International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Nudge Versus Sludge: Classifying Ai-Driven Ad-Personalization Techniques on A Manipulation–Transparency Spectrum

Ms. S. Manilakshmi[1], Dr. C. Vethirajan[2]

[1]Ph.D. Research Scholar (Full Time) Alagappa Institute of Management Faculty of Management Alagappa University, Karaikudi – 630 004
[2]Senior Professor & Head Department of Corporate Secretaryship Faculty of Management Alagappa University, Karaikudi – 630 004

Abstract
The rapid proliferation of generative and predictive artificial intelligence (AI) in digital advertising has sparked fears that personalization, once touted as a consumer benefit, has become a tool of manipulation. Behavioral economics provides a ready-made distinction between a nudge choice architecture that steers behavior while preserving autonomy and a sludge friction or influence that steers behavior against a person's interest. This paper argues that the distinction, conceived for public policy and consumer protection, has not yet been operationalized for AI-driven ad-personalization techniques. Drawing on the nudge/sludge literature, dark-pattern taxonomies, online-manipulation theory, and recent empirical work on AI ad disclosure, the paper develops a conceptual classification framework, the Manipulation–Transparency Spectrum, that positions specific AI personalization techniques dynamic creative optimization, algorithmic emotion-targeting, AI-generated scarcity claims, synthetic testimonials, conversational upselling, and personalized consent design along five operational dimensions: disclosure of AI logic, reversibility of influence, exploitation of cognitive bias, symmetry of data-consent design, and alignment of beneficiary interest. The paper develops eight theoretical propositions linking spectrum position with consumer trust and perceived manipulation and regulatory exposure. The paper discusses the implications for the design of AI ad-transparency tools, platform governance and disclosure regulation within the framework of frameworks such as the transparency obligations in Article 50 of the EU AI Act and India's Digital Personal Data Protection Act, 2023. The paper concludes with a research agenda for empirically validating and extending the framework.

Keywords: AI-driven advertising, ad transparency, consumer manipulation, nudge, sludge, dark patterns, algorithmic disclosure, consumer trust

1. INTRODUCTION

Digital advertising has become substantially automated. Real-time bidding, generative creative production, and algorithmic audience targeting now mediate most of what a consumer sees, when they see it, and in what emotional frame it is presented. Industry bodies have responded to the resulting trust deficit with disclosure guidance — the Interactive Advertising Bureau (IAB) released its first AI Transparency and Disclosure Framework in 2025 after surveys showed a sharp decline in Generation Z trust toward AI-

generated advertising (IAB, 2025). Regulators have moved in parallel: the European Union's AI Act imposes transparency obligations under Article 50 for AI-generated or manipulated content, with enforcement beginning in August 2026, while India's Digital Personal Data Protection Act, 2023 (DPDP Act) and its 2025 Rules impose consent and purpose-limitation requirements that bear directly on personalized advertising built on behavioral data.

However, neither industry frameworks nor emerging regulations give researchers a way to classify which AI-driven personalization techniques are manipulative and which are benign. A recommendation engine and a fake countdown to urgency are both 'AI-driven personalization,' but they're not the same thing in terms of consumer welfare. Behavioral economics has long distinguished between a nudge, a choice-preserving intervention that nudges behavior toward a person's own interest, and a sludge, friction or influence that serves the choice architect at the expense of the chooser (Thaler & Sunstein, 2008; Sunstein, 2021). However, this nudge/sludge distinction has been developed mainly for public-policy settings (retirement savings, organ donation, administrative burden), and has only recently been extended to commercial settings, in a general cognitive framework (Luo, Soman, & Zhao, 2023). Yet it has not yet been systematically mapped onto the technical repertoire of AI-driven ad personalization.

This paper addresses that gap conceptually. It asks: how can AI-driven ad-personalization techniques be classified along a manipulation–transparency spectrum in a way that is theoretically grounded, empirically operationalizable, and useful to the three audiences who most need it — researchers building AI ad-transparency tools, platforms designing disclosure interfaces, and regulators drafting or enforcing disclosure standards?

