

Ethical Issues in AI-Based Digital Marketing: A Corporate Social Responsibility (CSR) Perspective

GAHAN K NAIK

Dayananda sagar college of engineering

Abstract - Artificial intelligence has moved digital marketing from broadcasting messages to inferring people. Segments are built from behavioral traces instead of declared preferences, creative is generated and varied automatically, and bidding decisions are made by systems no individual marketer inspects. The ethical problems that follow are familiar enough in outline: opaque consent, inferred sensitive attributes, discriminatory delivery, manipulative interface design, undisclosed synthetic content. What is less familiar is how firms manage them. The usual apparatus is a pre-launch compliance checklist, a legal sign-off, and a broadly uniform allocation of review effort that treats a contextual retail campaign and an autonomously optimized credit campaign as equivalent risks. This chapter asks whether machine learning and explainable AI (XAI) can be turned on the problem they helped create, and what a corporate social responsibility (CSR) framing adds to a compliance framing. The argument draws on the marketing ethics, algorithmic fairness and privacy literatures, on three strategic analysis lenses (PESTEL, Porter's Five Forces and SWOT), and on an empirical illustration built from a 1,920-record campaign governance dataset covering eight product categories, six targeting methods and four consent bases. Gradient-boosted and random forest models predicted realized complaint intensity with a mean absolute error of 3.18 complaints per million impressions, against 5.77 for the compliance-checklist baseline, a reduction of 44.8 per cent. Campaigns that went on to produce a material ethical incident were identified with an ROC-AUC of 0.904. SHAP analysis put consent clarity, audience vulnerability, allocated assurance effort, disclosure practice and model autonomy at the top of the attribution; product category accounted for 2.7 per cent of the total and the three commercial variables together for 3.7 per cent. A reallocation simulation then showed that redistributing an unchanged budget of assurance review hours by predicted risk cut unmitigated ethical exposure by 50.4 per cent, with no additional governance headcount. The chapter offers an integrated framework linking prediction, explanation and marketing decision-making, together with guidance for firms at different levels of governance maturity.

Introduction

Advertising has always been in the business of persuasion, and persuasion has always raised ethical questions. What has changed is the resolution. A television advertisement in 1990 addressed a demographic. A programmatic advertisement in 2026 addresses an inferred individual, selected in low milliseconds by a system that has assembled a behavioral profile from sources the recipient could not enumerate if asked. The persuasive claim may be identical in both cases. The apparatus behind its delivery is not, and most of the ethical

difficulty now attaching to digital marketing belongs to the apparatus.

None of this is speculative. Sweeney (2013) showed that searches for names statistically associated with Black Americans returned advertisements suggesting arrest records far more often than searches for names associated with white Americans. Datta et al. (2015), using an automated experimental tool, found that simulated male job seekers saw advertisements for high-paying executive roles more frequently than otherwise identical female profiles. Ali et al. (2019) reported something more uncomfortable: skewed delivery along racial and gender lines could emerge from a platform's own optimization machinery even where the advertiser had specified entirely neutral targeting. Lambrecht and Tucker (2019) then traced an apparently discriminatory outcome in STEM career advertising to an economic mechanism. Younger women were a more expensive audience to reach, so a cost-minimizing algorithm reached fewer of them. No prejudice was required anywhere in the system. That last result is the one conventional compliance is least equipped to handle, because conventional compliance looks for intent.

A parallel literature has documented the design side. Mathur et al. (2019) crawled roughly eleven thousand shopping websites and identified 1,818 dark patterns across fifteen types. Nouwens et al. (2020) examined consent management platforms on popular United Kingdom websites and found that only a small minority met even a minimal reading of the applicable law; implicit consent and asymmetric button design were routine. Luguri and Strahilevitz (2021) moved from observation to experiment, and found that mild dark patterns more than doubled acceptance of a dubious paid programme while aggressive ones came close to quadrupling it. The cumulative implication is awkward for anyone relying on consent records. In a large fraction of cases those records measure interface design as much as they measure preference. Against all this, the governance apparatus has barely moved. Most firms still run a pre-launch compliance checklist, completed by the campaign team and reviewed by legal or brand, scored at the level of product category and channel. A credit card campaign on paid social is flagged as riskier than a homeware campaign on search. The checklist asks whether consent was obtained, whether the creative complies with the applicable advertising code, and whether prohibited targeting categories were avoided. It is completed once, before launch, and it returns a category.

