functions, and platforms deliberately induce compulsive use through addictive design; (4) vulnerability of protected class—minors have immature cognitive development and heightened impulsivity; (5) foreseeability of severe

4. Alternative Pathway: The Special Relationship Doctrine

For courts hesitant to extend fiduciary duties, ABT provides an alternative through the special relationship doctrine.¹⁰⁵ Under *Tarasoff*, special relationships create affirmative duties based on knowledge and control.¹⁰⁶ Platforms satisfy these criteria: comprehensive knowledge of user vulnerabilities (exceeding even therapeutic relationships), discretionary control over information environment, and design features targeting identified vulnerabilities. Whether characterized as “fiduciary” or “special relationship,” the legal consequence is identical: platforms owe duties of care to users whose cognitive processes they systematically influence.¹⁰⁷

harm—platforms’ internal research documents foreseeable psychological injury; and (6) public policy favoring protection—protecting children’s cognitive development from commercial exploitation serves compelling interests repeatedly recognized by courts. These factors establish a “special relationship” imposing duties of reasonable care, loyalty to user well-being, and candor about material conflicts. The substance of duty is identical to the fiduciary framework; only the label differs. Courts should choose the framing that best fits their jurisdiction’s precedents, confident that either route reaches sound results. *See* RESTATEMENT (SECOND) OF TORTS § 315 (1965) (special relationships create affirmative duties); RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 40 (2012) (special relationships imposing duty to protect from risk of physical harm); *Tarasoff v. Regents of Univ. of Cal.*, 551 P.2d 334, 342–43 (Cal. 1976) (special relationship between therapist and patient creates affirmative duty to exercise reasonable care to protect foreseeable victims from serious danger); *Rowland v. Christian*, 443 P.2d 561, 564 (Cal. 1968) (articulating factors for duty determination: foreseeability, certainty of injury, closeness of connection, moral blame, policy of preventing harm, burden on defendant, and availability of insurance—all of which favor imposing duty on platforms). *See also* Balkin, *Information Fiduciaries*, *supra* note 101, at 1190–1201.

¹⁰⁵RESTATEMENT (SECOND) OF TORTS § 315 (Am. L. Inst. 1965) (“There is no duty so to control the conduct of a third person as to prevent him from causing physical harm to another unless (a) a special relation exists between the actor and the third person which imposes a duty upon the actor to control the third person’s conduct, or (b) a special relation exists between the actor and the other which gives to the other a right to protection.”). Though originally framed in terms of physical harm, courts have extended these principles to psychological and emotional harms where special relationships exist. *See* DAN B. DOBBS, *THE LAW OF TORTS* § 319 (2000).

¹⁰⁶*Tarasoff v. Regents of Univ. of Cal.*, 551 P.2d 334, 343 (Cal. 1976) (holding that when a therapist determines, or pursuant to professional standards should determine, that a patient presents a serious danger of violence to a foreseeable victim, the therapist bears a duty to exercise reasonable care to protect that victim). The special relationship doctrine has since expanded beyond therapeutic contexts to encompass relationships characterized by knowledge asymmetry, control, and foreseeable harm. *See* RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 41 (Am. L. Inst. 2012); *Rowland v. Christian*, 443 P.2d 561, 564 (Cal. 1968) (articulating factors for determining when duty arises: “the foreseeability of harm to the plaintiff, the degree of certainty that the plaintiff suffered injury, the closeness of the connection between the defendant’s conduct and the injury suffered, the moral blame attached to the defendant’s conduct, the policy of preventing future harm, the extent of the burden to the defendant and consequences to the community of imposing a duty to exercise care with resulting liability for breach, and the availability, cost, and prevalence of insurance for the risk involved”).

¹⁰⁷*See* Elettra Bietti, *A Genealogy of Digital Platform Regulation*, 7 GEO. L. TECH. REV. 1 (2023) (tracing regulatory approaches to platform governance); *see also* Kathryn Judge, *The Limits of Liability*, 87 U. CHI. L. REV. 2045 (2020) (analyzing structural limitations of tort law for systemic harms); *see* Jack M. Balkin, *How to Regulate*

D. The Manipulation-Vulnerability Matrix: ABT's Central Analytical Tool

The Manipulation-Vulnerability Matrix is this Article's most practical contribution. Courts can use it today—without legislation, without academic consensus, without delay. It translates ABT's theory into operational doctrine that judges can apply to the cases already on their dockets.¹⁰⁸

Like any good legal tool, the Matrix provides structure without rigidity. It offers systematic breach analysis while preserving judicial discretion for case-specific facts.

The Matrix serves three functions. First, it operationalizes the fiduciary duties by scaling breach analysis to manipulation severity and user vulnerability. Second, it calibrates remedies proportionally to harm. Third, it provides predictability—both plaintiffs and platforms know which design choices cross the line from permissible optimization to tortious manipulation.

The Matrix Framework

ABT's Manipulation-Vulnerability Matrix provides a two-axis analytical tool for breach analysis:

Axis 1: Manipulation Intensity (scored Low/Medium/High/Extreme)

Covertiness of technique: Does it work by bypassing conscious deliberation?

(and Not Regulate) *Social Media*, 1 J. FREE SPEECH L. 71 (2021) (framework for platform regulation); see also Ellen P. Goodman & Karen Kornbluh, *Social Media Platforms Need an Alternative to the Orthodoxy of Free Speech Maximalism*, Knight First Amend. Inst. (2021).

¹⁰⁸The Matrix operationalizes breach analysis in manner analogous to Judge Learned Hand's famous formula for negligence in *United States v. Carroll Towing Co.*, 159 F.2d 169, 173 (2d Cir. 1947) (liability where $B < PL$). Like the Hand formula, the Matrix provides structured reasoning while preserving judicial discretion for case-specific application. Tiered risk-assessment matrices are well-established in tort law. See John W. Wade, *On the Nature of Strict Tort Liability for Products*, 44 MISS. L.J. 825, 837–38 (1973) (articulating seven-factor risk-utility test for design defect widely adopted by courts); RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2 cmt. d (1998) (adopting risk-utility balancing as analytical framework for determining whether foreseeable risks could have been reduced by reasonable alternative design). Similarly, regulatory frameworks employ tiered classification systems. See 21 C.F.R. pt. 860 (FDA medical device classification into Class I, II, and III based on degree of risk to patients). The Matrix translates these proven analytical methods to the digital manipulation context.

Opacity to users: Can they understand or resist it?

Rejection of safer alternatives: Did the platform choose harm over known safer designs?

Axis 2: User Vulnerability (scored Low/Medium/High/Extreme)

Minor status: Children are heightened-duty beneficiaries

Known mental-health risk factors: Depression, anxiety, eating disorders

Design-induced dependence: Addiction features

Figure 1: The Manipulation-Vulnerability Matrix

Manipulation Intensity	User Vulnerability: Low	User Vulnerability: Medium	User Vulnerability: High	User Vulnerability: Extreme
Extreme	Viable Claim	Strong Liability	Presumption of Breach	HIGHEST LIABILITY
High	Closer Analysis	Viable Claim	Presumption of Breach	Presumption of Breach
Medium	Unlikely Liability	Closer Analysis	Viable Claim	Strong Liability
Low	No Liability	Unlikely Liability	Closer Analysis	Viable Claim

Reading the Matrix: Locate the intersection of manipulation intensity (rows) and user vulnerability (columns). The cell indicates liability exposure and remedy availability. The framework operates like a risk matrix in products liability or a sentencing grid in criminal law—systematic enough for predictability, flexible enough for case-specific judgment. Movement toward the upper-right quadrant increases liability exposure; movement toward the lower-left reduces or eliminates it. The boundaries are not arbitrary; they reflect the proportionality principle that distinguishes negligent design from tortious exploitation.

Matrix Application Principles:

Red Zone (High × High and above): Presumption of breach; heightened remedies including punitive damages

Yellow Zone (Medium combinations): Viable claims requiring full breach analysis;
standard compensatory damages

Green Zone (Low × Low): No liability (legitimate optimization within fiduciary bounds)

Boundary cases: Medium × Medium requires careful fact-specific analysis

The Matrix scales three elements proportionally with quadrant placement: (1) intensity of duty owed, (2) burden of proof on breach, and (3) available remedies. This proportionality ensures that the most egregious cases (high manipulation of highly vulnerable users) face the strictest scrutiny and most substantial consequences, while edge cases receive appropriately nuanced treatment.

Threshold Definitions

Courts require operational criteria for Matrix placement. The following tables provide the administrable standards necessary for consistent application.

Manipulation Intensity Thresholds

Level	Defining Characteristics	Indicative Features
Low	User-initiated engagement; transparent ranking criteria; meaningful opt-outs that function as advertised; no exploitation of psychological vulnerabilities	Chronological feeds; explicit preference settings; email digests; search results ranked by stated relevance
Medium	Algorithmic personalization present; some engagement optimization; opt-outs exist but require effort to locate; mild variable reward elements	Default algorithmic feeds with accessible alternatives; notification prompts; basic recommendation engines; “trending” features
High	Systematic amplification regardless of harm potential; documented internal research showing negative effects ignored; opt-outs ineffective	Infinite scroll; autoplay chains; emotion-weighted ranking (e.g., 5× anger); recommendation “rabbit

	or reset without notice; variable reward schedules deliberately deployed	holes”; dark patterns in settings
Extreme	Active identification and targeting of vulnerable users; content progressively escalates toward harm; internal research documents specific group harms yet feature retained; addiction mechanics as core design; no meaningful user agency	Sub-3-minute pathways to self-harm content; pro-eating-disorder amplification to flagged teens; “challenges” pushed to minors; systems that identify “insecure” users for targeted content

User Vulnerability Thresholds

Level	Defining Characteristics	Indicative Populations
Low	Adults with no documented risk factors; usage within normal parameters (under 2 hours daily); strong offline social networks; no platform-induced dependence	Occasional users; professionals using platforms instrumentally; adults with robust non-digital social lives
Medium	Adults with some risk factors; extended daily usage (2-4 hours); emerging dependence patterns; some susceptibility based on life circumstances	Heavy users; adults with mild anxiety or depression; users during divorce, job loss, or grief; isolated individuals
High	Documented psychological vulnerabilities; adolescents ages 13-17; prior history of addiction, eating disorders, or mental health treatment; platform-induced functional impairment evident	Teenagers generally; users with diagnosed conditions; users exhibiting compulsive checking, sleep disruption, or social withdrawal
Extreme	Young adolescents (13-15) at developmental peak susceptibility; users in acute mental health crisis; users whose own data reveals active distress signals; users platform systems have flagged as at-risk	13-14 year olds; users whose posts/searches indicate self-harm ideation; users platform’s safety systems have identified but continued serving harmful content

Zone-Specific Remedies

Zone	Liability Standard	Appropriate Remedies
------	--------------------	----------------------

Green (Low × Low)	No liability	No ABT claim lies; fiduciary duties satisfied through transparency and functional user controls
Yellow (Medium combinations)	Fact-intensive inquiry	Compensatory damages if breach proven; injunctive relief limited to specific harmful features; no presumption of punitive damages; defendant bears production burden on internal research
Red (High × High and above)	Presumption of breach	Full compensatory, hedonic, and dignitary damages; design-modification injunctions; mandatory algorithmic audits; punitive damages where internal knowledge documented; burden shifts to defendant to prove non-breach
Extreme × Extreme	Maximum exposure (rare, egregious cases)	All Red Zone remedies plus: punitive multiplier of 3-5× compensatory; court-appointed algorithmic monitor; mandatory safety-by-design modifications; suitable for class treatment and MDL consolidation

Concrete Matrix Hypotheticals

Hypothetical 1: Alexis Spence (Instagram Eating-Disorder Recommendations) (High × Extreme)

Facts (alleged): Alexis Spence created her first Instagram account at age eleven (despite Instagram’s 13+ minimum). She alleges Instagram’s recommendation system “almost immediately” steered her into an Explore-page “echo chamber” of eating-disorder content—

”thinspo” and accounts offering purging and extreme dieting instructions—and that she later developed anorexia requiring hospitalization.¹⁰⁹

Matrix Analysis: Manipulation Intensity: HIGH (recommendation-driven escalation toward harmful content; internal research documented teen-girl body-image harms; safeguards for minors allegedly insufficient).¹¹⁰ User Vulnerability: EXTREME (minor at peak susceptibility; eating disorders have highest mortality rate of any mental illness).¹¹¹ Quadrant: High × Extreme → Presumption of breach.

