AI REGULATION: TROJAN HORSES, SHIFTING WINDOWS AND ESTRANGING THE FAMILIAR

Krishna Deo Singh Chauhan[*] *and Anupriya Singh*[**]

ABSTRACT

On 25 March 2026, a California jury found Meta and YouTube liable for negligently designing social media platforms that addicted and harmed a young woman. The verdict, which applied product liability doctrine to platform architecture rather than content, was not an isolated event. It was one of three developments that this article examines – the Trojan horse problem, where AI-powered technologies present a benign exterior while concealing architectures designed to exploit users, the rapid shift in the Overton window of

[*] Krishna Deo Singh Chauhan is an Associate Professor of Law and Assistant Dean, Jindal Global Law School, O.P. Jindal Global University and Associate Director, Cyril Shroff Centre for AI, Law and Regulation (CSCAILR). The author may be reached at KDSingh@jgu.edu.in.
[**]Anupriya Singh is a Lecturer, and Assistant Dean, Jindal Global Law School, O.P. Jindal Global University. The author may be reached at anupriya@jgu.edu.in.

regulatory acceptability, as actions once considered unthinkable, such outright bans on social media for minors, multi-billion Euro fines for addictive design, jury verdicts treating algorithms as defective products have materialised in quick succession across multiple jurisdictions, and the implications of these developments for how regulators should approach AI.

Drawing on Viktor Shklovsky's literary concept of ostranenie, or defamiliarisation, the article argues that the central failure in AI regulation is not a failure of law but a failure of vision. Legal instruments are available, but what is missing is the willingness to see AI-enabled systems for what they functionally are, rather than what they formally resemble. Regulators, courts, and legislators see familiar objects and apply familiar categories. The harms accumulate in the gap between appearance and reality.

The article proposes that regulators must institutionalise a standing disposition to look again at what appears familiar – to ask whether the category that once applied still holds. Three specific demands follow from this:

categorical scepticism, functional classification, and pre-emptive scrutiny. The article concludes by examining how this framework applies to India's developing AI regulatory landscape.

Keywords: *AI Regulations, Reciprocal Tariffs, International Emergency Economic Powers Act (IEEPA), Trade Policy.*

I. INTRODUCTION

On 25 March 2026, a jury in the Los Angeles County Superior Court found Meta and Google liable for negligently designing platforms that addicted a young woman and damaged her mental health.[1] The verdict awarded six million dollars in compensatory and punitive damages.[2] In the grand scheme of things, for two companies whose combined annual revenues exceed three hundred and fifty billion dollars, this is not a highly significant amount. But its significance lies in the legal re-orientation it entails. What the jury affirmed was not simply that social media can harm children. It affirmed that the very architecture of a

[1] *KGM v Meta Platforms, Inc.*, Los Angeles County Superior Court, (verdict 25 March 2026); Cecilia Kang, Ryan Mac and Eli Tan, 'Meta and YouTube Found Negligent in Landmark Social Media Addiction Case' (*The New York Times*, 25 March 2026) <https://www.nytimes.com/2026/03/25/technology/social-media-trial-verdict.html> accessed 4 April 2026.

[2] ibid.

digital product, such as infinite scroll, autoplay loops, notification scheme calibrated to exploit users' subconscious biases, can constitute a defect.[3] In other words, the same legal category that applies to faulty brakes and exploding gas tanks was applied, for the first time in a jury verdict, to the engineering of attention itself.

This was not rather unthinkable in the legal paradigm that pre-existed. For nearly three decades, Section 230 of the Communications Decency Act had functioned as an almost impenetrable shield, deflecting lawsuits against platforms with the argument that they merely hosted content they did not create.[4] The plaintiff's legal team in *KGM v Meta Platforms, Inc* (hereinafter referred to as **'KGM'**) innovated on the legal case they built.[5] They shifted the target from content to conduct, from what users post to how the platform is built.[6]

This article takes the *KGM* verdict as a point of departure, but its argument is broader. The verdict, it is submitted, is symptomatic of three developments in the legal treatment of AI-powered technologies that deserve systematic attention. The first concerns what might be

[3] 'Jury Orders $3 Million in Punitive Damages Against Meta and YouTube in Landmark Social Media Addiction Trial' (*The Lanier Law Firm*) <https://www.lanierlawfirm.com/news/jury-orders-3-million-in-punitive-damages-against-meta-and-youtube-in-landmark-social-media-addiction-trial/> accessed 4 April 2026.

[4] Communications Decency Act 1996, 47 USC § 230(c)(1); *Lemmon v Snap, Inc.*, 995 F.3d 1085 (9th Cir. 2021).

[5] *KGM* (n 1).

[6] Kristin Stoller, 'A Court Just Ruled That Tech Addiction Is Real—and Dangerous. It Could Be Meta and YouTube's Big Tobacco Moment' (*Fortune*) <https://fortune.com/2026/03/25/meta-google-instagram-youtube-tech-addiction-lawsuit-kgm/> accessed 11 April 2026.

called the Trojan horse problem, is the tendency of AI-driven products to present an (often benign) exterior, while it contains deliberate architectures and emergent properties,[7] and that can lead to manipulation, exploitation, or harm through their design. Social media platforms are the paradigmatic instance, but they are far from the only one. The second concerns the shifting Overton window of legal and political acceptability. Actions against technology companies that would have been inconceivable five years ago, such as outright bans on social media for minors, multi-billion dollar fines for platform design, jury verdicts treating algorithms as defective products, have materialised in rapid succession. The window of what is legally thinkable is in motion, and it is widening. The third, and the one that emerges from the second, concerns the implications of this, and the lessons that must be drawn from them, particularly by regulators. The central thesis of this paper is that the first too phenomena necessitate an evaluation not just of specific regulations, but of the regulatory approach towards AI and AI powered systems, towards a pre-emptive stance. Metaphorically drawing on the concept of *ostranenie*, or defamiliarization from the works of the literary theorist Viktor Shklovsky, the paper suggests that regulators must defamiliarize themselves of the conventional approaches and categories of regulation

[7] 'What Is Emerging in Artificial Intelligence Systems?' (*Max Planck Law*) <https://law.mpg.de/perspectives/what-is-emerging-in-artificial-intelligence-systems/> accessed 11 April 2026.

when dealing with AI. Some specific implications of this process are then discussed.