2. Literature Review

2.1 Nudge and sludge: theoretical foundations

Thaler and Sunstein (2008) defined nudge as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (p. 6). They also stated that a real nudge should be easy and inexpensive to avoid. Sunstein (2021) later coined sludge as the mirror-image concept: friction, complexity, or influence that makes desirable action harder or undesirable action easier, typically to the benefit of the choice architect rather than the chooser. Luo, Soman, and Zhao's (2023) meta-analytic cognitive framework shows that nudges and sludges can be computationally distinguished by the cognitive mechanisms they engage. Dual-process accounts suggest that nudges and sludges both most often operate on System-1, intuitive judgment, but diverge in whether the behavior that follows serves the actor's own stated goals. The same point is made in the Cambridge literature on "nudge/sludge symmetry" in Behavioral Public Policy: the same behavioral technique (a default, a delay, a framing choice) can be a nudge or a sludge depending entirely on whose interest it serves and how transparent its operation is to the person affected. The theoretical anchor of the framework proposed in Section 4 is this logic of transparency and interest alignment.

2.2 AI-driven personalization in digital advertising

AI personalization in advertising is based on a few different technical mechanisms: real-time assembly of ad creative from behavioral signals (dynamic creative optimization), algorithmic audience segmentation and lookalike targeting, content produced by AI (AI-written copy, synthetic imagery and increasingly synthetic testimonials or endorsements), and conversational or agentic upselling in chat-based commerce interfaces. Research is growing meta-analytically. Personalization is a consistent driver of ad effectiveness and click-through, but the trust effects are more mixed and moderated by how personalization is perceived

(helpful customization or surveillance) (meta-analytic review of personalized-advertising persuasion, Journal of Advertising family literature, 2025). Work that explicitly discusses AI personalization as perceived surveillance indicates that the same targeting mechanism can be perceived by consumers as either service or intrusion dependent on disclosure and perceived control (Systemic Analytics, 2025/2026). Specifically, in the Indian context, recent survey data among urban educated Gen Z consumers points to considerable unease with generative-AI-powered hyper-personalized advertising even when the personalization is not overtly deceptive, mostly due to opacity about how personal data is used (Frontiers in Communication, 2026) – an insight that is directly relevant to any framework designed for implementation in South Asian regulatory and market settings.

2.3 Dark patterns and online manipulation

A parallel and more design-focused literature has catalogued "dark patterns" — interface and interaction designs that manipulate users into choices they would not otherwise make. Mathur and colleagues' large-scale crawls of e-commerce sites produced an influential empirical taxonomy of dark-pattern types (urgency, scarcity, social proof, forced action, obstruction, sneaking, and misdirection), later extended to e-commerce datasets used for automated dark-pattern detection (Mathur et al., 2019; subsequent e-commerce dataset work, 2022–2023). Susser, Roessler, and Nissenbaum's (2019) philosophical account of "online manipulation" supplies the normative vocabulary this paper borrows: manipulation, on their account, is influence that works by exploiting a target's cognitive or emotional vulnerabilities rather than by appealing to their capacity for reason, and it is distinguished from legitimate persuasion precisely by its reliance on concealment or indirection. Recent advertising-specific work applies this lens directly to consent interfaces, finding that dark patterns in data-consent disclosures provoke measurable consumer reactance even when the underlying data practice is disclosed somewhere in the interface (Journal of Advertising, 2025), and social-media-specific analyses extend the same critique to algorithmic feed manipulation (Asian Journal of Business Ethics, 2026). A systematic review of the regulatory scholarship on dark patterns further shows that legal and human–computer-interaction perspectives on dark-pattern regulation have developed largely in parallel, with limited cross-pollination into marketing or consumer-behavior theory — precisely the gap this paper's framework is intended to help close (systematic review of dark-pattern regulation scholarship, 2025).