ABT Claim: Breach of loyalty, care, and candor. Causation: engagement logs + recommendation pathways + internal research. Harm: clinically diagnosed eating disorder.¹¹²

Hypothetical 2: The Outrage Amplifier (High × Medium)

Facts: David, age 42, uses Facebook primarily for family updates. The News Feed ranking system detects that he occasionally clicks on political posts. It begins serving increasingly polarizing political content, prioritizing posts that triggered “angry” reactions in similar users (the 5× weighting disclosed by Haugen). Over 18 months, David’s feed becomes

¹⁰⁹Complaint at ¶ 12, *Spence v. Meta Platforms, Inc.*, No. 3:22-cv-03294 (N.D. Cal. filed June 6, 2022) (consolidated into *In re Social Media Adolescent Addiction/Pers. Injury Prods. Liab. Litig.*, MDL No. 3047, 4:22-md-03047-YGR). The complaint alleges that despite Instagram’s Terms of Service requiring users to be at least 13 years old, the platform “encouraged” Spence to create an account at age 11 and “made no meaningful effort to verify her age or remove her account.”; Alex Chandler, *Family Sues Meta, Blames Instagram for Eating Disorder and Self-Harm*, NBC NEWS (June 8, 2022), <https://www.nbcnews.com/tech/social-media/meta-lawsuit-instagram-caused-eating-disorder-self-harm-rcna32221> (last visited Dec. 23, 2025) (quoting Spence complaint allegations and attorney statements); Complaint at ¶¶ 15, 18, 22, *Spence*, No. 3:22-cv-03294 (alleging that Instagram’s algorithm curated Spence’s Explore page to display “thinspo” imagery and systematically recommended pro-anorexia accounts). See also April Dembosky, *Alexis Spence Says Instagram Steered Her to Anorexia Content—Now She’s Suing Meta*, Slate (June 9, 2022), <https://slate.com/technology/2022/06/spence-lawsuit-instagram-meta.html> (last visited Dec. 23, 2025).

¹¹⁰See Wells, Horwitz & Seetharaman, *supra* note 4; see also Haugen, *supra* note 21.

¹¹¹Jon Arcelus et al., *Mortality Rates in Patients With Anorexia Nervosa*, 68 ARCH. GEN. PSYCHIATRY 724 (2011).

¹¹²Alex Chandler, *Family Sues Meta, Blames Instagram for Eating Disorder and Self-Harm*, NBC NEWS (June 8, 2022), <https://www.nbcnews.com/tech/social-media/meta-lawsuit-instagram-caused-eating-disorder-self-harm-rcna32221> (last visited Dec. 23, 2025) (quoting Spence complaint allegations and attorney statements).

dominated by outrage-inducing political content. His family reports he has become “angrier,” more polarized, and withdrawn from in-person social activities. He reports anxiety and sleep disruption.

Matrix Analysis: Manipulation Intensity: HIGH (5× anger weighting was deliberate; internal research warned platform was becoming “angrier, more polarizing”; safer alternatives rejected). User Vulnerability: MEDIUM (adult, but platform-induced dependence; documented anxiety/withdrawal). Quadrant: High × Medium → Viable claim.

ABT Claim: Breach of loyalty and care. Causation: Facebook’s emotional contagion research + engagement logs. Harm: Psychological distress with functional impairment.

Hypothetical 3: The Gaming Rabbit Hole (Extreme × Extreme)

Facts: Marcus, age 13, creates a TikTok account. He watches one video about a viral “challenge.” Within 20 minutes, TikTok’s “For You” recommendation system floods his feed with increasingly dangerous challenge videos: the “Blackout Challenge” (intentional asphyxiation), the “Benadryl Challenge” (overdose for hallucinations), and similar content. The system treats his extended viewing time on these videos as a strong engagement signal and infers interest. Marcus attempts the Blackout Challenge and suffers severe brain injury requiring hospitalization.

Matrix Analysis: Manipulation Intensity: EXTREME (rabbit holes within minutes; system is optimized to maximize “time spent” above all; self-harm content pushed to teens in under 3 minutes).¹¹³ User Vulnerability: EXTREME (13-year-old with limited impulse control, immature risk assessment). Quadrant: Extreme × Extreme → Maximum liability exposure.

¹¹³See BEN SMITH, *How TikTok Reads Your Mind*, N.Y. TIMES (Dec. 5, 2021) (analyzing leaked internal document “TikTok Algo 101” revealing algorithm’s “‘ultimate goal’ [is] adding daily active users” through optimization for “‘retention’—that is, whether a user comes back—and ‘time spent’”); see also Bobby Allyn, Sylvia Goodman &

ABT Claim: Heightened duty to minor. All three breaches present. Causation: Direct recommendation pathway to injury. Harm: Severe physical injury. This is ABT’s clearest application—if any case warrants recognition of ABT, it is this one.

The Green Zone protects platforms that provide genuine user controls. Where an adult user can meaningfully opt out of algorithmic curation—switching to chronological feeds, disabling notifications, curating followed accounts without dark patterns resetting preferences—manipulation intensity drops to Low, and no ABT claim lies.¹¹⁴

Hypothetical 4: The Boundary Case—Fitness Influencer (Medium × Medium)

Facts: Riley, age 28, voluntarily follows workout accounts on Instagram. The system recommends similar content, occasionally more extreme. She feels body image pressure but finds some content motivating; no diagnosed disorder.

Matrix Analysis: Manipulation intensity is MEDIUM (user-initiated with some amplification, not extreme). User vulnerability is MEDIUM (adult with agency, but documented pressure). Quadrant: Medium × Medium requires fact-specific analysis; outcome depends on evidence of platform intent and harm.

Dara Kerr, *TikTok Executives Know About App’s Effect on Teens, Lawsuit Documents Allege*, NPR (Oct. 11, 2024) (internal documents from multistate lawsuit showing TikTok employees stated goal is to “contribute to DAU and retention” rather than reduce screen time); see Center for Countering Digital Hate, *TikTok Pushes Harmful Content* (2022); see also ABC News (Dec. 14, 2022), <https://abcnews.go.com/GMA/Family/tiktok-pushes-harmful-content-teens-39-seconds-new/story?id=95357982> (last visited Dec. 23, 2025).

¹¹⁴THE GREEN ZONE LIMITING CASE: THE INFORMED PROFESSIONAL (LOW × LOW). Jennifer, age 45, uses Twitter/X to follow policy analysts. She switched to the chronological “Following” feed, disabled push notifications, and curated 200 accounts. “Suggested” posts can be dismissed; settings persist. She uses the platform 40 minutes daily with no compulsive checking. *Manipulation Intensity:* LOW—meaningful controls exist and function as advertised; genuine agency over content exposure; no dark patterns prevent opt-out. *User Vulnerability:* LOW—adult with no documented risk factors; normal usage parameters; strong offline networks; no platform-induced dependence. QUADRANT: LOW × LOW → No liability; fiduciary duties satisfied. This hypothetical illustrates ABT’s limiting principle: the same platform becomes Low-manipulation when meaningful controls exist and function. ABT does not condemn algorithmic curation as such—it condemns platforms that offer illusory controls while engineering around user choice. If Jennifer’s “Following” feed secretly reintroduced algorithmic ranking, or her notification settings reset without notice, manipulation intensity would shift toward Medium or High. The Green Zone is available to platforms that design for genuine user agency.

Analysis: ABT does not impose absolute liability for all algorithmic curation. Borderline Medium × Medium cases may fail to establish breach, demonstrating ABT’s proportionality and limiting principle.

Judicial Application of the Matrix

The Matrix operates as both analytical tool and procedural mechanism. Its application varies by litigation stage:

At the Pleading Stage: Plaintiff must allege facts sufficient to place defendant’s conduct outside the Green Zone. Allegations establishing only Low × Low placement fail to state an ABT claim. The complaint need not use Matrix terminology, but must allege facts supporting Medium-or-higher manipulation intensity *and* Medium-or-higher user vulnerability. Matrix placement is a legal conclusion requiring factual support—bare assertions of “manipulation” or “vulnerability” without specific design features and user characteristics are insufficient.

At Summary Judgment: The Matrix functions as a gatekeeping mechanism analogous to *Daubert*. Where undisputed material facts establish Green Zone placement, summary judgment for defendant is appropriate—ABT claims require threshold manipulation and vulnerability levels that such facts preclude. Where undisputed facts establish Red Zone placement (High × High or above), breach is presumed and the burden shifts to defendant to establish affirmative defenses or rebut the presumption through evidence of safety investments, meaningful user controls, or non-exploitative intent. Yellow Zone cases—where reasonable jurors could differ on manipulation intensity or vulnerability level—present genuine disputes of material fact requiring trial.

At Trial: Expert testimony informs both Matrix axes. On manipulation intensity: engineers, UX researchers, and behavioral psychologists testify regarding design features,

variable reward implementation, and opt-out functionality. On vulnerability: adolescent development specialists, psychiatrists, and clinical psychologists assess user susceptibility, platform-induced dependence, and harm manifestation. The jury applies the Matrix framework to determine (1) zone placement and (2) in Yellow Zone cases, whether breach occurred on the specific facts. In Red Zone cases, the jury determines whether defendant has rebutted the breach presumption.

On Appeal: Zone placement presents a mixed question of law and fact. The Matrix framework itself is reviewed *de novo* as a legal standard; the underlying factual determinations—what design features existed, what the user’s characteristics were—are reviewed for clear error. Whether particular facts satisfy a given threshold (e.g., whether 5× anger weighting constitutes “High” manipulation) is a legal question subject to *de novo* review, permitting appellate courts to develop consistent standards across cases.

E. Temporal Dimension: Cumulative Harm as Continuing Tort

ABT treats sustained algorithmic manipulation as a continuing tort—prolonged, uninterrupted exposure is a single ongoing wrong. This has three implications:

1. Statute of limitations: Accrual begins when plaintiff knows or should know of injury and its algorithmic cause (discovery rule), not at first exposure.
2. Damages: Cumulative exposure over months or years is relevant to injury severity.
3. Remedies: Forward-looking relief (injunctions, monitoring, algorithmic audits) is often more appropriate than backward-looking damages alone.

F. Evidence and Proof Map

ABT claims require evidence across three domains: (1) general causation—epidemiological studies, quasi-experimental research, and platform internal documents establishing population-level risk;¹¹⁵ (2) specific causation—exposure telemetry and clinical history linking platform use to individual harm; and (3) company knowledge—internal research, A/B test results, and communications documenting platform awareness of harms and rejection of safer alternatives.¹¹⁶

Evidentiary Sufficiency: The Matrix operates as a gatekeeping mechanism. Evidence *supporting* liability includes: internal documents showing safety-for-engagement trade-offs, A/B tests rejecting safer designs, algorithmic specifications targeting vulnerability indicators, and clinical records establishing temporal relationship between platform intensification and symptom onset. Evidence that will *not* support liability includes: bare temporal correlation without algorithmic targeting evidence, general dissatisfaction, content disagreement (ABT targets manipulation mechanics, not objectionable content), and Low $\times$ Low Matrix placement. The Green Zone (Low $\times$ Low) yields no liability; the Yellow Zone requires case-by-case analysis; the Red Zone (High/Extreme $\times$ High/Extreme) triggers presumption of breach.

Damages and Apportionment

¹¹⁵See *In re Silicone Gel Breast Implant Prods. Liab. Litig.*, 318 F. Supp. 2d 879, 893 (C.D. Cal. 2004) (accepting epidemiological evidence for general causation); see also Michael D. Green et al., *Reference Guide on Epidemiology*, in REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 549 (3d ed. 2011) (explaining role of epidemiological evidence in toxic tort litigation).

¹¹⁶See Fed. R. Civ. P. 37(e) (providing for sanctions if electronically stored information that should have been preserved is lost because a party failed to take reasonable steps to preserve it); cf. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 593–94 (1993) (establishing gatekeeping standards for expert testimony); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 141 (1999) (extending *Daubert* to technical and specialized knowledge). ABT’s interdisciplinary expert requirements ensure rigorous, scientifically grounded testimony.