The category is the weak link. It captures what is being advertised and almost nothing about how the campaign was actually built, and the risk now lives in how it was built. Two credit campaigns can share a regulatory category, share a creative approval, and differ by an order of magnitude in the

depth of inference applied to their audiences, the quality of the consent underlying the data, the autonomy granted to the bidding system, and the share of creative that was generated instead of written. A conventional checklist records none of those four. All four are choices the firm made.

Assurance effort is allocated with the same bluntness. Six or seven hours of ethics and legal review per campaign is a common figure, spread more or less evenly because spreading it evenly is administratively simple and easy to defend. The result is misallocation running in both directions at once. A contextual campaign for consumer electronics absorbs review it does not need. An autonomously optimized campaign for a gambling product, aimed at an audience selected partly on inferred financial stress, gets the same allocation and is badly under-reviewed. The total budget may be perfectly adequate. Its distribution is not.

Meanwhile the raw material for a better allocation has been accumulating. Ad servers record every impression, placement and frequency figure. Customer data platforms record which attributes built each audience and where those attributes came from. Consent platforms record the basis on which each record was collected and the interface through which the choice was made. Complaint queues, subject access requests, opt-out events, platform enforcement notices and moderation escalations all record what happened next. Two or three years of that gives a firm a transactional history linking configuration to consequence. Almost no one uses it for governance. Everyone uses it for optimization.

Artificial intelligence has begun to close the gap, and I want to acknowledge the irony rather than smooth it over: the same family of methods that produces the delivery skew Ali et al. (2019) documented can be used to predict which campaigns will produce it. Mittelstadt et al. (2016) mapped the ethical concerns raised by algorithms and concluded that auditing was the most developed of the available responses. Raji and Buolamwini (2019) showed that published audits changed commercial behavior. Metaxa et al. (2021) consolidated the field into a methodological programme. So the building blocks exist and are well documented.

The difficulty is what comes after. Much of the published work stops once a disparity has been measured or a fairness metric computed. That is a finding. A marketing organization cannot act on the statement that a campaign exhibits a 0.12 disparity in delivery rates. It can act on the statement that this campaign should not launch until the consent basis for its audience is replaced, or that forty campaigns out of six hundred warrant full ethical review while the rest proceed on a light-touch check. Getting from a measurement to an allocation of scarce human attention is where governance programmers usually fail, and it is not a modelling problem.

Trust is the second difficulty, and it runs in the opposite direction from the one usually assumed. A performance marketer told to pause a campaign that is hitting its cost-per-acquisition target, because a model has assigned it an elevated ethical risk score, will want to know why. If the answer is that the model returned a high number, the recommendation will be overridden, and it should be. An unexplained instruction to destroy measurable commercial value is not a control. It is an assertion of authority. Dietvorst et al. (2015) documented how

quickly people abandon algorithmic advice after seeing it err; Logg et al. (2019) found that lay users often prefer algorithmic judgement to human judgement. Read together, the two literatures suggest acceptance turns on whether the reasoning can be inspected.

This is where explainable AI stops being an academic nicety and becomes an operational requirement. Ribeiro et al. (2016) introduced LIME, approximating a complex model locally with an interpretable surrogate. Lundberg and Lee (2017) introduced SHAP, with additivity guarantees and an exact algorithm for tree ensembles. Applied to a campaign risk score, SHAP turns an opaque number into a decomposition: this campaign is flagged because the consent behind sixty per cent of its audience came from a bundled interface, because the inferred attributes include proxies for a health condition, and because the bidding system is running at maximum autonomy with no disclosure on the generated creative. Each clause points at a fix. Rudin (2019) has argued that post-hoc explanation of a black box is an inferior substitute for an interpretable model in high-stakes settings. I take that seriously and return to it in Section 6. The position here is narrower: a decomposed explanation is adequate when the decision it supports is where to send a human reviewer.

So this chapter treats ethical governance of AI-based marketing as an integrated system, not a compliance exercise and not a modelling exercise either. Three components are inseparable. Prediction at the level of the individual campaign, replacing a category-level checklist score with a configuration-sensitive estimate. Explanation, turning that estimate into named and addressable causes. And decision, turning the explained estimate into a reallocation of assurance effort the organization is already paying for. The CSR framing determines what the system is for. A compliance framing asks whether the firm can defend what it did; a CSR framing asks whether the firm should have done it. Those questions diverge most sharply where regulation has not yet arrived.