These three claims are developed across the three Parts that follow. *Part II* examines the Trojan horse dynamic through the lens of the *KGM* litigation and its legal innovations. *Part III* traces the shifting Overton window across multiple jurisdictions. *Part IV* develops the *ostranenie* framework as a regulatory method, operationalising it into three specific demands – categorical scepticism, functional classification, and pre-emptive scrutiny, before examining its implications for India's developing AI regulatory landscape.

II. TROJAN HORSES: THE ARCHITECTURE OF CONCEALED HARM

A. THE GIFT AND WHAT IT CONCEALS

The Trojan horse is perhaps one of the oldest metaphors for strategic deception in the Western tradition. The Greeks offered Troy a magnificent wooden horse as a gift. The Trojans, dazzled by its beauty and convinced of its innocence, brought it inside the city walls. What followed is well known.[8] The metaphor endures because it captures a specific kind of risk – that not all risks are exhausted by those that appear obvious, but may be part of something that appears to be a gift.

[8] 'Trojan Horse | Story & Facts | Britannica' (27 February 2026) <https://www.britannica.com/topic/Trojan-horse> accessed 4 April 2026.

Social media platforms arrived in precisely this fashion. Facebook offered connection. Instagram offered self-expression. YouTube offered access to infinite knowledge and entertainment. Every successful social media platform, brought a major benefit, which were, and remain, genuinely appealing propositions. They were also free, which made them irresistible.

What was concealed within these appealing exteriors was an architecture of engagement optimisation that operated on principles borrowed, sometimes quite consciously, from the behavioural science of addiction.[9] Variable-ratio reinforcement schedules, the same mechanism that makes slot machines compelling, were embedded in notification systems and content feeds.[10] Infinite scroll eliminated natural stopping points, mimicking the *ludic loop* documented by Natasha Dow Schüll in her study of machine gambling and autoplay features ensured that passivity, not activity, was the default state of engagement.[11] Lastly, but importantly to the present discussion, recommendation algorithms, powered by deep learning models trained on billions of interactions, refined the delivery of content with a precision that earlier media technologies could not have imagined.

[9] Shoshana Zuboff, *The Age of Surveillance Capitalism* (Profile Books 2019).

[10] CB Ferster and BF Skinner, 'Variable Ratio, *Schedules of reinforcement.*' (Appleton-Century-Crofts 1957) <https://doi.org/10.1037/10627-007> accessed 4 April 2026.

[11] Natasha Dow Schüll, *Addiction by Design: Machine Gambling in Las Vegas* (Princeton University Press 2014).

The *KGM* trial surfaced internal Meta documents in which executives discussed strategies to attract and retain users as young as eleven. One internal document stated that if the company wanted to succeed with teenagers, it had to recruit them as pre-teens. Another internal communication, disclosed during trial, likened Instagram to a drug, with the acknowledgement that the company's employees functioned as its distributors.[12] Sean Parker, Facebook's founding president, had publicly acknowledged as early as 2017 that the platform was designed to exploit psychological vulnerabilities.[13] Frances Haugen's disclosures in 2021 revealed internal research showing that Meta was aware of the harm its platforms inflicted on adolescent mental health, including evidence that Instagram worsened body image issues in a significant proportion of teenage girls.[14]

The plaintiff in the *KGM* case testified that she began using YouTube at the age of six and Instagram at nine. By the time she finished elementary school, she had posted 284 videos on YouTube. She described compulsive use, difficulty stopping, anxiety when separated

[12] Tyler Katzenberger, '"We'Re Basically Pushers": Court Filing Alleges Staff at Social Media Giants Compared Their Platforms to Drugs' (*POLITICO*, 23 November 2025) <https://www.politico.com/news/2025/11/22/were-basically-pushers-court-filings-allege-staff-at-social-media-giants-compared-their-platforms-to-drugs-00666181> accessed 4 April 2026.

[13] Olivia Solon, 'Ex-Facebook President Sean Parker: Site Made to Exploit Human "Vulnerability"' (*The Guardian*, 9 November 2017) <https://www.theguardian.com/technology/2017/nov/09/facebook-sean-parker-vulnerability-brain-psychology> accessed 4 April 2026..

[14] Megan Switzgable, 'How the Media Frames Mental Health and Social Media: A Case Study of the Facebook Whistleblower' (Senior Honors Thesis, University of New Hampshire 2022) https://scholars.unh.edu/honors/622/ accessed 4 April 2026.

from her devices, and an escalating need for engagement. She was later diagnosed with depression, anxiety, and body dysmorphic disorder.[15]

B. THE LEGAL INNOVATION: CONTENT VERSUS CONDUCT

The legal significance of *KGM* lies in the manoeuvre by which the plaintiff's legal team circumnavigated Section 230.[16] Traditional lawsuits against social media companies typically centred on harmful third-party content: a user was exposed to such material, suffered injury, and sought to hold the platform liable for hosting, publishing, or recommending it.[17] Section 230 was designed precisely to foreclose this line of attack. Platforms were treated as conduits, not publishers. The content belonged to the users who created it; the platform merely provided the infrastructure. Hundreds of lawsuits had been dismissed on this basis over the preceding decades.

The *KGM* litigation reframed the claim entirely. The harm, the plaintiff argued, arose not from any particular piece of content but from the platform's *design features* themselves: infinite scroll, autoplay, push

[15] Sarah Shamim, 'Jury Finds Meta, YouTube Liable for Social Media Addiction: What We Know' (*Al Jazeera*) <https://www.aljazeera.com/news/2026/3/26/jury-finds-meta-youtube-liable-for-social-media-addiction-what-we-know> accessed 4 April 2026; Dara Kerr, 'Meta and YouTube Designed Addictive Products That Harmed Young People, Jury Finds' (*The Guardian*, 25 March 2026) <https://www.theguardian.com/media/2026/mar/25/jury-verdict-us-first-social-media-addiction-trial-meta-youtube> accessed 4 April 2026.