2.4 Transparency, disclosure, and consumer trust

A distinct empirical stream asks whether disclosing AI's role in an advertisement changes consumer response, with mixed and context-dependent findings. Experimental work shows that role-disclosure transparency in AI-generated advertising influences persuasion, but not uniformly — the effect depends on how AI's role is disclosed (a generic AI label versus a specific explanation of what the AI did) and on the product category (Psychology & Marketing, 2025). Related work evaluating AI-generated ad disclosure finds effects on perceived authenticity and purchase intention that vary with consumer skepticism and prior AI attitudes (University of Manchester research repository working paper). At the interface-design level, research on targeted-advertisement explanations finds that different stakeholders — consumers, advertisers, and platforms — want structurally different explanations of why an ad was shown, complicating any single-format transparency solution (arXiv preprint on multi-stakeholder ad-explanation preferences, 2024). Industry data reinforce the urgency: IAB-commissioned survey work reports a nineteen-point drop in Generation Z trust toward AI-generated advertising, prompting the IAB's first formal AI Transparency and Disclosure Framework (IAB, 2025). Taken together, this literature

establishes that transparency is necessary but not sufficient for trust repair — which is precisely why a classification framework that goes beyond a binary disclosed/undisclosed distinction is needed.

2.5 Regulatory landscape

Regulation is moving faster than academic theory in this space. The EU AI Act's Article 50 imposes transparency obligations on providers and deployers of AI systems that generate or manipulate content, including a requirement to disclose AI-generated or AI-manipulated audio, image, video, or text content, with enforcement of these obligations beginning 2 August 2026 (EU AI Act transparency-obligations guidance, 2026). In India, the DPDP Act, 2023, together with its 2025 Rules, establishes consent and purpose-limitation requirements for processing personal data including the behavioral and inferential data that power ad personalization — though commentators note the Act does not yet speak directly to algorithmic personalization or automated-decision transparency in the way the EU AI Act does (DPDP Act and AI compliance commentary, 2025). This asymmetry between a data-protection-first regime (India) and an AI-content-disclosure-first regime (EU) is itself a source of comparative research opportunity, and is revisited in Section 6.3.

3. Research Gap

The reviewed literature establishes three things without yet connecting them: (1) nudge/sludge theory gives a principled way to distinguish autonomy-preserving from autonomy-undermining influence; (2) dark-pattern and online-manipulation research has produced granular taxonomies of manipulative interface techniques, largely outside advertising-specific personalization; and (3) AI-disclosure research shows that transparency's effect on trust is conditional rather than uniform. No published framework yet classifies the specific technical repertoire of AI-driven ad personalization — dynamic creative optimization, algorithmic emotion-targeting, AI-generated scarcity and urgency claims, synthetic testimonials, and conversational upselling — along an operational manipulation–transparency spectrum grounded in nudge/sludge theory. This is the gap the remainder of the paper addresses.

4. Proposed Conceptual Framework: The Manipulation–Transparency Spectrum

4.1 Classification dimensions

The framework positions each AI-driven personalization technique along five dimensions, each anchored at a "nudge" pole and a "sludge" pole:

Dimension	Nudge pole	Sludge pole
Disclosure of AI logic	Explicit, timely, specific	Absent, generic, or buried in policy text
Reversibility of influence	Easy to opt out of or undo	Difficult, delayed, or effectively irreversible
Cognitive-bias exploitation	Minimal reliance on bias	Deliberate use of scarcity, urgency, or fabricated social proof
Data-consent symmetry	Clear, opt-in, granular choice	Asymmetric, pre-checked, or friction-loaded refusal
Beneficiary alignment	Serves consumer's stated interest	Serves advertiser interest at consumer expense

A technique's aggregate position on the spectrum is not assumed to be a simple additive score; rather, each dimension functions as a necessary condition check, consistent with the online-manipulation literature's emphasis that concealment (low disclosure) combined with vulnerability-exploitation (bias exploitation) is the specific combination that constitutes manipulation, rather than either factor alone (Susser, Roessler, & Nissenbaum, 2019).