ABT damages follow established tort frameworks: Economic damages (medical expenses, future care, lost earning capacity);¹¹⁷ Hedonic damages (loss of enjoyment, impaired development—particularly serious for minors);¹¹⁸ Dignitary damages (violation of cognitive autonomy itself, compensable like privacy torts);¹¹⁹ and Punitive damages (for Red Zone conduct where platforms possess internal research documenting harm yet reject safer alternatives).¹²⁰

Cognizable Harms: ABT recognizes that cognitive injuries manifest in various forms. Compensable harms include: Clinical diagnoses (depression, anxiety, eating disorders, addiction, post-traumatic stress); Behavioral harms (compulsive use patterns, social withdrawal, academic/occupational impairment); Developmental harms (impaired identity formation, stunted social skills, distorted worldviews—especially for minors); and Severe psychological distress (even without formal diagnosis, substantial emotional distress with functional impairment suffices where causation is established). Clinical diagnosis is preferred but not required; expert testimony establishing substantial impairment of cognitive functioning or emotional well-being is sufficient.

Concrete Application: Alexis Spence (MDL 3047)

¹¹⁷See RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 46 cmt. e (2012) (economic damages include medical expenses and lost earning capacity).

¹¹⁸See *McDougald v. Garber*, 536 N.E.2d 372, 375–76 (N.Y. 1989) (recognizing hedonic damages for loss of enjoyment of life as distinct category); see also RESTATEMENT (SECOND) OF TORTS § 905 cmt. e (1979); see *Cacy v. City of New York*, 512 N.Y.S.2d 1015, 1017 (Sup. Ct. 1987) (recognizing that childhood injuries during critical developmental periods warrant enhanced damages given lifelong impact).

¹¹⁹See RESTATEMENT (SECOND) OF TORTS § 652B cmt. a (1977) (intrusion upon seclusion protects dignitary interest independent of consequential harm); see also NEIL M. RICHARDS, *Intellectual Privacy*, 87 TEX. L. REV. 387, 421–25 (2008) (arguing cognitive manipulation violates fundamental autonomy interest).

¹²⁰See RESTATEMENT (SECOND) OF TORTS § 908(2) (1979) (punitive damages appropriate for reckless or malicious conduct); see also *BMW of N. Am., Inc. v. Gore*, 517 U.S. 559, 575 (1996) (punitive damages serve deterrence and retribution functions).

Alexis Spence alleges that Instagram’s recommendation system steered her toward eating-disorder content beginning at age eleven, contributing to anorexia and repeated hospitalizations.¹²¹ In ABT terms, economic damages would include past and future treatment costs and diminished earning capacity; hedonic damages capture the loss of developmental years; dignitary damages reflect the violation of cognitive autonomy and Meta’s documented knowledge of teen-girl body-image harms; and punitive damages may be available for Red Zone conduct showing knowing disregard of documented risks.¹²²

Apportionment Among Multiple Causes

Mental health harms are multifactorial. ABT adopts several liability: each defendant liable only for its proportional contribution, quantifiable through exposure telemetry.¹²³ De minimis exposure does not meet the substantial factor threshold; substantial regular exposure clearly does.¹²⁴ When multiple platforms contributed, exposure-share or risk-contribution principles from toxic tort apply.¹²⁵ The “eggshell skull” rule ensures vulnerability increases duty

¹²¹Complaint at ¶¶ 12, 15, 18, 22, Spence v. Meta Platforms, Inc., *supra* note 109; Chandler, *supra* note 109; Dembosky, *supra* note 109.

¹²²See RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 46 cmt. e (2012) (economic damages include medical expenses and lost earning capacity); *see* McDougald v. Garber, 536 N.E.2d 372, 375–76 (N.Y. 1989) (recognizing hedonic damages for loss of enjoyment of life as distinct category); *see also* RESTATEMENT (SECOND) OF TORTS § 905 cmt. e (1979); *See* Cacy v. City of New York, 512 N.Y.S.2d 1015, 1017 (Sup. Ct. 1987) (recognizing that childhood injuries during critical developmental periods warrant enhanced damages given lifelong impact); *see* RESTATEMENT (SECOND) OF TORTS § 652B cmt. a (1977) (intrusion upon seclusion protects dignitary interest independent of consequential harm); *see also* NEIL M. RICHARDS, *Intellectual Privacy*, 87 TEX. L. REV. 387, 421–25 (2008) (arguing cognitive manipulation violates fundamental autonomy interest); *See* RESTATEMENT (SECOND) OF TORTS § 908(2) (1979) (punitive damages appropriate for reckless or malicious conduct); *see also* BMW of N. Am., Inc. v. Gore, 517 U.S. 559, 575 (1996) (punitive damages serve deterrence and retribution functions); *see* Wells, Horwitz & Seetharaman, *supra* note 4; *see* Testimony of Frances Haugen, *supra* note 21.

¹²³See RESTATEMENT (THIRD) OF TORTS: APPORTIONMENT OF LIABILITY § 17 (2000) (several liability appropriate where factfinder can reasonably determine proportionate responsibility).

¹²⁴See RESTATEMENT (SECOND) OF TORTS § 431 cmt. a (1965) (“substantial factor” requires material contribution, not de minimis causation).

¹²⁵See *Sindell v. Abbott Labs.*, 607 P.2d 924, 937 (Cal. 1980) (adopting market-share liability where individualized causation impossible but relative contribution measurable); *see* *Thomas v. Mallett*, 701 N.W.2d 523, 555–58 (Wis.

rather than negating causation—platforms targeting users they know to be “insecure” cannot escape liability by arguing predisposition.¹²⁶ Settlement with one platform does not bar claims against others.¹²⁷ Courts have ample tools to value and apportion ABT damages fairly.

PART II: OVERCOMING THE MAJOR LEGAL BARRIERS

Two defenses dominate algorithmic-harms litigation: Section 230 and the First Amendment. ABT survives both only if the claim targets platform conduct—vulnerability targeting and behavioral-conditioning mechanics—rather than third-party content or protected editorial judgment. The path through these doctrines is narrow, and precision is required.

A. Section 230: The Conduct-Content Distinction

For nearly three decades, Section 230 has been the most consequential statute in internet law. Section 230 was written to protect Good Samaritans; platforms have often treated it as near-absolute immunity. Enacted to encourage content moderation, courts interpreted it expansively, immunizing many claims framed as complaints about third-party content.¹²⁸ But the Supreme Court’s 2024 decision in *Moody v. NetChoice* sharpens the doctrinal distinctions recommender-system litigation must confront.

2005) (adopting risk-contribution theory in lead paint litigation where multiple manufacturers contributed to concentrated risk).

¹²⁶See RESTATEMENT (SECOND) OF TORTS § 461 (1965) (“eggshell skull” rule: tortfeasor liable for full extent of harm even if plaintiff’s pre-existing condition makes injury more severe).

¹²⁷See RESTATEMENT (THIRD) OF TORTS: APPORTIONMENT OF LIABILITY § 16 (2000) (settlement with one tortfeasor does not discharge other tortfeasors under several liability).

¹²⁸See, e.g., 47 U.S.C. § 230(c)(1); *Zeran v. Am. Online, Inc.*, 129 F.3d 327, 330 (4th Cir. 1997).

For ABT, the key point is doctrinal, not rhetorical. Section 230 bars treating a platform as the publisher or speaker of information “provided by another information content provider.”¹²⁹ It is not a general immunity for a platform’s own product design. When the gravamen of a claim is the platform’s own conduct—its extraction of behavioral data, inference of vulnerability, and deployment of behavioral-conditioning mechanics—liability need not depend on treating the platform as the publisher of another’s speech.

Navigating the Moody Tension

Section 230 does not immunize platforms for information they themselves create or develop. In *Moody v. NetChoice*, the Supreme Court held that algorithmic content curation constitutes the platform’s *own expressive activity*.¹³⁰ This creates tension: platforms claim their recommendations are their own act (for First Amendment protection) while arguing tort liability for those same recommendations treats them as the publisher of another’s information (for Section 230 immunity). Part II.B addresses *Moody*’s First Amendment implications in detail; here the focus is Section 230.

ABT shifts the gravamen from publication to conduct. The Section 230 question is not whether third-party content exists in the background—it always does—but whether liability requires treating the platform as the publisher of third-party speech. ABT claims target Actionable Behavioral Conditioning: vulnerability detection and engagement-maximizing mechanics operating irrespective of which messages happen to be in the feed. Section 230

¹²⁹47 U.S.C. § 230(c)(1) (“No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.”).

¹³⁰*Moody v. NetChoice, LLC*, 144 S. Ct. 2383, 2401–03 (2024) (holding that platforms’ “compilation[s] of others’ content” through algorithmic curation constitute the platforms’ “own expressive product” entitled to First Amendment protection as editorial discretion).

immunizes platforms for *hosting* content; it does not immunize them for *using* content as a delivery vehicle for covert behavioral modification the platform designed and optimized.¹³¹

The Circuit Split

The circuits are divided on how to apply these principles. In *Anderson v. TikTok*, the Third Circuit held that algorithmic recommendations constitute TikTok’s “own expressive activity”—first-party speech not immunized by Section 230.¹³² The court reasoned that when TikTok’s algorithm curates content, “the platform is speaking, not merely hosting others’ speech.”¹³³

In *M.P. v. Meta Platforms, Inc.*, decided in February 2025, the Fourth Circuit reached the opposite conclusion.¹³⁴ The court held that algorithmic curation remains “integral to the publishing function” and thus receives Section 230 immunity, reasoning that “decisions about whether and how to display certain information provided by third parties are traditional editorial functions of publishers.”¹³⁵

¹³¹See TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 5–18 (2018) (analyzing platform governance); see also Daphne Keller, *Intermediary Liability for Designing and Recommending Content*, STAN. L. SCH. CTR. FOR INTERNET & SOC’Y (2023) (distinguishing platform conduct from content publication).

¹³²*Anderson v. TikTok, Inc.*, 116 F.4th 180, 188–89 (3d Cir. 2024) (holding that TikTok’s algorithmic recommendations constitute the platform’s “own expressive activity”—first-party speech—rather than publication of third-party content immunized by Section 230).

¹³³*Anderson*, 116 F.4th at 189 (“TikTok’s curation and recommendation of videos to a user through the FYP Algorithm is TikTok’s own expressive activity—it is TikTok, not a user, that ‘speaks’ when the FYP Algorithm recommends certain videos to a specific user based on its predictions of what will keep that user engaged.”).

¹³⁴*M.P. v. Meta Platforms, Inc.*, 127 F.4th 516, 524–28 (4th Cir. 2025) (granting Section 230 immunity to Instagram and Facebook for algorithmic recommendation of third-party content, holding that such curation remains “integral to the publishing function”).

¹³⁵*M.P.*, 127 F.4th at 526 (“The decisions about whether and how to display certain information provided by third parties are traditional editorial functions of publishers. That these decisions are now made, in part, by algorithms does not change their essential character.”).

The Fourth Circuit’s reasoning sits in tension with *Moody*’s description of algorithmic curation as the platform’s “own expressive product.” But ABT need not depend on any single resolution of the circuit split. If a plaintiff challenges *which* content a platform chose to amplify, Section 230 arguments are at their strongest. If the plaintiff challenges *how* the platform conditioned behavior through vulnerability targeting and content-agnostic engagement mechanics, Section 230 arguments are at their weakest. ABT is constructed around the latter.

Judge Rushing’s partial dissent in *M.P.* identified the flaw. She argued that Facebook’s recommendation to join a particular group was “undoubtedly Facebook’s own conduct,” noting that “Facebook is not also immune when it conducts statistical analyses of that information and delivers a message based on those analyses.”¹³⁶ ABT targets the same category of conduct: platform *recommendations* based on algorithmic analysis of user vulnerability, not passive display of user-generated content.

Anderson represents the better-reasoned position, and *Moody* provides the key. The Fourth Circuit’s *M.P.* analysis contains an internal contradiction that *Moody* exposes. If algorithmic curation constitutes the platform’s “own expressive activity”—as the *Moody* Court held for First Amendment purposes¹³⁷—then by definition it is not “information provided by another information content provider” within Section 230’s immunity.¹³⁸ The platform cannot simultaneously claim its algorithm is its own protected speech (to invoke First Amendment

¹³⁶*M.P.*, 127 F.4th at 535–36 (Rushing, J., concurring in part and dissenting in part) (arguing that Facebook’s recommendation to join a particular group was “undoubtedly Facebook’s own conduct” and that “Facebook is not also immune when it conducts statistical analyses of that information and delivers a message based on those analyses”).