[16] Communications Decency Act (n 4).

[17] Diana Novak Jones and Diana Novak Jones, 'What Comes next after the Social Media Trial Verdicts?' (*Reuters*, 25 March 2026) <https://www.reuters.com/sustainability/boards-policy-regulation/what-comes-next-after-social-media-trial-verdicts-2026-03-25/> accessed 4 April 2026.

notifications calibrated to maximise re-engagement, algorithmic recommendation systems optimised for time-on-platform, and the absence of meaningful parental controls or usage limits. These were not acts of publishing or speaking. They were *product design decisions*, and they could be evaluated under the same product liability framework that governs any manufactured goods. Jurors were specifically instructed not to consider the content of the posts and videos Kaley had seen on the platforms, so that the case turned on whether the platforms' design and warnings, rather than third-party content, had caused her harm.[18]

Judge Carolyn B. Kuhl's November 5, 2025 ruling denying Meta's motion for summary judgment drew an important distinction. Features tied to the publication or display of third-party content may fall within Section 230's protection. But claims directed at the platform's own design features, such as infinite scroll and other features alleged to drive compulsive use, were not treated as automatically barred on that

[18] Kaitlyn Huamani, Barbara Ortutay and The Associated Press, 'Meta and YouTube Found Liable in Landmark Child Social Media Harm Case, Ordered to Pay $3 Million—with Punitive Damages Still to Come' (*Fortune*) <https://fortune.com/2026/03/25/meta-youtube-liable-child-harm-social-media-punitive-damages-3-million-case/> accessed 4 April 2026.

basis.[19] The ruling thus offered a possible roadmap for distinguishing publisher liability from product-design liability in future cases.[20]

The product-based framing translated, in this trial, into negligence and failure-to-warn questions focused on platform design and operation. The jury was asked whether Meta had been negligent in the design or operation of Instagram, whether that negligence was a substantial factor in causing harm, whether Meta knew or should have known of the danger to minors using the platform in a reasonably foreseeable manner, and whether Meta failed adequately to warn of that danger. The jury answered those questions in the plaintiff's favour.[21] Meta was assigned seventy per cent of the responsibility, and Google, thirty per cent. The punitive-damages finding further indicated that the jury concluded, by clear and convincing evidence, that the companies' conduct involved malice, oppression, or fraud, going beyond mere negligence and suggesting at least a finding of aggravated misconduct,

[19] *Christina Arlington Smith, et al v TikTok Inc, et al* (Superior Court of California, County of Los Angeles, Central District, Spring Street Courthouse, Department 12, No 22STCV21355, 5 November 2025, Hon Carolyn B Kuhl), https://www.courthousenews.com/wp-content/uploads/2025/11/social-media-lawsuits-kgm-motion-denied.pdf, accessed 11 April 2026

[20] Haim Ravia and Dotan Hammer, 'California State Court Jury Find Meta and Google Liable for Teen Addiction' (*Pearl Cohen*, 26 March 2026) <https://www.pearlcohen.com/california-state-court-jury-find-meta-and-google-liable-for-teen-addiction/> accessed 4 April 2026.

[21] 'Landmark Verdicts Against Meta and YouTube Signal New Era of Social Media Platform Liability' (*Crowell & Moring - Landmark Verdicts Against Meta and YouTube Signal New Era of Social Media Platform Liability*) <https://www.crowell.com/en/insights/client-alerts/landmark-verdicts-against-meta-and-youtube-signal-new-era-of-social-media-platform-liability> accessed 4 April 2026.

potentially including conscious disregard of known risks.[22] The platforms concerned were thus, liable.

This dynamic is not unique to social media. It is a recurring feature of AI-powered consumer technologies. These systems are introduced as conveniences, tools, or companions, often offered freely or cheaply, and presented as empowering. Yet embedded within their architecture are mechanisms that extract value from users in ways that are neither transparent nor fully understood at the moment of adoption. Smart speakers, navigation apps, AI chatbots, and generative AI tools all illustrate this pattern. Their surface-level value is real, but the business models and behavioural architectures underlying them often carry concealed costs, from surveillance and data extraction to psychological dependency and deskilling.

These technologies are not simply harmful, they are genuinely useful and harmful at the same time. Often, the very features that create value also create risk. But the risk is hidden. And finding it in a timely fashion requires a careful, even pre-emptive look.

III. SHIFTING WINDOWS: THE EXPANDING FRONTIER OF REGULATORY ACCEPTABILITY

A. THE OVERTON WINDOW AND LEGAL CHANGE

[22] ibid.

The concept of the Overton window, developed by Joseph Overton at the Mackinac Center for Public Policy in the 1990s, describes the range of policies that are considered politically acceptable at any given moment.[23] Ideas outside the window are dismissed as radical, impractical, or unthinkable. Ideas within it are treated as reasonable subjects for debate, even if controversial. The window is not fixed. It shifts over time in response to changing circumstances, public opinion, technological disruption, and the cumulative effect of events that render previously unthinkable propositions suddenly plausible.

The regulation of AI-powered technologies provides a striking illustration of how rapidly the Overton window can move. More importantly, it shows that what is shifting is not merely regulatory intensity, but the underlying legal and political imagination through which digital systems are understood. Consider the following thought experiment. In 2019, if a policy analyst had proposed that a liberal democracy would require major social media platforms to prevent those under sixteen from holding accounts, backed by penalties running into tens of millions of dollars and without a parental-consent exception, the proposal would likely have been dismissed as paternalistic overreach. If the same analyst had suggested that a California jury would hold Meta and YouTube liable on negligence-based theories focused on the design and operation of their platforms, the claim would have seemed like a category error. Social media

[23] 'The Overton Window' (*Mackinac Center*) <https://www.mackinac.org/OvertonWindow> accessed 4 April 2026.

platforms are not toasters, and their harms do not fit easily within older intuitions about defective products. And if the analyst had predicted that the European Commission would impose a fine exceeding one hundred million euros in part because of the deceptive design of a platform's verification system, that too would have seemed faintly absurd.[24]

Yet all three of these things happened within the span of a few months, between late 2025 and early 2026. The window shifted, and it shifted fast. And the shift is not simply towards harsher regulation. It is towards a new understanding of platforms and AI systems as designed environments whose architecture can itself be the object of legal judgment.