4.2 Illustrative mapping of techniques

Technique	Typical disclosure	Typical bias reliance	Illustrative spectrum position
Dynamic creative optimization (content/imagery matched to inferred preference)	Low–moderate	Low	Nudge-leaning, contingent on disclosure
Algorithmic emotion/mood targeting	Low	Moderate–high	Sludge-leaning
AI-generated scarcity or countdown claims	Low	High (urgency bias)	Sludge
Synthetic testimonials or AI-fabricated reviews	Very low (often undisclosed)	High (social-proof bias)	Sludge
Conversational AI upselling in chat commerce	Moderate	Moderate	Midpoint; depends on reversibility of the upsell
Personalized recommendation with visible "why am I seeing this" control	High	Low	Nudge

4.3 Positioning logic

The framework's central claim is that a technique's position on the spectrum is not fixed by the underlying AI method but by implementation choices layered on top of it — the same recommendation algorithm can be deployed as a nudge (disclosed, reversible, bias-neutral) or a sludge (undisclosed, hard to reverse, bias-exploiting). This reframes the transparency-tool design problem: the goal is not to label techniques as inherently good or bad, but to give consumers and auditors visibility into the implementation choices that move a technique along the spectrum.

5. Theoretical Propositions

1. **P1:** The lower a technique's score on AI-logic disclosure, the higher its perceived manipulativeness, holding targeting accuracy constant.
2. **P2:** Disclosure alone will not fully offset perceived manipulation for techniques that score high on cognitive-bias exploitation (e.g., fabricated urgency), because concealment is not the only driver of manipulativeness.
3. **P3:** Techniques positioned toward the sludge pole on beneficiary alignment will show a stronger negative relationship with brand trust than techniques positioned toward the sludge pole on reversibility alone.
4. **P4:** Consumer classification of a technique's spectrum position will diverge from practitioner

classification, with practitioners systematically underestimating the sludge-ward position of bias-exploiting techniques relative to consumer perception.

5. **P5:** AI ad-transparency tools that surface implementation-level attributes (disclosure specificity, reversibility, consent symmetry) will produce larger trust gains than tools that surface only a binary "AI-generated" label.
6. **P6:** The trust benefit of a given transparency-tool feature will be moderated by regulatory context, being larger in jurisdictions with mandatory AI-content disclosure (e.g., under EU AI Act Article 50) than in jurisdictions with data-consent-only regimes (e.g., under India's DPDP Act) — because ex-ante regulatory salience shapes baseline consumer expectations.
7. **P7:** Techniques that combine low disclosure with low consent symmetry (jointly sludge-ward on two dimensions) will show a super-additive, not merely additive, negative effect on perceived consumer autonomy.
8. **P8:** Over repeated exposure, consumers will recalibrate trust toward AI-personalized advertising more readily for techniques that remain stable in spectrum position across interactions than for techniques whose transparency is inconsistently applied.

6. Implications

6.1 Theoretical implications

The framework extends nudge/sludge theory, developed mainly for public-policy choice architecture, into the specific technical domain of AI-driven advertising, and it gives the dark-pattern and online-manipulation literatures a shared vocabulary with marketing and consumer-behavior research — addressing the cross-disciplinary gap identified in the dark-pattern regulation literature (Section 2.3). It also reframes "AI disclosure" research: rather than treating disclosure as a single binary variable, the framework decomposes it into disclosure specificity, reversibility, bias-reliance, consent symmetry, and beneficiary alignment as separately manipulable design variables, consistent with recent findings that disclosure effects on trust are conditional rather than uniform (Section 2.4).

6.2 Managerial and practitioner implications

For platforms and advertisers, the framework offers a self-audit tool: a technique that scores toward the sludge pole on two or more dimensions is a candidate for redesign before it becomes a regulatory or reputational liability. For designers of AI ad-transparency tools specifically, the framework argues against single-attribute solutions (a generic "AI-generated" badge) in favor of multi-attribute disclosure interfaces that expose disclosure specificity, reversibility, and consent symmetry separately — consistent with evidence that different stakeholders want structurally different explanations (Section 2.4).

6.3 Policy and regulatory implications

The framework can help reconcile the EU's AI-content-disclosure-first approach with India's data-consent-first approach by showing that both regimes are, in effect, regulating different dimensions of the same underlying spectrum — the EU AI Act's Article 50 principally targets the disclosure dimension, while the DPDP Act principally targets the consent-symmetry dimension. A comparative policy paper building on this framework could evaluate whether regulating only one dimension leaves the others (bias exploitation, reversibility, beneficiary alignment) unaddressed, and whether IAB-style industry self-regulation closes that residual gap in practice.