¹³⁷*Moody v. NetChoice, LLC*, 144 S. Ct. 2383, 2401–03 (2024) (holding that platforms’ “compilation[s] of others’ content” through algorithmic curation constitute the platforms’ “own expressive product” entitled to First Amendment protection as editorial discretion).

¹³⁸47 U.S.C. § 230(c)(1) (“No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.”).

protection) and third-party speech (to invoke Section 230 immunity). *Anderson* recognized this tension: when TikTok’s algorithm curates content, “the platform is speaking, not merely hosting others’ speech.”¹³⁹ That is the logical implication of *Moody*’s holding.

The circuit split creates opportunity rather than obstacle. Supreme Court review of the Section 230 question is increasingly likely given the acknowledged division. When that review comes, *Moody*’s characterization of algorithmic curation as platform expression provides the doctrinal foundation for holding that such expression falls outside Section 230’s immunity for “publishing” another’s speech. ABT is positioned for that resolution.

ABT’s Design: Targeting Tortious Conduct

ABT fits within this doctrinal seam by targeting conduct that is unambiguously the platform’s own. The claim is not: “The platform published harmful content created by a third party.” The claim is: “The platform detected psychological vulnerability in a minor user and exploited that vulnerability through behavioral conditioning techniques designed to maximize engagement at the cost of the user’s well-being.”

This is not merely a publishing function. This is psychological manipulation that *uses* content as a delivery vehicle. Consider: A bookstore may exercise editorial discretion in which books to stock—that is protected. But if the bookstore installs biometric monitoring that detects when customers display signs of psychological distress, then automatically repositions harmful materials in their sightline while obscuring the exits, the bookstore is not exercising editorial

¹³⁹Anderson, 116 F.4th at 189 (“TikTok’s curation and recommendation of videos to a user through the FYP Algorithm is TikTok’s own expressive activity—it is TikTok, not a user, that ‘speaks’ when the FYP Algorithm recommends certain videos to a specific user based on its predictions of what will keep that user engaged.”).

judgment. It is conducting psychological manipulation that happens to occur in a context involving books.

ABT targets the manipulation, not the merchandise. The platform's choice of *what content* to host may be immunized. Its choice to detect depression, exploit that vulnerability, and accelerate harmful content delivery is not. That is the platform's own tortious conduct—and *Moody* confirms it is the platform's own act.

The Narrow Core: Minors in the Red Zone

ABT need not be adopted at its outer boundaries to be workable. Courts can begin with the doctrine's narrowest core: cases involving minors (or users the platform knows to be acutely vulnerable) combined with Red Zone conduct (High or Extreme manipulation). In that core, the alleged wrong is content-agnostic conditioning mechanics plus vulnerability targeting supported by internal research and telemetry—not mere publication of third-party speech. This incremental path preserves Section 230's central function while permitting accountability for the most egregious, platform-designed behavioral exploitation.

B. First Amendment: Content-Neutral Regulation of Conduct

The First Amendment question is whether algorithmic manipulation constitutes protected editorial discretion. It does not—and ABT's design ensures constitutional survival. In *Moody v. NetChoice*, the Supreme Court held that algorithmic content curation constitutes protected editorial discretion, even when “exercised through algorithms or other automated systems.”¹⁴⁰

¹⁴⁰*Moody*, 144 S. Ct. at 2403 (“The principle does not change because the curation is conducted through an algorithm, rather than by a person. Deciding whether to allow certain content on a feed, or how to present it, is editorial discretion, regardless of whether it is exercised through algorithms or other automated systems.”).

Platforms will argue this forecloses ABT entirely. That conclusion does not follow; ABT survives only to the extent it targets content-neutral manipulation mechanics rather than editorial judgment about ideas.

The Moody Challenge

Moody protects editorial discretion over *content*—decisions about which ideas to propagate, which viewpoints to amplify, which messages to exclude. The Court analogized platform curation to the parade organizer in *Hurley v. Irish-American Gay, Lesbian and Bisexual Group of Boston*, who exercised protected discretion by choosing which groups could march.¹⁴¹ Under this analysis, algorithmic curation is expression, and expression receives First Amendment protection.

ABT does not challenge this holding. A platform may curate a conservative feed, a progressive feed, an entertainment feed, or any ideological mix. ABT is indifferent to the ideas platforms choose to propagate. What ABT prohibits is different: the exploitation of detected psychological vulnerabilities through behavioral conditioning techniques that operate *independently* of any editorial judgment about content.

Variable reward schedules do not select content; they hijack attention. Infinite scroll does not recommend posts; it eliminates stopping points. Vulnerability detection does not curate feeds; it targets users. These are not editorial choices about ideas. They are engineering decisions about psychological manipulation.¹⁴²

¹⁴¹*Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos.*, 515 U.S. 557, 573–74 (1995) (holding that parade organizer’s selection of participants constitutes protected First Amendment expression because “a private speaker does not forfeit constitutional protection simply by combining multifarious voices” into a compilation that “is a form of expression”).

¹⁴²See Andrew McStay, *Emotional AI and EdTech: Serving the Public Good?*, 40 LEARNING, MEDIA & TECH. 301 (2020) (analyzing emotional detection technology and its implications for autonomy); see also SHERRY TURKLE,

The Barrett Pivot: ABT's Constitutional Lifeline

Justice Barrett's concurrence in *Moody* provides the doctrinal opening ABT requires. She explicitly reserved the question ABT raises, distinguishing between algorithms that "implement human beings' inherently expressive choice" and those that operate differently.¹⁴³

"But what if a platform's algorithm just presents automatically to each user whatever the algorithm thinks the user will like—e.g., content similar to posts with which the user previously engaged? The First Amendment implications... might be different for that kind of algorithm."

Barrett asked the critical question: "Technology may attenuate the connection between content-moderation actions and human beings' constitutionally protected right to 'decide for [themselves] the ideas and beliefs deserving of expression.'" ¹⁴⁴ When an algorithm operates without "human-devised standards" about which ideas deserve propagation—when it simply maximizes engagement regardless of content—the human editorial judgment that justifies First Amendment protection is absent.

ABT's targets fall squarely within Barrett's unprotected category. Engagement-maximizing algorithms do not implement editorial standards about ideas. They implement a single, content-agnostic metric: maximize time on platform. If anger drives engagement, anger dominates. If anxiety drives engagement, anxiety dominates. The "editorial choice" is simply:

ALONE TOGETHER: WHY WE EXPECT MORE FROM TECHNOLOGY AND LESS FROM EACH OTHER 293–310 (2011) (documenting how digital mediation affects authentic connection).

¹⁴³*Moody*, 144 S. Ct. at 2408 (Barrett, J., concurring) ("But what if a platform's algorithm just presents automatically to each user whatever the algorithm thinks the user will like—e.g., content similar to posts with which the user previously engaged? The First Amendment implications of the laws might be different for that kind of algorithm.").

¹⁴⁴*Id.* ("Technology may attenuate the connection between content-moderation actions and human beings' constitutionally protected right to 'decide for [themselves] the ideas and beliefs deserving of expression, consideration, and adherence.'" (quoting *Turner Broad. Sys., Inc. v. FCC*, 512 U.S. 622, 641 (1994))).

more engagement. This is optimization of a business metric that happens to use content as fuel. It is not a choice about ideas deserving expression.¹⁴⁵

The Hurley Distinction

The *Moody* majority invoked *Hurley*, where a parade organizer with a *message* exercised protected discretion over which groups could march.¹⁴⁶ Engagement-maximizing systems have no message—they are indifferent to content except insofar as content serves the metric. A system that amplifies outrage does not *intend* to communicate outrage; it is optimized such that outrage tends to win because it drives engagement. The difference is between a curator with a vision and a machine with a metric. ABT regulates only the latter.

The Three-Part Test: When ABT Survives Constitutional Scrutiny

ABT claims survive First Amendment scrutiny when—and only when—the challenged conduct satisfies three criteria:

First, the challenged conduct must operate independently of content selection. Variable reward schedules, infinite scroll, autoplay, and notification timing function identically regardless of what content is displayed. These features manipulate attention through psychological mechanisms unrelated to the ideas conveyed. They would create compulsive use patterns

¹⁴⁵See YOCHAI BENKLER, ROBERT FARIS & HAL ROBERTS, NETWORK PROPAGANDA: MANIPULATION, DISINFORMATION, AND RADICALIZATION IN AMERICAN POLITICS 21–47 (2018) (analyzing information ecosystem dynamics); ELI PARISER, THE FILTER BUBBLE: WHAT THE INTERNET IS HIDING FROM YOU 7–19 (2011) (introducing personalization concerns).

¹⁴⁶*Hurley*, 515 U.S. at 574 (noting that the parade organizer “select[ed] contingents to make the parade more combative or combative-free, optimistic or pessimistic, comic or combative,” reflecting expressive choices about the parade’s message rather than content-agnostic optimization of any particular metric).

whether the content were political, entertainment, educational, or random. Content-agnostic manipulation cannot be protected as content-based expression.¹⁴⁷

Second, the challenged conduct must target detected user vulnerabilities rather than content preferences. When a platform identifies that a user is showing signs of depression—or is “insecure” or “worthless,” per Facebook’s advertising pitches—and adjusts the information environment to exploit that vulnerability, the platform is not exercising editorial discretion about ideas. It is conducting psychological targeting of *users*. The subject of the platform’s attention is the user’s mental state, not the content’s message.¹⁴⁸

Third, the challenged conduct must cause harm regardless of specific content delivered. If manipulation mechanics would cause harm whether the content were political news, fitness videos, or cat memes, then the harm flows from the mechanics, not the expression. The mechanism—not the message—is tortious. Regulating content-agnostic mechanisms is content-neutral regulation subject to intermediate scrutiny.¹⁴⁹

Where all three criteria are met, ABT regulates conduct that even *Moody* did not protect—conduct that, in Justice Barrett’s framing, operates “without reliance upon human-devised standards” about ideas deserving expression.

¹⁴⁷*Arcara v. Cloud Books, Inc.*, 478 U.S. 697, 706–07 (1986) (holding that conduct regulation not subject to First Amendment scrutiny when “it was conduct with a significant expressive element that drew the legal remedy in the first place,” and regulation of nonexpressive conduct does not implicate First Amendment even if it incidentally affects speech). Variable reward schedules and infinite scroll features operate identically regardless of content, distinguishing them from editorial judgment about ideas.

¹⁴⁸FTC unfairness authority extends to practices causing substantial injury that consumers cannot reasonably avoid and that is not outweighed by benefits. *See* 15 U.S.C. § 45(n). Targeting users based on detected psychological vulnerabilities—rather than content preferences—constitutes conduct-based manipulation, not expression. The platform’s “speech” is directed at the vulnerability, not conveying ideas.

¹⁴⁹Content-neutral regulations survive intermediate scrutiny. *See Ward v. Rock Against Racism*, 491 U.S. 781, 791 (1989). ABT targets manipulation mechanics that cause harm through their psychological operation, not through the ideas they convey. A platform using addiction mechanics on educational content violates ABT; one using transparent chronological feeds for extreme content does not. The harm derives from *how* content is delivered, not *what* is delivered.

Intermediate Scrutiny: The O'Brien Analysis

Even if algorithmic curation receives some First Amendment protection, ABT survives intermediate scrutiny under *United States v. O'Brien*.¹⁵⁰ The framework permits regulation of conduct with expressive components when the regulation: (1) is within constitutional power; (2) furthers an important or substantial governmental interest; (3) is unrelated to the suppression of free expression; and (4) restricts speech no more than necessary.¹⁵¹

Constitutional Power: States possess traditional police power to protect children. *See Ginsberg v. New York* (states may adjust protections for minors); *Ferber* (child protection justifies otherwise unconstitutional restrictions).¹⁵²

Substantial Interest: The Surgeon General's 2023 Advisory declared youth mental health a national crisis.¹⁵³ Platform internal research confirms *causation*—Meta's own studies quantified Instagram's harm to adolescent girls.¹⁵⁴

Unrelated to Expression: ABT's triggers are content-agnostic. Addiction mechanics operate identically regardless of content. A platform serving cat videos through addiction

¹⁵⁰*United States v. O'Brien*, 391 U.S. 367, 377 (1968) (establishing four-part test for regulations affecting conduct with expressive elements: regulation must be within constitutional power, further important governmental interest, be unrelated to suppression of free expression, and restrict no more speech than necessary).