In the forthcoming chapter, we look at these phenomena in greater details and delineate the implications.

B. AUSTRALIA: BANNING SOCIAL MEDIA FOR MINORS

The Online Safety Amendment (Social Media Minimum Age) Act 2024 received Royal Assent on 10 December 2024. Its age-restriction regime took effect on 10 December 2025,[25] applying age-restrictions

[24] CNN, 'Meta and Youtube Found Liable in Social Media Addiction Trial' < https://edition.cnn.com/2026/03/25/tech/social-media-addiction-trial-jury-decision> accessed 31 March 2026.

[25] 'eSafety Statement on the Online Safety Amendment' (Social Media Minimum Age) Act 2024 | eSafety Commissioner) <https://www.esafety.gov.au/newsroom/media-releases/esafety-statement-on-the-online-safety-amendment-social-media-minimum-age-act-2024> accessed 4 April 2026.

to social media platforms, including Facebook, Instagram, TikTok, YouTube, Snapchat, X, Reddit, Twitch, Threads, and Kick.[26] Critically, the prohibition cannot be overridden by parental consent. The onus is placed entirely on social media companies to take reasonable steps to prevent under-sixteens from holding accounts, with penalties of up to AUD 49.5 million for non-compliance.[27]

The genesis of the legislation is itself instructive. A key political catalyst appears to have been South Australian Premier Peter Malinauskas's response to Jonathan Haidt's The Anxious Generation, after his wife urged him to read it. The proposal did not emerge solely from a conventional law-reform process, but was shaped in significant part by a personal and political reaction to a broader moral panic about youth social media use.[28] Malinauskas subsequently commissioned a report, which provided the intellectual foundation for the ban. This origin story captures something important about how the Overton window moves: not always through institutional deliberation but

[26] 'Which Social Media Platforms Are Age-Restricted? ' (eSafety Commissioner) <https://www.esafety.gov.au/about-us/industry-regulation/social-media-age-restrictions/which-platforms-are-age-restricted> accessed 4 April 2026.

[27] Canberra Commonwealth Parliament, Parliament House, 'Online Safety Amendment (Social Media Minimum Age) Bill 2024' <https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/bd/bd2425/25bd039> accessed 4 April 2026.

[28] Josh Taylor, 'A US Psychologist Prescribed a Social Media Ban for Kids. How Did Australia Become the Test Subject?' (*The Guardian*, 7 December 2025) <https://www.theguardian.com/australia-news/2025/dec/07/how-australia-became-the-testing-ground-for-a-social-media-ban-for-young-people> accessed 4 April 2026.

through the intersection of personal experience, public anxiety, and political opportunity and brings forth harms that are all but fully visible.

What is most notable for present purposes is not whether the ban will succeed on its own terms.[29] It is that the ban was *conceivable* at all. The Overton window had shifted sufficiently for an outright prohibition to be politically viable in a mature liberal democracy. Similar proposals have since emerged in Denmark, Norway, France, Spain, Malaysia, and New Zealand.[30] Similarly, members of the European Parliament

[29] Josh Taylor, 'Reddit Launches High Court Challenge to Australia's under-16s Social Media Ban' (*The Guardian*, 12 December 2025) <https://www.theguardian.com/australia-news/2025/dec/12/reddit-high-court-challenge-social-media-ban-australia-under-16s> accessed 4 April 2026; Clare Armstrong, 'Lemon8 and Yope Put on Notice as Social Media Age Ban Targets' (*ABC News*, 2 December 2025) <https://www.abc.net.au/news/2025-12-02/lemon8-and-yope-put-on-notice-for-social-media-ban-targets/106093818> accessed 4 April 2026; Amnesty International, 'Australia: Government Should Strictly Regulate Social Media Platforms, Not Ban Children and Young People' (*Amnesty International Australia*, 9 December 2025) <https://www.amnesty.org.au/social-media-ban-for-children-and-young-people-an-ineffective-quick-fix-that-will-not-prevent-online-harms/> accessed 4 April 2026; Josh Butler, 'Two-Thirds of under-16s with Accounts on Instagram, Snapchat or TikTok Kept Access despite Ban' (*The Guardian,* 31 March 2026) <https://www.theguardian.com/australia-news/2026/mar/31/meta-tiktok-snapchat-google-under-investigation-australia-social-media-ban> accessed 4 April 2026.

[30] Ministry of Children and Families, 'Norway Moves Forward with Age Limit for Social Media' (*Government.no*, 11 June 2025) <https://www.regjeringen.no/en/whats-new/norway-moves-forward-with-age-limit-for-social-media/id3108682/> accessed 4 April 2026; Michaela Cabrera, 'French Senate Debates Social Media Ban for Children under 15' (*Reuters* , 31 March 2026) <https://www.reuters.com/technology/french-senate-debates-social-media-ban-children-under-15-2026-03-31/> accessed 4 April 2026; 'Denmark Set to Ban Social Media Platforms for Children under 15' (*Reuters*, 7 November 2025) <https://www.reuters.com/sustainability/society-equity/denmark-set-ban-social-media-platforms-children-under-15-2025-11-07/> accessed 4 April 2026; 'Malaysia Says It Plans to Ban Social Media for Under-16s from 2026' (*Reuters*, 24 November 2025) <https://www.reuters.com/world/asia-pacific/malaysia-says-it-plans-ban-social-media-under-16s-2026-2025-11-24/> accessed 4 April 2026; Renju Jose, 'New Zealand Parliament to Debate Teen Social Media Ban' (*Reuters*, 23 October

announced support for an EU-wide minimum age of sixteen for access to social media, video-sharing platforms, and AI companions.[31] Closer home, the State of Karnataka has moved towards implementing a similar ban.[32] The development is not merely Australian, it is global.