7. Limitations and Future Research Directions

As a conceptual paper, this framework has not yet been empirically validated; the dimensions and illustrative technique mapping in Section 4 are theoretically derived and require validation through the two-phase mixed-methods design outlined in the companion research concept note — an expert-panel or Delphi study to refine and weight the dimensions, followed by a scenario-based consumer study testing the propositions in Section 5. Future work should also test whether the five dimensions are conceptually and statistically distinct (via exploratory and confirmatory factor analysis) rather than collapsing into one or two latent factors, examine cross-cultural variation in spectrum perception (extending the India-specific evidence noted in Section 2.2 to other jurisdictions), and track how spectrum positioning shifts as platforms adapt to EU AI Act enforcement after August 2026.

8. Conclusion

AI-driven ad personalization is neither uniformly beneficial nor uniformly manipulative, and treating "AI in advertising" as a single undifferentiated phenomenon obscures the design choices that determine consumer welfare. By grounding a classification framework in nudge/sludge theory and operationalizing it with five implementation-level dimensions, this paper offers researchers, platform designers, and regulators a shared, testable vocabulary for distinguishing AI-driven nudges from AI-driven sludge — and a research agenda for validating that vocabulary empirically.

References

1. Digital Personal Data Protection Act, 2023, and DPDP Rules, 2025 (India) — compliance commentary. (2025).
2. European Union. Artificial Intelligence Act, Article 50 — Transparency obligations. Enforcement guidance. (2026).
3. Interactive Advertising Bureau (IAB). (2025). AI Transparency and Disclosure Framework.
4. Luo, Y., Li, A., Soman, D., & Zhao, J. (2023). A meta-analytic cognitive framework of nudge and sludge. Royal Society Open Science, 10(11), 230053.
5. Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. Proceedings of the ACM on human-computer interaction, 3(CSCW), 1-32.
6. Yada, Y., Feng, J., Matsumoto, T., Fukushima, N., Kido, F., & Yamana, H. (2022, December). Dark patterns in e-commerce: a dataset and its baseline evaluations. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 3015-3022). IEEE.
7. Multi-stakeholder preferences for targeted advertisement explanations. (2024). arXiv preprint.
8. Nudge/sludge symmetry: On the relationship between nudge and sludge and the resulting ontological, normative, and transparency implications. Behavioural Public Policy (Cambridge Core).
9. Personalization and privacy perceptions among urban, educated Indian Gen Z consumers in generative-AI-driven hyper-personalized advertising. (2026). Frontiers in Communication.
10. Role disclosure transparency in AI-generated ads and persuasion. (2025). Psychology & Marketing.
11. Darda, P., & Gupta, O. J. (2026). Shadows in the feed: a critical examination of dark patterns and user manipulation on social media. Asian Journal of Business Ethics, 15(2), 295-315.
12. Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. Geo. L. Tech. Rev., 4, 1.

13. Sunstein, C. R. (2021). Sludge: What stops us from getting things done and what to do about it. MIT Press.
14. Nafei, A. (2026). When Personalization Feels Like Surveillance: A Marketing Framework for Artificial Intelligence-Personalized Advertising. Systemic Analytics, 4(2), 102-117.
15. Thaler, R. H., Sunstein, C. R., & Milligan, S. (2009). Nudge: Improving decisions about health, wealth, and happiness (Vol. 47, p. 9). London: Penguin books.
16. Yao, Y. (2025, July). Transparent technology: Evaluating the impact of AI-generated ad disclosure on consumer trust, authenticity, and purchase intention. In Global Marketing Conference: 2025 Global Marketing Conference at Hong Kong (p. 289).
17. Ju, I., Ham, C. D., & Yel, E. (2026). Dark patterns in data-consent disclosures and consumer reactance to online behavioral advertising. Journal of Advertising, 55(3), 307-325.
18. Yi, W., & Li, Z. (2025). Mapping the scholarship of the regulation of dark patterns: A systematic review of concepts, regulatory paradigms, and solutions from law and HCI perspectives. Computer Law & Security Review, 59, 106225.