¹⁵¹*See also* *Turner Broad. Sys., Inc. v. FCC*, 512 U.S. 622, 662 (1994) (applying intermediate scrutiny to content-neutral cable regulations); *Ward v. Rock Against Racism*, 491 U.S. 781, 791 (1989) (content-neutral time, place, manner restrictions survive intermediate scrutiny when narrowly tailored to significant governmental interests).

¹⁵²*Ginsberg v. New York*, 390 U.S. 629, 638–39 (1968) (upholding variable obscenity standard permitting greater restrictions on minors); *New York v. Ferber*, 458 U.S. 747, 756–57 (1982) (recognizing compelling state interest in protecting children justifies speech restrictions that would be unconstitutional for adults).

¹⁵³*See* Surgeon General's Advisory, *supra* note 2, at 4–9 (declaring adolescent mental health crisis and identifying social media as contributing factor requiring urgent public health response); *see* CENTERS FOR DISEASE CONTROL AND PREVENTION, YOUTH RISK BEHAVIOR SURVEY DATA SUMMARY & TRENDS REPORT 2011–2021, at 49–52 (2023) (documenting 50% increase in persistent feelings of sadness or hopelessness among high school students; 51% increase in female adolescent suicide attempts).

¹⁵⁴*See* Wells, Horwitz & Seetharaman, *supra* note 4 (reporting internal Facebook research finding “We make body image issues worse for one in three teen girls” and “Teens blame Instagram for increases in the rate of anxiety and depression”).

mechanics violates ABT; one serving extreme content through non-manipulative design does not.

Narrow Tailoring: ABT leaves unlimited alternatives—chronological feeds, non-exploitative recommendations, adult optimization with consent. Platforms retain discretion over *what* to show; they lose only the right to psychologically manipulate delivery to children.

ABT's Limiting Principles

Claims targeting editorial discretion (*which* content to amplify) fail—*Moody* protects those choices. Claims targeting manipulation mechanics (vulnerability exploitation, addiction features, concealment of known harm, minor-specific addictive design) survive because they address content-agnostic psychological manipulation, not editorial judgment. The Matrix operationalizes this distinction.

The Candor Duty

ABT's duty of candor requires disclosure of material conflicts, algorithmic mechanics, and known risks. *Moody* preserved space for transparency mandates.¹⁵⁵ Under *Zauderer*, compelled disclosure of “purely factual and uncontroversial information” in commercial contexts survives First Amendment scrutiny.¹⁵⁶ ABT's required disclosures are factual (algorithm ranking, data collection, internal research), uncontroversial (platforms advertise these capabilities

¹⁵⁵*Moody*, 144 S. Ct. at 2408–09 (noting that the Court's analysis did not disturb Texas provisions requiring platforms to provide explanations for content moderation decisions, preserving regulatory space for transparency mandates distinct from editorial control).

¹⁵⁶*Zauderer v. Office of Disciplinary Counsel*, 471 U.S. 626, 651 (1985) (holding that compelled disclosure of “purely factual and uncontroversial information” in commercial contexts survives First Amendment scrutiny when “reasonably related to the State's interest in preventing deception of consumers”).

to advertisers), and reasonably related to preventing harm.¹⁵⁷ The pharmaceutical parallel is precise: FDA risk disclosures are unquestioned.

The Slot Machine Analogy

Casinos cannot deploy slot-machine designs targeting minors—not because of the machine’s “message” but because of its manipulative mechanics.¹⁵⁸ ABT applies the same principle. *Tornillo* protects what newspapers *say*. It does not protect how they design the ink. The algorithm is the addictive ink. The platform’s editorial choices are expressive. Its deployment of dopamine-hijacking mechanics to create compulsive use in children is conduct—and regulable conduct at that.

C. Anticipated Defenses

Platforms will raise predictable defenses. None withstands scrutiny under ABT.¹⁵⁹

Consent and assumption of risk fails because consent requires knowledge of what is being assumed—you cannot consent to deliberately concealed manipulation. Minors lack capacity for knowing consent to cognitive exploitation.¹⁶⁰ The “just log off” intuition—that users

¹⁵⁷*NIFLA v. Becerra*, 138 S. Ct. 2361, 2371–72 (2018) (striking down compelled ideological speech unrelated to services provided, while distinguishing *Zauderer*-type disclosures of factual information about commercial services); *See* 21 C.F.R. § 201.57 (FDA drug labeling requirements); *Sorrell v. IMS Health Inc.*, 564 U.S. 552, 579 (2011) (distinguishing commercial speech regulations from content-based speech restrictions). The analogy between pharmaceutical disclosure and algorithmic transparency requirements is doctrinally precise.

¹⁵⁸*See, e.g.,* Schüll, *Addiction by Design*, *supra* note 19, at 52–78 (2012) (documenting regulatory restrictions on slot machine designs targeting psychological vulnerabilities through variable reward schedules, near-miss effects, and other manipulation techniques); *see also* 15 U.S.C. § 1175 (prohibiting gambling devices in certain contexts regardless of expressive content).

¹⁵⁹The analysis that follows distills six categories of anticipated defenses, each of which fails to defeat ABT claims in High × High or Extreme × Extreme Matrix quadrants. Platform counsel will predictably raise these arguments; courts should recognize their fundamental mismatch with ABT’s structure and purpose.

¹⁶⁰CONSENT AND ASSUMPTION OF RISK: RESTATEMENT (SECOND) OF TORTS § 892B cmt. b (1979) establishes consent is ineffective if obtained through misrepresentation or concealment of material facts. Platforms’ Terms of Service do not disclose algorithmic detection of adolescent vulnerability or amplification of harmful content. *See*

who dislike platforms can simply leave—founders on the realities of platform design. Network effects make exit costly: a teenager’s social life, school communications, and peer relationships exist on the platform; departure means social exile.¹⁶¹ Addiction mechanics make exit difficult: variable reward schedules, notification systems, and infinite scroll are specifically designed to prevent departure. Most fundamentally, a platform cannot design systems that create compulsive use and then invoke user choice as a defense. A casino that drugs patrons to keep them gambling cannot claim they freely chose to stay.¹⁶² Platforms that identify vulnerable users, exploit their psychological susceptibilities, and deploy addiction mechanics to maximize engagement have forfeited the assumption of risk defense. Users cannot “assume” risks they were never told about, cannot consent to manipulation deliberately concealed, and cannot freely choose when choice itself has been engineered away.

Editorial discretion misapprehends ABT’s target. ABT regulates manipulation mechanics, not ideas. Justice Barrett’s *Moody* concurrence distinguishes human editorial judgment from algorithms that “automatically generate outputs based on a set of rules”—engagement optimization falls in the latter category.¹⁶³

Helling v. Carey, 519 P.2d 981, 983 (Wash. 1974) (assumption of risk unavailable where defendant creates inherent risk plaintiff cannot reasonably avoid). Design-induced addiction through variable rewards and infinite scroll negates voluntariness. Minors lack capacity for knowing consent to cognitive manipulation. See RESTATEMENT (SECOND) OF TORTS § 892A(2) (1979); Parham v. J.R., 442 U.S. 584, 602–03 (1979). Network effects and platform lock-in make exit structurally impossible for adolescents whose social life migrates to platforms. See Balkin, Information Fiduciaries, *supra* note 101. Assumption of risk may apply in Low × Low quadrants (minimal manipulation of informed adults) but fails where ABT’s core concerns arise.

¹⁶¹See JONATHAN HAIDT, THE ANXIOUS GENERATION: HOW THE GREAT REWIRING OF CHILDHOOD IS CAUSING AN EPIDEMIC OF MENTAL ILLNESS 156–62 (2024) (documenting network effects and social costs of platform departure for adolescents).

¹⁶²*Cf.* Schüll, Addiction by Design, *supra* note 19, at 52–78 (2012) (documenting how casinos design environments to prevent patron departure through variable reward schedules and environmental manipulation).

¹⁶³ *Moody*, 144 S. Ct. at 2408 (Barrett, J., concurring).

Parental responsibility does not defeat ABT claims. Platform wrongdoing is independent of parental supervision; comparative fault does not apply when injury transcends alleged comparative negligence.¹⁶⁴

Business judgment rule does not shield knowing harm. Platform directors face disqualifying conflicts (engagement profit vs. user well-being); internal research documenting harm constitutes bad faith.¹⁶⁵

Federal preemption finds no support in statutory text. COPPA addresses data collection, not manipulation. ABT operates in traditional state tort territory.¹⁶⁶

Standing exists where platforms have induced addiction. The compulsion itself is the ongoing injury.¹⁶⁷

¹⁶⁴ See RESTATEMENT (THIRD) OF TORTS: APPOINTMENT OF LIABILITY § 7 cmt. m (2000) (comparative responsibility inapplicable where defendant's conduct is intentional or reckless).

¹⁶⁵ *Garfield v. Allen*, 277 A.3d 296, 336 (Del. Ch. 2022) (holding that business judgment rule does not apply when directors knowingly exceed authority or act in bad faith, emphasizing that rule protects discretionary judgments only when they fall within directors' authority and are made in good faith); *Kahn ex rel. DeKalb Genetics Corp. v. Roberts*, 679 A.2d 460 (Del. 1982) (clarifying that business judgment rule protects lawful actions taken in good faith and without conflicts of interest, but when directors act in response to threat to their control, heightened judicial scrutiny applies to ensure actions not self-interested).

¹⁶⁶ FEDERAL PREEMPTION: 15 U.S.C. § 6501 *et seq.* (COPPA) addresses data collection from children under 13, not psychological manipulation or cognitive integrity violations. No express preemption language exists. See *Silkwood v. Kerr-McGee Corp.*, 464 U.S. 238, 251 (1984) (federal regulatory schemes do not preempt state tort law absent express congressional intent). State UDAP laws address deceptive trade practices; ABT addresses dignitary and psychological harms from manipulative conduct. These are complementary regimes operating in different domains. Products liability, medical malpractice, and other state torts routinely coexist with federal regulation. Section 230, as discussed in Part II.A, does not apply to platform conduct claims.

¹⁶⁷ STANDING FOR INJUNCTIVE RELIEF: Article III standing requires injury in fact, causation, and redressability. See *Lujan v. Defenders of Wildlife*, 504 U.S. 555, 560–61 (1992). Where platforms deliberately induce compulsive use disorder, future harm is imminent absent relief—telling addicted users to “just stop” is like telling drug addicts to exercise willpower. The addiction itself is the ongoing injury. See *United States v. Trucking Empls., Inc.*, 561 F.2d 313, 317–18 (D.C. Cir. 1977) (injunctive relief appropriate where defendant's conduct creates ongoing harm plaintiff cannot self-remedy). For minors, structural barriers (network effects, peer pressure, developmental immaturity) make self-remedy impossible. Injunctive relief protecting vulnerable populations from ongoing systemic harm is a core equitable remedy. See *Brown v. Plata*, 563 U.S. 493, 511 (2011).

Floodgates concerns are answered by ABT's limiting principles: Matrix gatekeeping, substantial factor thresholds, genuine injury requirements, and demanding evidentiary burdens.¹⁶⁸

Overbreadth ("Gmail Problem") is avoided by ABT's three-part test: liability only for conduct (1) operating independently of content, (2) targeting vulnerabilities, and (3) causing content-agnostic harm. Gmail, Venmo, and Etsy serve user interests and do not exploit psychological vulnerabilities.

PART III: PLATFORM-SPECIFIC APPLICATIONS

Theory becomes tangible through application. This Part shows how ABT analyzes the documented design choices of four major platforms across varied fact patterns. The applications progress from direct emotional amplification (Facebook) through content velocity escalation (TikTok) to identity-based targeting (Instagram) and ideological transformation (YouTube), illustrating ABT's breadth and practical applicability.