C. THE EUROPEAN UNION: THE DSA ENFORCEMENTS

The European Union's Digital Services Act (hereinafter referred to as **'DSA'**), which entered into force for Very Large Online Platforms in August 2023, represented the most comprehensive regulatory framework for online platforms in the world.[33] In furtherance of this law, the Commission opened formal proceedings against TikTok in February 2024, focusing on areas that bear a striking resemblance to the design features at issue in *KGM*. This included the assessment and mitigation of risks stemming from the design of TikTok's system, including algorithmic systems that may stimulate behavioural addictions and create so-called 'rabbit hole effects'.[34] The proceedings

2025) <https://www.reuters.com/world/asia-pacific/new-zealand-parliament-debate-teen-social-media-ban-2025-10-23/> accessed 4 April 2026.

[31] 'Children Should Be at Least 16 to Access Social Media, Say MEPs' (*News | European Parliament*, 26 November 2025) <https://www.europarl.europa.eu/news/en/press-room/20251120IPR31496/children-should-be-at-least-16-to-access-social-media-say-meps> accessed 4 April 2026.

[32] 'Karnataka Govet Bans Social Media for Children - Google Search' (*Reuters*)<https://www.reuters.com/technology/indias-tech-state-karnataka-bans-social-media-children-under-16-2026-03-06/> accessed 4 April 2026.

[33] Single Market for Digital Services (the Digital Services Act), 2022, Regulation (EU) 2022/2065, European Parliament and of the Council, 19 October 2022.

[34] 'European Commission - Press Release - Commission Opens Formal Proceedings against TikTok under the Digital Services Act'

also targeted TikTok's measures to ensure the privacy, safety, and security of minors, particularly the default privacy settings for minors as part of the design and functioning of the platform's recommender systems. The language of the Commission's press release framed the investigation as an inquiry into the *design of the system itself*, and specifically into the risks that design posed for the fundamental right to physical and mental well-being, rather than as a content moderation inquiry.[35]

The preliminary findings, issued after an in-depth investigation that included analysis of TikTok's internal data and documents, multiple requests for information, a review of scientific research on behavioural addiction, and interviews with domain experts, were damning. The Commission preliminarily found TikTok in breach of the DSA for its addictive design.[36] The features singled out were precisely those that the *KGM* jury had evaluated under American product liability doctrine, such as infinite scroll, autoplay, push notifications, and a highly personalised recommender system. The Commission's analysis identified that by constantly rewarding users with new content, certain design features of TikTok fuelled the urge to keep scrolling and shifted

<https://ec.europa.eu/commission/presscorner/api/files/document/print/en/ip_24_926/IP_24_926_EN.pdf> accessed 7 April 2026.

[35] ibid.

[36] 'Commission Preliminarily Finds TikTok's Addictive Design in Breach of the Digital Services Act' (*European Commission - European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/ip_26_312> accessed 7 April 2026.

the brains of users into what the Commission described as 'autopilot mode'.[37]

Two aspects of the preliminary findings deserve particular emphasis. *First*, the Commission found that TikTok's own risk assessment was inadequate. TikTok had failed to assess how its addictive design features could harm the physical and mental well-being of users, including minors and vulnerable adults, and had disregarded important indicators of compulsive use, such as the time that minors spent on the platform at night and the frequency with which users opened the app. *Second*, the Commission concluded that TikTok's mitigation measures were insufficient. Screen time management tools were found to be easy to dismiss and to introduce limited friction. Parental controls required additional time and skills that many parents did not possess. The Commission concluded that TikTok needed to change the *basic design* of its service, including by disabling key addictive features such as infinite scroll over time, implementing effective screen time breaks including during the night, and adapting its recommender system.[38]

The convergence between the *KGM* verdict and the Commission's TikTok proceedings is striking and, for the purposes of this article, deeply instructive. Both target the same design features. Both frame the harm in terms of behavioural addiction and its consequences for mental health. And both, crucially, locate the regulatory problem not in the

[37] ibid.
[38] ibid.

content that users encounter but in the *architecture* through which that content is delivered. Yet they arrive at this conclusion through entirely different institutional channels – a California jury applying state product liability law, and the European Commission applying provisions of the DSA. This convergence across legal traditions and regulatory modalities is perhaps the most powerful evidence that the Overton window has shifted. What was once a critique confined to advocacy organisations and academic commentary has become the operational premise of enforcement proceedings on two continents.

D. THE CONVERGENCE: FROM BIG TOBACCO TO BIG TECH

The comparison between the current wave of social media litigation and the tobacco lawsuits of the 1990s has become nearly ubiquitous, and the *KGM* verdict has only intensified the analogy.[39] There are several structural parallels. In both cases, internal corporate documents revealed that companies were aware of the addictive and harmful properties of their products. In both cases, the companies publicly denied or minimised these harms while privately strategizing to

[39] Andrew Ross Sorkin and others, 'Is Big Tech Facing a Big Tobacco Moment?' (*The New York Times*, 26 March 2026) <https://www.nytimes.com/2026/03/26/business/dealbook/meta-youtube-social-media-tobacco.html> accessed 7 April 2026.

maximise engagement. And in both cases, children and adolescents were a particularly vulnerable population.[40]

The tobacco analogy is not perfect. Tobacco litigation ultimately produced the Master Settlement Agreement of 1998, in which the four largest manufacturers agreed to restrict marketing, fund public health campaigns, and pay billions in ongoing damages.[41] Social media litigation is nowhere near that stage. Appeals are near-certain, and the legal viability of the content-versus-conduct distinction will be tested in appellate courts and potentially at the US Supreme Court. Moreover, social media platforms are not single-product companies in the way that tobacco manufacturers were. The design features at issue are deeply entangled with the features that users value.