Section 4 uses a campaign governance dataset of 1,920 records spanning eight product categories, eight channels, six targeting methods, four consent bases and four regulatory regimes. It is synthetic. I state that here, not in a closing limitations paragraph, because it governs what the results can support. The illustration demonstrates that the three components compose into a working system and that the structural claims survive the composition. It says nothing about the magnitude of any effect in a real marketing organization.

The chapter proceeds as follows. Section 2 sets out the theoretical framework, drawing on CSR and stakeholder theory, contextual integrity, legitimacy and institutional theory, the resource-based view, and the strategic analysis frameworks. Section 3 reviews the empirical literature and identifies four gaps. Section 4 presents the illustration. Section 5 examines how the components interact. Section 6 sets out the challenges, including those the framework itself creates. Sections 7 to 10 cover recommendations, industry cases, research directions and conclusions.

2. THEORETICAL FRAMEWORK

Carroll's (1979, 1991) account of corporate social responsibility gives the most direct reason for treating this as something other than a compliance problem. Carroll arranged corporate obligations into four layers: economic responsibility to be profitable, legal responsibility to obey the law, ethical responsibility to do what is right beyond what is required, and discretionary or philanthropic responsibility. The ethical layer is defined by the space between what the law demands and what stakeholders legitimately expect. AI-based marketing currently occupies an unusually large volume of that space. Inferring a consumer's likely pregnancy from purchase timing was lawful in most jurisdictions long before anyone tried to regulate it, and remains lawful in several. Establishing that it may be done does not answer whether it should be.

Stakeholder theory supplies the second lens. Freeman (1984) described the firm as a nexus of relationships with groups that affect or are affected by its objectives, and Donaldson and Preston (1995) separated the descriptive, instrumental and normative readings of that claim. In digital marketing the consumer occupies an odd dual position: the audience is both the stakeholder whose interests are at issue and the raw material from which the advertising transaction is manufactured. Mitchell et al. (1997) explain why this stakeholder gets deprioritized so consistently. Their typology scores stakeholders on power, legitimacy and urgency, and the individual data subject scores high on legitimacy and close to zero on the other two. Regulators supply the missing power. That is why regulatory arrival changes corporate behavior so much more reliably than ethical argument does, and why a CSR programmer built purely in anticipation of regulation is hard to distinguish from compliance with better branding.

These two lenses pull against each other, and a framework that pretends otherwise will not survive contact with a marketing team. Personalization is not intrinsically hostile to consumers. Relevant advertising lowers search costs and surfaces useful offers. Bleier and Eisenbeiss (2015) found that personalized banner advertising raised click-through where the retailer was trusted. Tucker (2014) found that after a social network gave users finer privacy controls, users became more responsive to personalized advertising, not less. The question is not whether to personalize but on what basis, with what disclosure, and within what limits. Treat all personalization as suspect and the framework will be ignored, deservedly, because it strips value from both sides of the exchange without identifying where harm actually occurs.

Nissenbaum's (2004, 2010) theory of contextual integrity identifies that mechanism more precisely than any competing account. Informational norms are context-relative, and a privacy violation consists in a flow that breaches the norms of the context where the information was generated. This explains a pattern consent-based analysis cannot. Someone who searches for information about a medical condition has not thereby agreed that the fact may travel to an advertising exchange and shape their insurance offers, even where a lawful basis can be constructed from an accepted privacy policy. The theory is operationally useful because it directs attention to the provenance of each attribute in an audience definition. Provenance is a recordable property of a campaign. A consent flag is a much weaker signal.

Legitimacy and institutional theory explain why firms adopt ethics programmes and why so many underperform. Suchman (1995) defined legitimacy as a generalized perception that an entity's actions are proper within a socially constructed system of norms. DiMaggio and Powell (1983) identified coercive, mimetic and normative pressures as the mechanisms of institutional isomorphism. Both predict what we observe: responsible marketing commitments cluster in time, resemble one another closely in language, and appear fastest at the firms most exposed to scrutiny. Both also predict ethics-washing. A commitment adopted to secure legitimacy is optimized for perception, and perception and effect diverge as soon as the effective version becomes expensive.