¹⁶⁸FLOODGATES AND INDETERMINACY: ABT contains multiple structural limitations preventing boundless liability. First, the Matrix eliminates Low × Low quadrant claims (minimal manipulation of non-vulnerable users). Second, the substantial factor threshold bars de minimis exposure claims—trivial platform use does not meet the materiality requirement. *See* RESTATEMENT (SECOND) OF TORTS § 431 cmt. a (1965). Third, the harm requirement demands genuine psychological injury (clinical diagnosis or substantial functional impairment); mere dissatisfaction or disagreement with algorithmic choices does not suffice. Fourth, the proof burden is demanding: sophisticated interdisciplinary expert testimony, discovery of internal platform documents, and analysis of exposure telemetry data. This is expensive, complex litigation, not a strike suit. These limitations confine ABT to its proper domain: egregious exploitation of vulnerable users causing serious, documented psychological harm. Other torts have successfully implemented similar limiting principles to prevent liability proliferation. *See* *Thing v. La Chusa*, 771 P.2d 814, 826–30 (Cal. 1989) (limiting bystander emotional distress claims to close family members present at scene, observing that “in order to avoid limitless liability out of all proportion to the degree of a defendant’s negligence,” recovery must be cabined by clear rules); *Palsgraf v. Long Island R.R.*, 162 N.E. 99 (N.Y. 1928) (Cardozo, C.J.) (limiting foreseeable plaintiff rule to those within zone of danger); RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 47 (2010) (confining direct infliction emotional distress claims to danger of immediate bodily harm or specified relationship categories). ABT’s Matrix performs the identical gatekeeping function.

A. Facebook/Meta: The Anger Amplification Engine

Day in the Feed: Open the app. A post with 500 angry reactions sits on top. Your friend comments; you click; the replies are hotter. The room gets louder.

Design: Facebook’s News Feed ranking system assigns five times greater weight to “angry” reactions than “likes.”¹⁶⁹ The company conducted experiments on 700,000 users proving “massive-scale emotional contagion.”¹⁷⁰ Data scientists warned Meaningful Social Interactions (MSI) weighting was making the platform “angrier, more polarizing”¹⁷¹—leadership rejected safer alternatives because they would decrease engagement.¹⁷²

Matrix Analysis: High Manipulation × High Vulnerability → Red Zone. Presumption of breach applies.

ABT Claim Structure: Duty to billions, especially minors. Breach through loyalty (engagement profit over well-being) and care (failed to implement safer design). Causation through platform’s own research linking amplified outrage to anxiety, depression, and polarization. Harm through psychological distress and relationship breakdown.

Forensic note: To advertisers, “precision emotional targeting.” To the board, “engagement optimization.” To courts, “we can’t control our own systems.”¹⁷³

¹⁶⁹See Hao, *supra* note 3.

¹⁷⁰See Kramer et al., *supra* note 52.

¹⁷¹See Haugen, *supra* note 21.

¹⁷²Jeremy Merrill & Will Oremus, *More internal documents show how Facebook’s algorithm prioritized anger*, WASH. POST (Oct. 25, 2021).

¹⁷³*Cf.* Testimony of Mark Zuckerberg, *Hearing on Cambridge Analytica and Facebook*, S. Comm. on Com., Sci., and Transp., 115th Cong. (2018) (statements reflecting a claimed inability to fully control the platform’s operations).

Remedies: Injunctive relief prohibiting disproportionate anger weighting. Compensatory damages for psychological harm. Punitive damages for documented internal knowledge plus conscious rejection of safer alternatives.

B. TikTok: The Rabbit-Hole Algorithm

Day in the Feed: Search “sad.” Pause for three seconds on a confessional clip. Within minutes, the “For You” page starts offering darker themes and hashtags. You keep watching because it now feels personal.

Design: TikTok’s “For You” page (FYP) represents hyper-personalized algorithmic curation at its extreme. The system learns from every micro-behavior: gaze duration, video replays, scrolling pauses.¹⁷⁴ Academic research confirms that TikTok’s system learns from granular behavioral signals—gaze duration, replays, scrolling pauses—to personalize recommendations in ways that sustain compulsive engagement.¹⁷⁵ Its primary function is creating “rabbit holes”—once it infers nascent interest, it floods the feed with increasingly intense content.

Evidence of Harm: Amnesty International’s 2023 audit concluded TikTok’s design is “manipulative and addictive,” with test accounts posing as 13-year-olds drawn into self-harm content within minutes.¹⁷⁶ The Center for Countering Digital Hate documented TikTok pushing

¹⁷⁴Steffen Zannettou et al., *Analyzing User Engagement with TikTok’s Short Format Video*, in PROCEEDINGS OF THE 2024 ACM WEB CONFERENCE 1–12 (2024).

¹⁷⁵*Id.*

¹⁷⁶AMNESTY INTERNATIONAL, *DRIVEN INTO DARKNESS* (2023); AMNESTY INTERNATIONAL, *TIKTOK FAILS TO ADDRESS RISKS* (2023), <https://www.amnestyusa.org/reports/tiktok-fails-to-address-risks-to-children-and-young-peoples-mental-health-despite-past-warnings/> (last visited Dec. 23, 2025).

suicide content to teen accounts in as little as 2.6 minutes.¹⁷⁷ A 2024 U.K. study found the recommendation system amplified misogynistic content four-fold in five days, “actively indoctrinating” boys.¹⁷⁸ In “blackout challenge” lawsuits, *Anderson v. TikTok* (3d Cir. 2024) held that algorithmic promotion of deadly content is the platform’s own conduct—Section 230 does not bar these claims.¹⁷⁹

Matrix Analysis:

Manipulation Intensity: EXTREME: Extreme sensitivity to subconscious signals, harmful feedback loops at unprecedented speed, reward-system optimization for compulsive use

User Vulnerability: EXTREME: User base predominantly young minors in critical cognitive-emotional development stages

Result: Extreme × Extreme quadrant → Clearest possible liability case

ABT Claim Structure:

Duty: Heightened fiduciary duty to minor users

Breach: Designing systems that identify and exploit developmental vulnerabilities; deploying lethal content recommendations to children

Causation: Direct causal chain—recommendation → user viewing → user attempting → death

Harm: Death (most extreme imaginable harm)

¹⁷⁷CENTER FOR COUNTERING DIGITAL HATE, TIKTOK PUSHES HARMFUL CONTENT (2022); *see also* ABC News (Dec. 14, 2022), <https://abcnews.go.com/GMA/Family/tiktok-pushes-harmful-content-teens-39-seconds-new/story?id=95357982> (last visited Dec. 23, 2025).

¹⁷⁸KAITLYN REGEHR ET AL., SAFER SCROLLING (UCL/ASCL 2024), <https://www.ascl.org.uk/ASCL/media/ASCL/Help%20and%20advice/Inclusion/Safer-scrolling.pdf> (last visited Dec. 23, 2025).

¹⁷⁹*See Anderson v. TikTok, Inc.*, *supra* note 132.

Forensic note: A system that can identify a vulnerable teen within minutes took years to face accountability. ABT fixes that timeline.

Zone Classification and Remedies: TikTok’s rabbit-hole system sits at Extreme Manipulation × Extreme Vulnerability—the clearest ABT case. Injunctive relief should prohibit recommendation systems that escalate based on distress signals and ban FYP for minors without structural rabbit-hole interventions. In death cases, wrongful death damages apply. Punitive damages warranted: knowledge of 2.6-minute harm pathways + continued deployment + minor targeting = reckless disregard.

C. Instagram: The Body Image Distortion Field

Day in the Feed: Follow one fitness account. Like two diet posts. By day three, the Explore grid is extreme deficits, thigh gaps, and calorie-burn hacks. Your friends’ likes become your mirror.

Design: Instagram’s visual-content recommendation system infers user interests and connects them with similar content. Interest in diet or exercise is treated as appetite for increasingly extreme material, leading users from mainstream fitness toward pro-anorexia (“pro-ana”) and self-harm networks.

Evidence of Harm: Meta’s internal research found “we make body image issues worse for 1 in 3 teen girls”¹⁸⁰—then buried the findings and continued unchanged, as Haugen’s testimony documented.¹⁸¹ Independent 2024 research corroborated Meta’s findings: the recommendation system amplifies eating disorder content at rates exceeding 4,300% for

¹⁸⁰See Wells, Horwitz & Seetharaman, *supra* note 4.

¹⁸¹See Haugen, *supra* note 21.

vulnerable users.¹⁸² A Danish study found Instagram “actively contributes to the formation and spread of self-harm networks” by linking vulnerable teens into pro-ana clusters.¹⁸³ Eating disorders carry the highest mortality rate of any mental illness; anorexia sufferers are five times more likely to die than peers.¹⁸⁴

Matrix Analysis:

Manipulation Intensity: HIGH: Systematic exploitation of adolescent social comparison psychology, quantified validation (likes) as variable reward, active promotion of harmful networks

User Vulnerability: HIGH: Teenage girls at peak susceptibility to body image issues and eating disorders

Result: High × High quadrant → Presumption of breach with punitive exposure

ABT Claim Structure:

Duty: Heightened duty to minor users

Breach: Loyalty (deployed algorithms known to worsen body image), Care (failed to filter pro-eating-disorder content), Candor (actively concealed findings)

Causation: General (Meta’s research), specific (user data documenting algorithmic pathway into pro-anorexia echo chambers)

Harm: Clinically diagnosed eating disorder and attendant consequences

¹⁸²Aislinn Rowan, Miriam Redi & Claudia Stöckl, *Does TikTok Contribute to Eating Disorders? A Comparison of the TikTok Algorithms Belonging to Individuals with Eating Disorders versus Healthy Controls*, 124 APPETITE 106937 (2024).

¹⁸³Miranda Bryant, *Instagram actively helping spread of self-harm*, THE GUARDIAN (Nov. 30, 2024), <https://www.theguardian.com/technology/2024/nov/30/instagram-actively-helping-to-spread-of-self-harm-among-teenagers-study-suggests> (last visited Dec. 23, 2025).

¹⁸⁴Jon Arcelus et al., *Mortality Rates in Patients With Anorexia Nervosa*, 68 ARCH. GEN. PSYCHIATRY 724 (2011).

Forensic note: Internal slide deck, page 15: “We make body image issues worse for 1 in 3 teen girls.” Page 16: “We need to better understand...” Translation: we knew; we studied; we kept the design.

Zone Classification and Remedies: Instagram’s body image distortion sits at High Manipulation × High Vulnerability in the Red Zone. Concealment of internal research elevates to punitive territory. Injunctive relief should prohibit Explore algorithms amplifying eating disorder content to vulnerable users and recommending pro-eating-disorder networks. Economic damages include hospitalization plus ongoing care. Punitive damages warranted: documented harm + active concealment + unchanged deployment = reckless disregard.

D. YouTube: The Radicalization Pipeline

Day in the Feed: Watch one dramatic clip about social decline. Autoplay offers something edgier. “Next up” is tuned to keep you there. After an hour, the tone has shifted. So have you.

Design: YouTube’s recommendation system drives over seventy percent of total platform watch time,¹⁸⁵ optimizing primarily for “watch time.” The system learns user interests then recommends increasingly extreme or sensational content, creating a “radicalization pipeline” systematically funneling users toward extremist ideologies and misinformation.

Evidence of Harm: A 2025 experimental study found YouTube’s recommendation system systematically amplifies negative emotions—especially anger—creating “emotional filter

¹⁸⁵See, e.g., Kevin Roose, *The Making of a YouTube Radical*, N.Y. TIMES (June 8, 2019) (explaining that “more than 70 percent of the time users spend on YouTube is driven by its recommendations”).

bubbles” that escalate outrage over time.¹⁸⁶ Former engineer Guillaume Chaslot reported internal proposals to “tame” extremist tendencies were rejected because they would reduce watch time—initiatives cutting viewing by approximately one percent were shelved.¹⁸⁷ Academic studies document users moving from mainstream videos to extremist content via YouTube’s recommendations.¹⁸⁸

Matrix Analysis:

Manipulation Intensity: HIGH: Sophisticated opaque recommendation systems, documented echo-chamber creation, auto-play features encouraging passive prolonged consumption

User Vulnerability: MEDIUM: Extremely broad user base; heightened to HIGH for minors and users with known susceptibility

Result: High × Medium quadrant → Viable claims; for minors, High × High triggers presumption of breach

ABT Claim Structure:

Duty: Fiduciary duty to users, heightened for minors

Breach: Designing algorithms foreseeably leading users down radicalization pathways to maximize watch time

¹⁸⁶Hussam Habib & Rishab Nithyanand, *YouTube Recommendations Reinforce Negative Emotions* (2025), <https://arxiv.org/html/2501.15048v1> (last visited Dec. 23, 2025).