Nevertheless, the direction of developments is instructive. It took roughly four decades from the first credible scientific warnings about smoking to the Master Settlement Agreement of 1998.[42] During that interval, the tobacco industry maintained its denials and continued to

[40] Carolina Rossini, 'Two Verdicts in Two Days: How American Courts Are Rewriting the Rules for Big Tech and Children' (*The Conversation*, 26 March 2026) <https://doi.org/10.64628/AAI.rmjdp5qcm> accessed 7 April 2026.

[41] 'The Master Settlement Agreement and Attorneys General' (*National Association of Attorneys General*) <https://www.naag.org/our-work/naag-center-for-tobacco-and-public-health/the-master-settlement-agreement/> accessed 7 April 2026.

[42] CDC, 'A History of the Surgeon General's Reports on Smoking and Health' (*Tobacco - Surgeon General's Reports*, 2 May 2024) <https://www.cdc.gov/tobacco-surgeon-general-reports/about/history.html> accessed 7 April 2026; Richard Doll and A Bradford Hill, 'Smoking and Carcinoma of the Lung' (1950) 2 BMJ 739 <https://doi.org/10.1136/bmj.2.4682.739>.

profit while the evidence accumulated.[43] The decisive shift, when it came, was not incremental but a cascade, with the release of internal documents, the collapse of industry credibility, and a rapid political consensus that permitted interventions once dismissed as confiscatory overreach.[44] The harms had always been present; what changed was the threshold at which they could no longer be treated as contested. It is not inconceivable that AI-powered platform design is in an analogous transitional moment. If the tobacco precedent is any guide, the measures that currently appear harsh, such as mandatory algorithmic audits, strict liability for addictive design, outright prohibitions on specific engagement-maximising features, may come to be seen, within a shorter time horizon than one might suppose, as the predictable consequence of harms that were visible far earlier than they were addressed.

IV. ESTRANGING THE FAMILIAR: REGULATION AS PERCEPTUAL SHIFT

[43] Naomi Oreskes and Erik M Conway, *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming* (Bloomsbury press 2010).

[44] Walter J Jones and Gerard A Silvestri, 'The Master Settlement Agreement and Its Impact on Tobacco Use 10 Years Later' (2010) 137 Chest 692 <https://doi.org/10.1378/chest.09-0982>; Robert S Wood, 'Tobacco's Tipping Point: The Master Settlement Agreement as a Focusing Event' (2006) 34 Policy Studies Journal 419 <https://doi.org/10.1111/j.1541-0072.2006.00180.x>; Lisa Bero, 'Implications of the Tobacco Industry Documents for Public Health and Policy' (2003) 24 Annual Review of Public Health 267 <https://doi.org/10.1146/annurev.publhealth.24.100901.140813>.

Before we discuss a guiding principle for dealing with these tectonic shifts in the regulatory space for AI, it is important to address an important objection to the arguments made in this paper. That the aforementioned matters involved not AI, but design of platforms, and the logic of the two are not the same. However, there are two reasons why the reasoning developed in the above-mentioned cases can be equally, if not more so, applied to AI and AI powered systems.

The first is that recommender systems implicated in above discussed matters are 'AI'.[45] The personalised content feeds are not a generic software feature. They are the outputs of machine learning models trained on vast behavioural datasets, continuously optimised to predict and exploit individual psychological responses.[46] When courts and regulators held platforms accountable for addictive design, they were in large part, even if implicitly, holding AI systems accountable. Thus, the verdict and the DSA enforcement proceedings are, in significant part, already AI regulation.

[45] Zhenhua Dong and others, 'A Brief History of Recommender Systems' (*arXiv*, 5 September 2022) <http://arxiv.org/abs/2209.01860> accessed 1 October 2024.

[46] Gediminas Adomavicius and others, 'Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects' (2013) 24 Information Systems Research 956 <https://doi.org/10.1287/isre.2013.0497>; Megan A Brown and others, 'Echo Chambers, Rabbit Holes, and Ideological Bias: How YouTube Recommends Content to Real Users' (*Brookings*, 13 October 2022) <https://www.brookings.edu/articles/echo-chambers-rabbit-holes-and-ideological-bias-how-youtube-recommends-content-to-real-users/> accessed 22 October 2025; Silvia Milano, Mariarosaria Taddeo and Luciano Floridi, 'Recommender Systems and Their Ethical Challenges' (2020) 35 AI & SOCIETY 957 <https://doi.org/10.1007/s00146-020-00950-y>.

The second reason is that AI, as it develops beyond recommender systems into "Chat-GPT" like systems, such as conversational agents, companions, and personalised assistants, it carries the same engagement-maximising logic further and deeper. Where a feed exploits attention, a conversational AI can exploit emotional dependence. A large language model can engage in sycophancy, calibrating its responses to what the user wants to hear rather than what is accurate or useful.[47] The paraphernalia of engagement optimisation has expanded, and its potential for harm has intensified accordingly. Several such cases have already emerged.[48]

These developments do not yield a ready-made regulatory blueprint. The analysis that follows is offered as reorientation rather than solution. But at a moment when AI regulatory frameworks are being actively fashioned across jurisdictions, and when foundational categories have not yet hardened nor assumptions have calcified into settled doctrine, even reorientation has value.

A. THE PERCEPTUAL PROBLEM

[47] Ssean Goedecke, 'Sycophancy Is the First LLM "Dark Pattern"' (28 April 2025) <https://www.seangoedecke.com/ai-sycophancy/> accessed 14 December 2025; Mrinank Sharma and others, 'Towards Understanding Sycophancy in Language Models' (*arXiv*, 2023) <https://doi.org/10.48550/ARXIV.2310.13548> accessed 14 December 2025.

[48] Lauren Walker, 'Belgian Man Dies by Suicide Following Exchanges with Chatbot' (28 March 2023) <https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt> accessed 4 May 2025; Kevin Roose, 'Can A.I. Be Blamed for a Teen's Suicide?' (*The New York Times*, 23 October 2024) <https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html> accessed 11 April 2026.