The manipulation literature supplies the normative core. Calo (2014) described digital market manipulation as the systematic exploitation of cognitive bias, made possible once a firm knows a consumer well enough to find the moment and the frame in which their judgement is weakest. Susser et al. (2019) refined this into an account of online manipulation as covert influence that targets and exploits vulnerabilities, distinguishing it from persuasion, which addresses rational agency, and from coercion, which overrides it. Yeung (2017) named the dynamic, continuously updated variant the hypernudge. The distinction these authors draw is the one that matters for practice. Mass persuasion addresses a population and can be evaluated by inspecting the message. Personalized persuasion addresses an individual through a vulnerability the system discovered by itself, and inspecting the message tells you very little. Creative approval sits at the Centre of most marketing compliance processes, and this is why it has become such a poor proxy for ethical review.

Zuboff (2019) offers the broadest critique, arguing that behavioral surplus extraction constitutes a novel economic logic. I include it because it frames the stakes accurately, and I do not rely on it operationally, because a systemic critique of an economic order does not generate campaign-level decision rules. A marketing director who accepts the diagnosis in full still has to decide on Monday which of four hundred live campaigns to review. This chapter works at that level deliberately. Doing so brackets a question the systemic literature is right to keep open.

The resource-based view explains why an ethics-led approach might produce durable advantage. Barney (1991) held that sustained advantage requires resources that are valuable, rare, imperfectly imitable and non-substitutable; Teece et al. (1997) extended this to the capability of reconfiguring resources as conditions change. The models and fairness metrics in this chapter satisfy none of those conditions. They are published, and a competent team can reproduce them in an afternoon. What does satisfy them is a first-party data asset resting on consent that was actually given, linked to a multi-year record of which configurations produced complaints, and governed by a process the organization genuinely follows. Martin et al. (2017) support the value side of the claim, finding that data privacy performance affected firm value and that breaches produced negative spillovers onto rivals, which gives the asset a public-good character that complicates the competitive story. Consumer trust, once lost, is slow and expensive to rebuild. That is the practical definition of imperfect imitability.

Strategic analysis frameworks test organizational feasibility. PESTEL describes a macro-environment that has shifted decisively. Politically and legally, the General Data Protection Regulation (2016) made consent and purpose limitation the defaults; the Digital Services Act (2022) prohibited advertising based on profiling with special categories of personal data, and profiling-based advertising directed at minors; the EU Artificial Intelligence Act (2024) prohibited manipulative techniques that materially distort behavior, prohibited exploitation of vulnerabilities arising from age, disability or economic situation, and imposed transparency obligations on synthetic content. India's Digital Personal Data Protection Act (2023) requires consent that is free, specific, informed, unconditional and unambiguous, and restricts behavioral tracking and targeted advertising aimed at children. The Central Consumer Protection Authority (2023) named thirteen specific dark patterns. Economically, the decline of stable cross-site identifiers has raised the cost of behavioral measurement and improved the relative economics of contextual approaches. Socially, privacy expectations have risen, unevenly by generation and market. Technologically, on-device processing, privacy-enhancing techniques and generative creative are changing what is possible faster than governance is changing what is permitted. Environmentally, the compute intensity of continuous auction-time optimization is small but no longer negligible in marketing sustainability reporting.

Porter's (1980) Five Forces clarifies where the ethical leverage actually sits. Supplier power dominates, and it belongs to the platforms, which control the data, the auction, the delivery algorithm and the rules governing all three. An advertiser worried about delivery skew is a buyer who cannot inspect the mechanism being purchased. Buyer power is concentrated among the largest spenders, which is why coordinated action by major advertisers has historically moved platforms faster than regulatory pressure of comparable intensity. Substitutes are real and rising: contextual advertising, retail media on declared first-party data, and creator partnerships all substitute for behavioral targeting, and their relative performance improves as identifiers become scarcer. New entry into privacy-preserving advertising technology is easy at the tooling layer and very hard at the data layer. Rivalry is intense and conducted largely on short-horizon performance metrics. That last point is the structural reason ethical restraint is competitively costly in the short run, and the reason collective mechanisms matter.

SWOT summaries the internal picture. The strengths of a prediction-led approach are campaign-level granularity, the ability to learn from every complaint and enforcement event, and an auditable record that is genuinely useful when a regulator asks how a decision was reached. The weaknesses are dependence on complaint data that is noisy, incomplete and biased toward consumers who know how to complain, and a real risk that quantification confers false precision on contestable judgements. The opportunities are regulatory readiness ahead of enforcement deadlines, differentiation where trust is a purchase criterion, and reuse of governance data to improve targeting quality instead of only constraining it. The threats are ethics-washing, gaming by the teams being

measured, and the possibility that a documented risk score becomes evidence of knowledge in later litigation.