¹⁸⁷See, e.g., Guillaume Chaslot, *How YouTube’s Algorithm is Politicizing Us*, THE GUARDIAN (Oct. 2, 2019), <https://www.theguardian.com/commentisfree/2019/oct/02/how-youtubes-algorithm-is-politicizing-us> (last visited Dec. 23, 2025); see also Zeynep Tufekci, *YouTube, the Great Radicalizer*, N.Y. TIMES (Mar. 10, 2018); see Jack Nicas, *How YouTube Drives People to the Internet’s Darkest Corners*, WALL ST. J. (Feb. 7, 2018) (reporting on internal discussions where proposals to reduce the spread of conspiratorial videos were rejected due to concerns about their impact on engagement metrics).

¹⁸⁸See, e.g., Manoel Horta Ribeiro et al., *Auditing Radicalization Pathways on YouTube*, PROC. ACM ON HUMAN-COMPUTER INTERACTION (CSCW 2020); Kevin Roose, *The Making of a YouTube Radical*, N.Y. TIMES (June 8, 2019).

Causation: Documentation through user data of specific algorithmic pathway from mainstream to extremist content; expert testimony establishing content diet as substantial factor

Harm: Psychological (severe paranoia, ideological fixation) or physical (injuries from violent acts by radicalized users)

Forensic note: Proposals to turn down sensationalism were internally called “the taming of the algorithm.” They would have reduced watch time by approximately one percent. They didn’t ship. That’s the grim calculus.

Zone Classification and Remedies: YouTube’s radicalization pipeline sits at High Manipulation × Medium Vulnerability—Yellow-to-Red Zone depending on user. For minors and users with known susceptibility, analysis shifts firmly into the Red Zone (High × High). Injunctive relief should prohibit autoplay for minors without affirmative opt-in and block recommendation algorithms escalating content extremity based on emotional engagement. Compensatory damages include deprogramming therapy, hedonic damages for relationship loss, and dignitary damages. Punitive damages warranted: documented safer alternatives rejected for minimal watch time preservation = reckless disregard.

PART IV: INTERNATIONAL VALIDATION

ABT’s principles have achieved remarkable convergence with independent regulatory initiatives across the democratic world, validating both the feasibility and policy wisdom of this approach.

A. The European Union and United Kingdom

The EU’s Digital Services Act (DSA), effective February 2024, requires very large online platforms to conduct systemic risk assessments for harms to “physical and mental well-being” and minors, implement mitigation measures including “adapting recommender systems,” and provide non-profiling feed options.¹⁸⁹ Article 34 specifically mandates assessment of “any actual or foreseeable negative effects” from “the design, functioning or use of their services”—language that directly parallels ABT’s focus on algorithmic mechanics rather than content. Platforms must evaluate how their “recommender systems and any other relevant algorithmic systems” create or amplify risks, particularly those affecting minors.¹⁹⁰ The Commission has opened formal proceedings against TikTok for addictive design targeting minors, alleging the “For You” feed and reward mechanisms violate DSA requirements—the first major enforcement action treating algorithmic addiction as a regulatory violation.¹⁹¹ Separately, the Commission has requested information from Meta regarding Instagram’s addictive-design features affecting minors.¹⁹² Meta has responded to European regulatory pressure by implementing design changes for EU users—including chronological feed options and reduced notification intensity—that it has not extended to American users, demonstrating both the feasibility of safer design and the role of legal pressure in motivating it.¹⁹³

¹⁸⁹DSA, *supra* note 8.

¹⁹⁰*Id.* art. 34(1)-(2) (requiring very large online platforms to “identify, analyse and assess... any actual or foreseeable negative effects for the exercise of fundamental rights” and specifically to evaluate “the design, functioning or use, including, where relevant, the interaction design, of their services”).

¹⁹¹See Press Release, European Commission, *Commission opens formal proceedings against TikTok under the Digital Services Act* (Feb. 19, 2024), https://ec.europa.eu/commission/presscorner/detail/en/ip_24_926 (last visited Dec. 23, 2025).

¹⁹²Press Release, European Comm’n, *Commission Sends Request for Information to Meta Under the Digital Services Act* (Nov. 14, 2024), https://ec.europa.eu/commission/presscorner/detail/en/ip_24_5765 (last visited Jan. 30, 2026).

¹⁹³See Natasha Lomas, *Meta tweaks EU ad-free subscription and adds ‘less personalized ads’ option*, TECHCRUNCH (Nov. 12, 2024), <https://techcrunch.com/2024/11/12/meta-tweaks-eu-ad-free-subscription-and-adds-less-personalized-ads-option/> (last visited Jan. 30, 2026) (documenting Meta’s implementation of chronological feed options and reduced personalization for EU users in response to DSA requirements); Meta, *Updates to Teen*

The UK’s Online Safety Act (2023) goes further, imposing a statutory “duty of care” requiring platforms to protect children from foreseeable algorithmic harms.¹⁹⁴ Ofcom’s implementing codes mandate specific design interventions: restricting autoplay for minors, limiting notification frequency, prohibiting personalized recommendation of harmful content to children, and requiring algorithmic impact assessments.¹⁹⁵ Unlike America’s content-focused approach, both the DSA and OSA treat algorithmic design as the primary vector of harm—precisely ABT’s insight. Penalties reach six percent (DSA) and ten percent (OSA) of global turnover, creating meaningful deterrence where American litigation has failed to deter.

B. Australia: The World’s First Social Media Ban for Minors

In November 2024, Australia enacted an outright ban on social media for children under sixteen—the most aggressive response any democracy has attempted.¹⁹⁶ The Online Safety Amendment Act 2024 permits no parental consent exception, rejecting the premise that parents can adequately consent to algorithmic manipulation their children cannot understand and

Accounts in the EU (Dec. 2024) (describing notification restrictions and content controls implemented for European users).

¹⁹⁴OSA, *supra* note 8.

¹⁹⁵See Ofcom, *Protecting children from online harm: Our proposed codes of practice* (Nov. 7, 2023), <https://www.ofcom.org.uk/news-centre/2023/protecting-children-from-online-harm-our-proposed-codes-of-practice> (last visited Dec. 23, 2025).

¹⁹⁶Online Safety Amendment (Social Media Minimum Age) Act 2024 (Cth) (Austl.). The legislation received Royal Assent on November 29, 2024, with enforcement provisions taking effect December 10, 2025. For comprehensive analysis, see eSafety Commissioner, *Social Media—Loss of Access Age Restrictions* (2024), <https://www.esafety.gov.au/industry/social-media-services> (last visited Dec. 23, 2025).

platforms deliberately conceal.¹⁹⁷ Penalties reach \$49.5 million AUD per violation, with platforms bearing the burden of age verification rather than users.¹⁹⁸

The legislative record makes the rationale explicit: the Senate inquiry found that platforms’ “addictive design features” and “algorithmic curation” create “significant and serious risks” to youth mental health that platforms have failed to self-correct.¹⁹⁹ Australian mental health data validates ABT’s premises: between 2007-2022, youth mental disorders increased forty percent (males) and sixty percent (females), tracking social media adoption curves with timing that cannot be explained by diagnostic changes alone.²⁰⁰ Scientific observers characterize the ban as a “natural experiment” that will produce evidence platform-funded studies have systematically obscured—evidence American courts will eventually receive.²⁰¹

¹⁹⁷See Haidt, *supra* note 55, at 187–215 (arguing that smartphone-based childhood has fundamentally altered adolescent development in ways that duty-based regulation cannot adequately address). Haidt’s work was explicitly cited in Australian parliamentary debates as providing the evidentiary foundation for categorical prohibition; Online Safety Amendment (Social Media Minimum Age) Act 2024 § 63B (Cth) (Austl.). Covered platforms include those with over one million Australian users that enable interaction between users, specifically: Facebook, Instagram, TikTok, Snapchat, Reddit, X (formerly Twitter), YouTube, Twitch, Threads, and Kick. Messaging services (WhatsApp, Messenger) and gaming platforms (Roblox) are exempt.

¹⁹⁸*Id.* § 63C. Parliament explicitly rejected parental consent exceptions after evidence that parents cannot meaningfully assess algorithmic risks without access to platform design specifications. See Senate Inquiry into Online Safety Amendment Bill, Submissions Summary 12–14 (2024) (summarizing expert testimony on information asymmetry between platforms and parents); Online Safety Amendment Act 2024 § 63H (Cth) (Austl.) § 63H (establishing maximum civil penalty of 50,000 penalty units). Under current valuation, this equals approximately \$49.5 million AUD (~\$32 million USD) per systemic violation. The eSafety Commissioner has enforcement discretion to pursue penalties proportionate to platform revenue.

¹⁹⁹Parliament of Australia, Senate Standing Committees on Environment and Communications, *Social media’s influence on Australian society* (2024), https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Environment_and_Communications/Socialmedia (last visited Dec. 23, 2025) (finding “the design features of platforms that facilitate harmful content and encourage compulsive use” are “a key driver” of youth mental health harms).

²⁰⁰See Australian Institute of Health & Welfare, *Mental Health of Young People* (2023), <https://www.aihw.gov.au/reports/children-youth/mental-health-of-young-people> (last visited Dec. 23, 2025) (documenting increases in mental disorder prevalence from 2007 baseline). The forty-percent male and sixty-percent female increases represent age-adjusted rates controlling for improved diagnostic practices; *Id.*; see also Australian Bureau of Statistics, *Causes of Death, Australia* (2023) (suicide data); Orygen, *Youth Mental Health Statistics* (2024), <https://www.orygen.org.au/About/News-And-Events/2024/Youth-Mental-Health-Report> (last visited Dec. 23, 2025) (self-harm hospitalization data). These statistics parallel American trends documented in *supra* notes 2–5.

²⁰¹Dyani Lewis, *Australia’s world-first social media ban is a ‘natural experiment’ for scientists*, NATURE (Dec. 2025), <https://www.nature.com/articles/d41586-025-04069-2> (last visited Dec. 23, 2025) (describing research

Australia chose prohibition; ABT offers the tort alternative. Both rest on identical premises: platforms exploit vulnerabilities in minors, algorithmic manipulation causes cognizable harm, and legal intervention is constitutionally permissible. The bill passed with seventy-seven percent public approval, reflecting a democratic consensus American law has yet to express.²⁰²

C. Converging Global Principles

Across three continents and four legal systems, five principles have achieved consensus:²⁰³

1. Platforms owe duties to users. Whether framed as statutory duty of care (UK), systemic risk obligations (EU), or algorithmic fiduciary duties (ABT), platforms can no longer disclaim responsibility for the consequences of their design choices.

2. Algorithms are primary risk vectors. The EU, UK, and Australia each identified recommender systems—not user-generated content—as the central mechanism of harm. ABT’s conduct-content distinction reflects this international consensus.

3. Transparency enables accountability. All three systems mandate disclosure of algorithmic functioning, though with varying specificity. ABT’s candor duty imports this principle into tort law.

opportunities created by Australia’s ban); *Id.* (“The law creates conditions that would be impossible to replicate in controlled research settings—a population-level intervention affecting millions of adolescents.”).

²⁰²YouGov Australia, *Social Media Age Restrictions Poll* (Oct. 2024) (finding 77% of Australian adults support under-16 social media ban, with support crossing partisan lines); eSafety Commissioner, *Online Safety Amendment Bill: Public Consultation Summary* (Nov. 2024) (reporting 15,247 public submissions—the largest response to any online safety consultation in Australian history).

²⁰³See Online Safety Act 2021 (Cth) (Austl.); see also Amy Kamarainen, *The Australian eSafety Commissioner and Algorithmic Accountability*, 27 INFO. TECH. L. J. 41 (2022); see Bill C-63, An Act to enact the Online Harms Act, 44th Parl., 1st Sess. (Can. 2024).

4. Vulnerable populations warrant special protection. Every regulatory system treats minors as categorically entitled to heightened protection, rejecting platforms’ arguments that parental supervision suffices.

5. Meaningful enforcement requires substantial penalties. Fines measured in percentage of global revenue (EU, UK) or eight-figure absolute amounts (Australia) replace the token penalties that have failed to deter.