The regulatory failure, or at least the significant delay, that permitted the harms documented in Parts II and III to accumulate was not, at its root, a failure of law. The legal instruments were available: tort liability, product safety obligations, consumer protection duties. The failure was one of vision. Regulators, courts, and legislators saw what they expected to see. Social media platforms were treated as bulletin boards that had grown faster and more popular; recommendation algorithms were understood as editorial tools that had become more sophisticated. In each case, the surface continuity with a familiar category obscured a functional transformation beneath it. The technology had changed in kind, not merely in degree. That transformation went unseen until the harms it generated became impossible to ignore.

The *KGM* verdict, the Australian ban, and the European Commission's enforcement proceedings all rest, beneath their distinct legal reasoning, on the same foundational recognition, that reimagines these categories. The architecture of a social media feed in 2026 is not a faster version of its 2012 predecessor. It is categorically different, animated by deep learning systems trained on billions of interactions, optimised for engagement in ways that exploit psychological vulnerabilities with a precision no earlier media technology could have achieved.[49] Yet it looks the same. It presents itself in the same idiom of connection, convenience, and self-expression and that surface continuity is

[49] Adomavicius and others (n 45).

precisely the difficulty. The regulatory task is not to navigate an unfamiliar landscape. It is to see a familiar one clearly, in a new light.

What is required, then, is not a new map but a new way of looking. It is submitted that the most precise intellectual framework for understanding this regulatory demand is Viktor Shklovsky's concept of *ostranenie* – defamiliarisation, or estrangement.[50]

B. *OSTRANENIE* AND THE RECOVERY OF SIGHT

Viktor Shklovsky was a Russian literary theorist and a founding figure of Russian Formalism, the critical movement that sought to identify what made literary language distinctively literary.[51] Writing in the turbulent context of revolutionary Russia, his 1917 essay 'Art as Technique' introduced the concept of *ostranenie*, or defamiliarisation, arguing that art's essential function is to restore perception to a world that habit has rendered invisible.[52]

Shklovsky observed that habituation renders perception automatic.[53] We cease to see the things we encounter repeatedly and instead process them through recognition of 'type' rather than apprehension of the particulars. The object is not perceived but merely identified, slotted

[50] Viktor Shklovsky, 'Art as Technique' in Julie Rivkin (ed), *Literary theory: an anthology* (6. print, Blackwell 2004).
[51] 'Viktor Shklovsky | Biography & Facts | Britannica' <https://www.britannica.com/biography/Viktor-Shklovsky> accessed 7 April 2026.
[52] Ian Buchanan, 'Ostranenie' *A Dictionary of Critical Theory* (Oxford University Press 2018) <https://www.oxfordreference.com/display/10.1093/acref/9780198794790.001.0001/acref-9780198794790-e-501> accessed 7 April 2026.
[53] Shklovsky (n 49) 15.

into a familiar category and passed over. 'Art,' he wrote, "exists that one may recover the sensation of life; it exists to make one feel things, to make the stone stony."[54] The purpose of the estranging technique is to arrest this automatism, to restore to the familiar object its full weight and strangeness, to compel the viewer to actually look.

Shklovsky illustrated the technique through Tolstoy's prose, which possessed an extraordinary capacity to describe familiar phenomena as though encountering them for the first time, stripping away inherited labels and rendering objects strange, unclassified, and therefore newly visible. The estranging move was not ignorance but deliberate refusal: a refusal to allow the ready-made category to substitute for genuine perception. Crucially, Shklovsky understood that this refusal had to be built into the work itself, because the automatism it sought to overcome was not a failure of intelligence but a structural feature of how perception operates under conditions of familiarity. The danger is not that we are careless. It is that we are too fluent.

The application to AI regulation is exact, and the parallel runs deeper than metaphor. Most AI powered systems, unlike outliers such as ChatGPT or other prominent large language models, do not announce its transformations. They arrive in the idiom of the familiar, presenting themselves as faster, smarter, more convenient versions of something already known, while their functional logic has shifted in ways that inherited categories cannot capture. This is precisely the condition that

[54] ibid 16.

ostranenie was designed to address. If the artistic task is to restore genuine perception to a world that habit has rendered automatic, the regulatory task is structurally identical: to insist on actually seeing what is there, rather than confirming what was expected. This is, at root, a perceptual demand, and it has a specific institutional implication. It requires not new legal categories ready-made, but a standing disposition to treat the appearance of continuity as a reason for heightened scrutiny rather than assumed equivalence. Where a system presents itself as a familiar object operating within a familiar frame, but brings in the unique capabilities of AI, that familiarity should trigger the estranging second look, not foreclose it. The thesis of this paper is that this disposition, institutionalised as regulatory method, is what the current moment requires.

C. ESTRANGEMENT AS REGULATORY METHOD

Operationalised as regulatory method, estrangement implies three specific demands:

The first is *categorical scepticism*. Regulatory frameworks inherit their categories from prior technological moments. Section 230 of the Communications Decency Act was designed for a world in which platforms were passive conduits of user-generated speech.[55] Similarly, product liability doctrine developed for a world of physical artefacts with identifiable design defects. AI does not fit neatly within any of

[55] Communications Decency Act (n 4).

these frames, and the attempt to apply them without modification may generate sub-optimal outcomes. The estranging move is to resist the default categorisation, to ask not what an AI system *resembles* but what it *does*, and to let the answer to that question determine the applicable framework. The *KGM* litigation's central manoeuvre, treating platform design features as product design decisions rather than as speech or publication, was precisely this kind of categorical estrangement.

The second demand is *functional classification*. Regulatory frameworks naturally inherit their categories from prior technological moments, and those categories are typically organised around formal resemblance: what a system looks like, what it calls itself, what established type it most closely approximates. The estranging move is to displace this question entirely and ask instead what the system *does*, how it acts on users, what it optimises for, what it extracts, and what it conceals. The *KGM* litigation's central manoeuvre was precisely this – treating platform design features not as speech or publication, which is what they formally resembled, but as product design decisions, which is what they functionally were. Functional classification as a regulatory demand requires that this move be made as a matter of institutional standing rather than left to the ingenuity of individual litigants. It means building into regulatory architecture the habit of asking not what category a system belongs to by appearance, but what category it earns by operation.