Complex adaptive systems thinking supplies the final caution. The advertising ecosystem is a network of semi-autonomous agents, consumers and advertisers and platforms and intermediaries and regulators and now generative systems, each responding to the others with a lag. Consumers who learn that a firm infers sensitive attributes adjust what they disclose, which degrades the data. Teams who learn which variables drive the score adjust those variables first, sometimes exactly as intended and sometimes as pure appearance management. Platforms change delivery algorithms without notice and invalidate historical relationships. Any measure that becomes a target degrades as a measure, and a governance model is a measure deliberately made into a target. Section 6 treats this as a live risk.

3. LITERATURE REVIEW

The personalization-privacy tension is the oldest strand of the relevant literature, and its central finding is a paradox. Consumers state strong privacy preferences and then behave inconsistently with them. Reviewing the behavioral evidence in Science, Acquisti et al. (2015) attributed this to three properties of privacy preferences: uncertainty about the consequences of disclosure, dependence on context, and malleability in response to how choices are presented. The third has the sharpest implication for practice. A consent rate is partly a measure of interface design and only partly a measure of preference, which undermines the evidentiary value of exactly the artefact most compliance processes lean on.

Marketing scholarship has examined when personalization is welcomed. Aguirre et al. (2015) separated covert from overt personalization and found that covert collection reduced click-through by triggering feelings of vulnerability, while overt collection paired with transparent explanation increased it. Tucker's (2014) control finding points the same way, as does Martin and Murphy's (2017) synthesis, which argues that data privacy should be handled as a marketing strategy variable and not as a legal constraint. Running through all three is a useful practical point: the ethically preferable configuration is frequently also the commercially superior one. Not always, though, and the exceptions are precisely the cases that need governance instead of persuasion.

The discrimination literature developed separately and with a more adversarial method. Sweeney (2013) used systematic search queries to demonstrate racially patterned delivery of arrest-record advertising. Datta et al. (2015) built AdFisher to detect differential delivery automatically. Speicher et al. (2018) showed that platform targeting tools could reach discriminatory audiences through correlated proxies even where explicit protected attributes were unavailable. Ali et al. (2019) then established the result that matters most for governance: the platform's own relevance and cost optimization could skew delivery of housing and employment advertisements with no discriminatory input from the advertiser whatsoever. Lambrecht and Tucker (2019) identified the economic mechanism in a related case. Collectively these papers show

that intent-based compliance cannot detect a substantial class of the harms at issue.

Barocas and Selbst (2016) supplied the legal-theoretical account connecting these findings, identifying the mechanisms by which data mining produces disparate impact without discriminatory intent: definition of the target variable, labelling of training examples, collection of training data, feature selection, and the use of proxies. Mittelstadt et al. (2016) mapped the wider ethical terrain and identified auditing as the most tractable response. Sandvig et al. (2014) had already set out the methodological options for auditing from outside, Metaxa et al. (2021) consolidated the field, and Raji and Buolamwini (2019) demonstrated that published audits changed commercial systems. That last paper remains the strongest available evidence that external scrutiny functions as a control.

Work on dark patterns is more recent and unusually well quantified. Gray et al. (2018) produced the first widely adopted taxonomy, distinguishing nagging, obstruction, sneaking, interface interference and forced action. Mathur et al. (2019) supplied measurement at scale and connected each observed pattern to the cognitive mechanism it exploits. Nouwens et al. (2020) applied the approach to consent interfaces and found that design choices with no legal justification, such as omitting a reject-all option from the first layer, produced large changes in recorded consent. Luguri and Strahilevitz (2021) established causal effect sizes experimentally, and found that less educated respondents were disproportionately susceptible to aggressive patterns, which turns a design question into a distributive one. Waldman (2020) argued that the phenomenon is a structural feature of an attention-based business model and not a collection of individual bad choices. That argument sits uncomfortably with any firm-level remedy, including the one proposed here.