Dimension	EU DSA	UK OSA	Australia 2024	ABT (Proposed)
Regulatory Model	Ex ante risk regulation	Statutory duty of care	Categorical prohibition	Ex post tort liability
Primary Mechanism	Systemic risk assessments	Ofcom codes	Age verification + ban	Private litigation
Maximum Penalty	6% global turnover	10% global turnover	\$49.5M AUD	Compensatory + punitive

These frameworks are complementary rather than competing: the EU regulates ex ante, the UK imposes ongoing duties, Australia prohibits access for minors, and ABT provides ex post remediation and compensation. Only tort law can make victims whole. Only tort law creates individual accountability for platform executives who made the design choices. And only American courts can apply these principles to American plaintiffs harmed by American design decisions.

CONCLUSION: COGNITIVE INTEGRITY AS LEGAL RIGHT

We have enclosed the land. We have enclosed the commons. We have enclosed culture and information through what scholars have called a “second enclosure movement” in intellectual property. We have even tried to enclose the genetic code itself—until courts drew limits. The last frontier of enclosure is the human mind. The fundamental principle is simple, and

must be stated plainly: the human mind is not a commodity to be mined.²⁰⁴ This is not a novel proposition. It is the logical extension of rights the law has protected for centuries—the right to bodily integrity, the right to reputation, the right to privacy.²⁰⁵ What distinguishes this moment is not the principle but the threat: never before has technology enabled such systematic, covert, and profitable exploitation of human cognitive vulnerabilities on this scale.²⁰⁶

The Algorithmic Behavioral Tort does three things:

First, it names the protected interest—cognitive integrity, the right to mental self-determination free from systematic manipulation.²⁰⁷ The Cartesian self, the “I” who thinks, must remain autonomous. When algorithms become the Cartesian demon, law must provide the shield philosophy cannot.²⁰⁸

²⁰⁴See Inclosure Act 1773, 13 Geo. 3 c. 81 (Gr. Brit.), <https://www.legislation.gov.uk/apgb/Geo3/13/81/contents> (last visited Dec. 23, 2025) (authorizing enclosure of land and extinguishing common rights); Inclosure Act 1845, 8 & 9 Vict. c. 118 (Gr. Brit.), <https://www.legislation.gov.uk/ukpga/Vict/8-9/118/contents> (last visited Dec. 23, 2025) (creating Inclosure Commissioners and streamlining enclosures); James Boyle, *The Second Enclosure Movement and the Construction of the Public Domain*, 66 LAW & CONTEMP. PROBS. 33, 33–34 (2003) (describing a “second enclosure movement” in intellectual property); *Diamond v. Chakrabarty*, 447 U.S. 303, 309–10 (1980) (broadly construing patent-eligible subject matter for biotechnology); *Ass’n for Molecular Pathology v. Myriad Genetics, Inc.*, 569 U.S. 576, 590–95 (2013) (limiting patents on naturally occurring DNA while allowing claims to cDNA); Richards, *Intellectual Privacy*, *supra* note 6 (arguing that intellectual/cognitive privacy is a foundational civil liberty and rejecting commodification of thought); Zuboff, *Surveillance Capitalism*, *supra* note 10, at 8–11 (describing extraction and monetization of human experience as raw material).

²⁰⁵See RESTATEMENT (SECOND) OF TORTS § 13 (1965) (battery protects bodily integrity); *id.* §§ 558–59 (1977) (defamation protects reputation); *id.* §§ 652A–E (1977) (privacy torts, including intrusion upon seclusion and public disclosure of private facts); cited *supra* notes 58, 62–66, 104.

²⁰⁶See FED. TRADE COMM’N, BRINGING DARK PATTERNS TO LIGHT: AN FTC STAFF REPORT (Sept. 2022), <https://www.ftc.gov/reports/bringing-dark-patterns-light-ftc-staff-report> (last visited Oct. 11, 2025) (documenting manipulative design), cited *supra* note 5; Georgia Wells, Jeff Horwitz & Deepa Seetharaman, *Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show*, WALL ST. J. (Sept. 14, 2021), <https://www.congress.gov/117/meeting/house/114054/documents/HHRG-117-IF02-20210922-SD003.pdf> (last visited Oct. 11, 2025), cited *supra* note 4; Frances Haugen, Testimony Before the U.S. SENATE COMM. ON COMMERCE, SCIENCE, AND TRANSPORTATION (Oct. 5, 2021), cited *supra* note 21; PEW RESEARCH CENTER, TEENS, SOCIAL MEDIA AND TECHNOLOGY 2023 (Dec. 11, 2023), <https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/> (last visited Oct. 11, 2025).

²⁰⁷See *supra* Part I.B (defining the protected interest as cognitive integrity and grounding it in intellectual privacy and dignitary torts); see also Richards, *Intellectual Privacy*, *supra* note 6, at 421–25.

²⁰⁸See *supra* Part I.B.4 (developing the Cartesian demon framework and arguing that engagement-maximizing algorithms function as the modern *malin génie*).

Second, it creates duties scaled to power and vulnerability. Platforms wielding unprecedented informational advantages and algorithmic control over billions—including millions of children—owe duties commensurate with that power.²⁰⁹ Whether courts frame these as fiduciary obligations or duties arising from special relationships, the substance is identical: loyalty, care, and candor toward users whose cognitive processes are subject to the platform’s discretionary influence.²¹⁰

Third, it supplies workable tools for immediate implementation. The Manipulation-Vulnerability Matrix translates theory into practice. The damages methodology provides valuation guidance. The proof map organizes complex evidence.²¹¹ Courts need not wait for legislative action or doctrinal consensus. ABT is ready to apply today.²¹²

The Rage Machine can be stopped. The addiction mechanics can be disabled.²¹³ The choice is not between innovation and stagnation—it is between innovation that serves human flourishing and innovation that treats adolescent psychology as exploitable raw material for engagement optimization. The former is commerce; the latter is systematic harm masquerading as inevitability.

²⁰⁹See *supra* Part I.C (algorithmic fiduciary relationship; special-relationship pathway); see also PEW RESEARCH CENTER, *Teens, Social Media and Technology 2023* (Dec. 11, 2023), <https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/> (last visited Dec. 23, 2025) (platform reach to minors).

²¹⁰See *supra* Part I.C (duties of loyalty, care, and candor regardless of label); Balkin, *Information Fiduciaries*, *supra* note 101, at 1190–1201.

²¹¹See *supra* Part I.D–F (Matrix, damages, proof map); see also *supra* Part III (applications).

²¹²See *supra* Part IV; Lemmon, 995 F.3d at 1092–93, *supra* note 15; see also 15 U.S.C. § 45(a), (n).

²¹³See Ofcom, *supra* note 195; DSA art. 38(1)–(2), *supra* note 8.

The evidence of harm is overwhelming.²¹⁴ The economic incentives will not self-correct—profit and user well-being are misaligned by design.²¹⁵ The international community has already acted, from the European Union’s Digital Services Act to the United Kingdom’s Online Safety Act.²¹⁶ American law, which once led the world in protecting fundamental rights against evolving threats, must not abdicate that role now.²¹⁷

At stake is nothing less than cognitive autonomy in the digital age. Without cognitive integrity, the First Amendment protects only the right to speak what an algorithm has nudged us to say. Without it, the “marketplace of ideas” becomes a company town where the currency is outrage, the rent is attention, and the landlord is the Rage Machine.²¹⁸

Platforms can infer a teen’s vulnerability within minutes when she feels “worthless.”²¹⁹ Accountability must not take years.²²⁰ ABT offers courts, legislators, and regulators the necessary tools to protect the most fundamental of human interests—the integrity of thought itself—while preserving the potential for genuine innovation and free expression.²²¹

²¹⁴See Surgeon General’s Advisory, *supra* note 2, at 4–9; Odgers & Jensen, *supra* note 36; Wells, Horwitz & Seetharaman, *supra* note 4; Haugen, *supra* note 21.

²¹⁵See Zuboff, Surveillance Capitalism, *supra* note 10, at 93–128; Schüll, Addiction by Design, *supra* note 19, at 5–8, 321–28; Alter, Irresistible, *supra* note 19, at 49–53.

²¹⁶See Part IV.A–B; DSA arts. 34–35, 38, *supra* note 8; OSA ss. 11–12, *supra* note 8.

²¹⁷See 15 U.S.C. § 45(a); Lemmon, 995 F.3d at 1092–93, cited *supra* note 15; *cf.* 47 U.S.C. § 230.

²¹⁸See *Abrams v. United States*, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting) (articulating the “marketplace of ideas” rationale for free speech); *Marsh v. Alabama*, 326 U.S. 501, 506–09 (1946) (treating a privately owned company town as the functional equivalent of a municipality for First Amendment purposes); *see also supra* Part I.B; Richards, Intellectual Privacy, *supra* note 6.

²¹⁹See Levin, *supra* note 37; *see also* Haugen, *supra* note 21.

²²⁰See DSA art. 74(1), cited *supra* note 189; Press Release, European Commission, Commission Opens Formal Proceedings Against TikTok Under the DSA (Feb. 19, 2024), https://ec.europa.eu/commission/presscorner/detail/en/ip_24_926 (last visited Oct. 11, 2025), cited *supra* note 191.

²²¹See Lemmon, *supra* note 15; Balkin, Information Fiduciaries, *supra* note 101; *see* Appendix (implementation remedies preserving innovation while curbing manipulation).

The law has always evolved to protect what society values most. In the agrarian age, it was land. In the industrial age, it was labor. In the information age, it must be the mind.

The Rage Machine is not inevitable. It is a product of design, and design can be regulated. It is a product of law, and law can be changed. The question is not whether we have the tools to dismantle the machine. The question is whether we value our own minds enough to pick them up.

APPENDIX: IMPLEMENTATION PATHWAYS

ABT can move now on three fronts—litigation, legislation, regulation. Each pathway reinforces the others and creates real incentives to design for well-being.

A. Litigation: Playbook for Plaintiffs' Counsel

Pleading Strategy: Plead conduct, not content. Frame claims around platform's design choices—algorithmic amplification, addiction mechanics, vulnerability exploitation—not user-generated posts. This framing navigates Section 230 by targeting the platform's own tortious conduct.

Elements to Establish:

1. Duty: Defendant operates a “covered platform” using personalized algorithmic curation, establishing the fiduciary relationship.
2. Breach: Matrix analysis demonstrates manipulative design; internal documents show knowledge of harms and rejection of safer alternatives.
3. Causation: Substantial-factor test using population-level evidence (platform research, epidemiology) and individual exposure data (engagement logs, recommendation pathways).

4. Harm: Clinical diagnosis preferred; severe psychological distress with functional impairment suffices.

Discovery Priorities: Seek design documents, internal research on mental health impacts, A/B testing data (especially tests of safer alternatives that were rejected), engagement metrics, algorithmic decision-making code, and communications regarding safety-vs-engagement trade-offs.

Expert Team: Computer scientist (decode algorithmic mechanics), psychologist/psychiatrist (establish cognitive harm and causation), epidemiologist (population-level risk evidence).

B. Legislative Framework

Model legislation should:

Codify algorithmic fiduciary duties for platforms using personalized recommendation systems

Enumerate prohibited manipulative practices: variable reward schedules targeting minors, vulnerability detection without disclosure, autoplay without affirmative opt-in for minors

Establish burden-shifting causation frameworks: once plaintiff demonstrates Red Zone Matrix placement, burden shifts to platform to prove its design was not a substantial factor

Provide private rights of action with attorney's fees to ensure access to justice

Create safe harbors for platforms offering non-profiling feeds for minors and conducting independent algorithmic audits

C. Regulatory Enforcement

The FTC and state attorneys general already possess authority under unfair and deceptive acts and practices (UDAP) statutes to pursue ABT-style claims:

“Unfair” conduct: Manipulative design causing substantial unavoidable injury satisfies unfairness prong

“Deceptive” omissions: Failure to disclose engagement-vs-wellbeing conflicts, concealment of internal research documenting harms

Enforcement tools: Consent decrees mandating algorithmic audits, disgorgement of profits from manipulative design, civil penalties

State attorneys general have proven effective, as demonstrated by ongoing litigation in California, New York, and the multi-state MDL. Coordinated enforcement across federal and state authorities maximizes pressure for systemic reform.