The third demand is *pre-emptive scrutiny*. Habituation is not static, but deepens over time. The longer a technology operates within familiar

frames, the more entrenched the assumption of its innocuousness becomes, and the harder the estranging effort required to dislodge it. This means that the moment of deployment, precisely when a system presents itself most fluently in the idiom of the familiar, is also the moment when the estranging second look is most necessary and most difficult to compel. A regulatory stance oriented around *ostranenie* must therefore be pre-emptive by design, building the demand for genuine seeing into the architecture of the regulatory process itself, at the point of design and deployment, before familiarity has foreclosed the possibility of surprise.

D. INDIA AND THE ESTRANGEMENT IMPERATIVE

For jurisdictions like India, where AI adoption is accelerating across education, public services, financial technology, and agriculture,[56] the estrangement imperative carries particular urgency. The regulatory landscape is in motion, but the direction of that motion will be determined by whether new frameworks are built around functional reality or inherited appearance.

India's regulatory response to AI has accelerated considerably in recent months. The 2026 IT (Intermediary Guidelines) Amendment Rules introduced the concept of "synthetically generated information"

[56] 'AI Adoption Index: Tracking India's Sectoral Progress on AI Adoption' (*NASSCOM*) <https://www.ey.com/content/dam/ey-unified-site/ey-com/en-in/insights/ai/documents/ey-nasscom-ai-adoption-index.pdf#:~:text=The%20study%20intends%20to%20assess,Healthcare%2C%20and%20Industrials%20and%20Automotive.> accessed 11 April 2026.

(hereinafter referred to as **'SGI'**) and imposed new due diligence obligations on intermediaries that enable its creation or dissemination.[57] These include requirements to deploy technical measures preventing the generation of unlawful SGI, to label lawful SGI prominently, and to embed permanent metadata enabling traceability to the originating system. Separately, India's AI Governance Guidelines,[58] published in early 2026 under MeitY's aegis, have laid out a principles-based framework across seven governance *sutras*, recommended a new AI Governance Group and AI Safety Institute, and called for targeted amendments to the IT Act to address classification and liability gaps in relation to AI systems. Taken together, these represent the most substantive engagement with AI governance that India has yet undertaken.

And yet, viewed through the lens of the *ostranenie* argument, both developments reveal a familiar pattern. The SGI Rules, for all their technical sophistication, operate within the inherited architecture of the 2021 Intermediary Guidelines. They extend existing due diligence obligations to a new category of content rather than asking whether AI-enabled systems require a categorically different regulatory logic altogether. The AI Governance Guidelines acknowledge this gap,

[57] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026, GSR 120(E), Sec 3(i), 10 February 2026 (India), https://egazette.gov.in/WriteReadData/2026/269993.pdf, accessed 11 April 2026

[58] 'India AI Governance Guidelines: Enabling Safe and Trusted AI Innovation' (Ministry of Electronics and Information Technology, Government of India 2025) <https://indiaai.s3.ap-south-1.amazonaws.com/docs/guidelines-governance.pdf> accessed 7 December 2025.

noting that the current definition of 'intermediary' does not sit well with systems that generate or modify content autonomously, and that classification and liability frameworks need rethinking but the acknowledgment remains prospective. The estranging question — whether an AI-enabled intermediary is functionally different from its predecessor, not merely more capable — has been named but not yet systematically answered.

More regulatory developments are likely in the near term, particularly as the IT Act amendment process moves forward and the AI Governance Group becomes operational. The estrangement imperative suggests that these developments should be shaped by two commitments. *First*, functional classification should be built into any revised legislative framework. The governing question should be what a system actually does to users, not what category it formally resembles. The SGI Rules' focus on content traceability and labelling is a step in this direction, but it does not reach the deeper question of how engagement-optimising architectures should be regulated regardless of whether the content they deliver is AI-generated or not. *Second*, the pre-emptive scrutiny demand argues for embedding the second-look obligation into the regulatory process itself, through mandatory impact assessments for AI-enabled systems before deployment,[59] rather than waiting for harm to surface. India's DPI experience shows that building governance requirements into system

[59] Bernd Carsten Stahl and others, 'A Systematic Review of Artificial Intelligence Impact Assessments' (2023) 56 Artificial Intelligence Review 12799 <https://doi.org/10.1007/s10462-023-10420-8>.

architecture from the outset is achievable at scale.[60] The same logic can be applied here.

V. CONCLUSION

Shklovsky's insight was that art restores vision: it makes the stone stony again, returning to objects the weight and particularity that habit has removed. The present argument is that AI regulation must perform an analogous function. The regulatory challenge posed by AI is not that the terrain is unmapped. It is that the terrain looks the same as before while having become categorically different. What is required is not a better map. It is clearer sight: the willingness to make strange what has been made familiar, and to do so before familiarity has exacted its full cost.

The developments traced in this article, from the KGM verdict to the Australian ban, the European Commission's enforcement proceedings, and India's own emerging regulatory architecture, suggest that the Overton window has shifted, perhaps irreversibly. What was unthinkable five years ago is now operational doctrine on two continents. But a shifting window is not the same as a clear view. The risk, as regulatory frameworks consolidate around the harms already visible, is that the next generation of AI-enabled systems will again arrive in the idiom of the familiar, and again be waved through by categories that do not fit. The three demands developed in this article

[60] Clark, J., Marin, et al (2025) Digital Public Infrastructure and Development: A World Bank Group Approach. Digital Transformation White Paper, Volume 1.

— categorical scepticism, functional classification, and pre-emptive scrutiny — are not a regulatory programme. They are a prior condition, the perceptual posture without which any programme will remain, at best, a response to the last problem rather than a preparation for the next. It is submitted that this posture, institutionalised as a matter of regulatory method, is what the present moment requires. The window is open. The question is whether those responsible for what is built through it will choose to actually look.

.