Generative AI has introduced concerns the empirical literature has not caught up with, and this part of the review is thinner than I would like. Synthetic creative can now be produced and varied faster than humans can review it, which breaks the assumption underlying every creative approval workflow in use. Synthetic endorsers, cloned voices and AI-generated testimonial content raise disclosure questions that existing rules address only partially. The Federal Trade Commission's (2023) revised endorsement guides cover material connections and fabricated reviews, but were not written for content whose speaker does not exist. The EU Artificial Intelligence Act (2024) imposes transparency obligations on synthetic content, supplying a legal floor without settling what adequate disclosure looks like in a six-second format. Empirical work on consumer response to disclosed versus undisclosed synthetic advertising is sparse. Section 9 treats it as a priority.

The explainability literature developed outside marketing but transfers directly. Ribeiro et al. (2016) introduced LIME. Lundberg and Lee (2017) introduced SHAP, unifying several attribution methods under a game-theoretic formulation with an exact polynomial-time algorithm for tree ensembles, which is what makes the approach practical at campaign volume. Doshi-Velez and Kim (2017) pressed for rigour in defining what interpretability is for, separating application-grounded from functionally-grounded evaluation. Rudin (2019) offered the

strongest dissent, arguing that in high-stakes settings post-hoc explanations of black boxes can mislead and that inherently interpretable models should be preferred. That objection bears directly on the design in Section 4 and is addressed in Section 6.

Research on CSR outcomes supplies the business-case context, and it is considerably more equivocal than advocacy literature suggests. Sen and Bhattacharya (2001) found that consumer response depended on congruence between the initiative and the firm's core business, and that poorly matched initiatives could reduce evaluations. Luo and Bhattacharya (2006) found CSR associated with customer satisfaction and market value, but conditionally on innovation capability and product quality, so that CSR could actually destroy value at firms performing poorly on those dimensions. Du et al. (2010) identified the central communication dilemma: telling stakeholders about an initiative is necessary for it to shape perception, and simultaneously invites scepticism about motive. For responsible marketing the congruence finding is encouraging, since data ethics is unusually well matched to a business built on consumer data. The communication finding is a warning.

Work on trust in algorithmic recommendation completes the picture, and it does not support a simple story. Dietvorst et al. (2015) documented algorithm aversion. Logg et al. (2019) found the opposite pattern elsewhere, with lay participants weighting algorithmic advice more heavily than human advice. Reconciling them requires attention to whether the user can inspect the reasoning and whether they can adjust the output; both raise acceptance. That is the empirical basis for the choice in Section 4 to expose attributions instead of scores, and to treat model output as an allocation of attention instead of a verdict.

Regulatory scholarship has focused heavily on explanation. Wachter et al. (2017) argued that the General Data Protection Regulation creates no legally binding right to explanation of specific automated decisions, only a narrower right to information about system logic; the reading has been contested but not displaced. Kaminski (2019) argued for a more expansive interpretation grounded in the Regulation's accountability architecture. Veale and Zuiderveen Borgesius (2021) assessed the Artificial Intelligence Act proposal and noted that the prohibition on manipulative techniques carried a harm threshold high enough to exclude most ordinary commercial persuasion. That observation matters here, because it locates the legal floor well below the ethical questions a marketing organization actually faces. Which is another description of Carroll's ethical layer.

Two further strands deserve brief mention. Applied machine learning in marketing is overwhelmingly concerned with optimization, with attribution and budget allocation and churn and propensity, and only marginally with governance, so the methods needed for the framework proposed here are mature while their application to this problem is not. And the Indian regulatory literature is developing quickly around the Digital Personal Data Protection Act (2023) and the dark patterns guidelines, with the children's provisions imposing constraints that have no direct European analogue.

Four gaps follow. First, the discrimination and dark patterns research documents harms thoroughly and at scale, but almost never connects a measured harm to an allocation of the firm's

finite governance capacity. Second, ethical assessment in practice operates at the level of the product category or the campaign brief, while the evidence indicates that the operative variables are configuration choices made after the brief is approved. Third, explainability research has established the methods but has rarely treated the marketer as the user, so explanations arrive in the vocabulary of features instead of decisions. Fourth, and most important, the CSR literature treats responsible marketing as a communicative and strategic posture while the operational literature treats it as compliance. Very little work connects a CSR commitment to a mechanism that changes which campaign gets scrutinized on a given Tuesday.

4. TECHNOLOGY-WISE IMPACT ANALYSIS

This section runs the framework end to end: exploratory analysis of ethical exposure across campaign configurations, predictive modelling of complaint intensity and material incidents, SHAP-based explanation of the predictions, and conversion of those predictions into a reallocation of assurance effort measured against existing practice.

4.1 Dataset and Approach

The analysis uses 1,920 campaign records, representing roughly the volume a mid-sized advertiser or agency group accumulates over three years. Each record is one campaign, described by its configuration at launch and by what happened afterwards. I have taken the campaign as the unit of analysis, not the impression or the individual, because the campaign is the level at which a marketing organization makes decisions, allocates review effort, and can actually intervene.

Predictors fall into four groups. Campaign and route characteristics cover the channel (eight classes, from paid social and programmatic display through retail media to in-app and push), the product category (eight classes, three of which - financial services and credit, real-money gaming, and health and wellness - are treated as sensitive by most advertising codes), and the operating regulatory regime (four regions). Data and targeting characteristics cover the targeting method (six classes, from contextual through first-party lookalike and third-party segment to fully autonomous optimization), personalization depth, the number of distinct data sources feeding the audience, the share of the audience derived from third-party data, the consent basis (explicit opt-in, legitimate interest, bundled or pre-ticked, partner-inherited), and a consent clarity score capturing the quality of the interface through which consent was obtained. Automation characteristics cover the model autonomy level on a five-point scale, running from human approval of every audience and creative decision to autonomous audience construction, bidding and creative generation, plus the share of creative assets that were machine-generated. Governance characteristics cover assurance review hours allocated before launch, presence of an AI or advertising disclosure, availability of explainability tooling to the team, completion of a data protection impact assessment, the weekly frequency cap, and two indices: audience vulnerability, describing how exposed the target audience was, and inference sensitivity, describing

how far the audience definition depended on inferred instead of declared attributes. Three commercial variables, spend and impressions and click-through rate, are included specifically so their contribution can be measured instead of assumed.

There are two outcomes. The first is realized complaint intensity, measured as verified consumer complaints and privacy escalations per million impressions, treated as a regression target. The second is a binary material-incident flag, defined structurally instead of by category: a campaign is flagged where the excess of realized complaint intensity over the pre-launch checklist expectation exceeded the assurance effort allocated to absorb it. One hour of pre-launch assurance is treated as absorbing about one point of excess complaint intensity, a modelling convention adopted so that exposure and mitigation share units and can be compared directly. The definition matters because it makes the target relative to what the organization planned for. It identifies the campaigns that surprised the organization, not the campaigns that were simply in a difficult category. Material incidents represented 30.7 per cent of records.

A note on the data belongs here, before the results and not in a limitations paragraph at the end. The dataset is an illustration, generated from a specified stochastic process in which I imposed the relationships between configuration and outcome. The models recover those relationships, and it would reflect poorly on the methods if they did not. The purpose is to show that the pipeline composes, that the explanations are legible in a marketer's vocabulary, and that the decision layer converts a prediction into a defensible reallocation. No coefficient, error magnitude or incident rate reported below corresponds to any real advertiser. Section 6 says which findings would and would not be expected to survive contact with real complaint data. Section 9 treats validation on real data as the first priority for subsequent work.

4.2 Exploratory Analysis

Ethical exposure was classified into three bands by the excess of realized complaint intensity over the checklist expectation: within expectation (up to two complaints per million impressions above the checklist figure), elevated (two to ten) and severe escalation (above ten). Exactly half the campaigns fell into the first band, 33.5 per cent into the second and 16.5 per cent into the third.

Across product categories the distribution is anything but uniform.

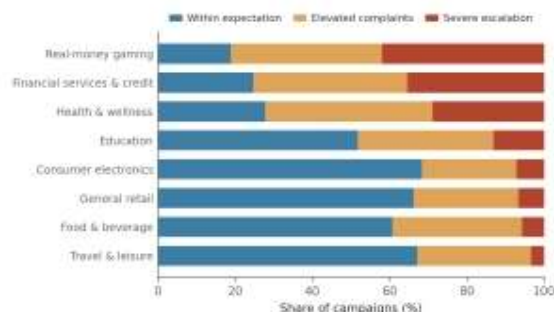


Figure 1 Ethical Exposure Band by Product Category
Real-money gaming is the obvious outlier. Fewer than one campaign in five stayed within the checklist expectation and

over forty per cent produced severe escalation. Financial services and credit, and health and wellness, follow at slightly lower intensity, with within-expectation rates of roughly a quarter and a little over a quarter. At the other end, travel and leisure, general retail and consumer electronics kept between sixty-six and sixty-eight per cent of campaigns within expectation. Measured as material incidents, the spread runs from 61.5 per cent for real-money gaming and 55.1 per cent for financial services and credit down to 12.6 per cent for travel and leisure.

This is exactly the pattern a conventional checklist already captures, which is why it is the least interesting result in the section. Any competent compliance function knows that credit and gambling advertising carries more risk than homeware advertising, and the checklist expectation used as the baseline here already encodes that knowledge. What matters for governance is the part left unexplained once category is accounted for. To see it, the same data has to be split along dimensions the checklist does not record.

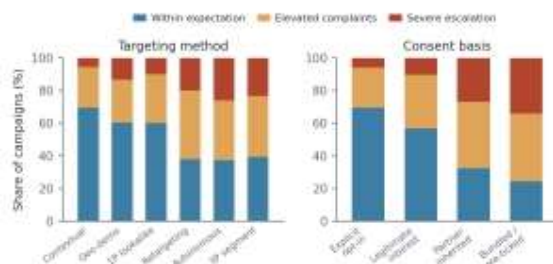


Figure 2 Ethical Exposure by Targeting Method and Consent Basis

Splitting by targeting method and consent basis sharpens the picture considerably. Mean excess complaint intensity for contextually targeted campaigns was negative, at -0.38, meaning those campaigns typically did slightly better than their checklist expectation. First-party lookalike and geo-demographic targeting sat at 1.08 and 1.35. Third-party segment targeting, behavioral retargeting and autonomous optimization clustered at 4.72, 4.52 and 5.11, an order-of-magnitude gap from the contextual case. Material incident rates followed: 14.8 per cent for contextual campaigns against 38.9, 40.8 and 41.1 per cent for the three data-intensive methods.

The consent split is starker, and it is the most useful descriptive result here. Campaigns built on explicit opt-in recorded a mean excess of -0.16 and a material incident rate of 14.2 per cent. Legitimate interest recorded 1.63 and 23.7 per cent. Partner-inherited consent recorded 5.71 and 45.8 per cent. Bundled or pre-ticked consent recorded 6.85 and 53.7 per cent. A campaign resting on bundled consent was therefore around three and a half times more likely to produce a material ethical incident than a comparable campaign resting on explicit opt-in.

That figure is worth sitting with, because it is a marketing argument and not an analytical one. Bundled consent gets chosen because it produces a larger addressable audience at lower acquisition cost, and within the illustration it does exactly that. What the split shows is that the saving is borrowed, not earned. It is repaid later in complaint volume, escalation handling, and review capacity consumed by campaigns that

should never have been configured that way. The regional split says the same thing from another angle: mean excess of 0.68 in the European and United Kingdom regime against 3.93 in the United States, 3.39 in India and 3.45 across the remaining Asia-Pacific markets. In the generating process that gap is driven by the consent practices each regime makes available, not by any property of the consumers.

Correlation analysis across the numerical variables identified the relationships the models would later exploit.

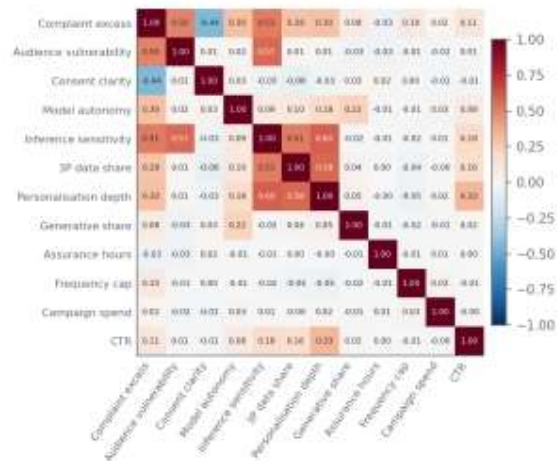


Figure 3 Correlation Matrix of Campaign and Governance Variables

Excess complaint intensity correlated with inference sensitivity at 0.51, audience vulnerability at 0.50, consent clarity at -0.44, personalization depth at 0.30, model autonomy at 0.30 and third-party data share at 0.28. The commercial variables were close to irrelevant: spend at 0.02, click-through rate at 0.11. Assurance review hours showed almost no marginal correlation with excess at all, at -0.03. That null result is worth stating explicitly, because it is the empirical signature of the misallocation this framework exists to correct. Effort spread evenly cannot correlate with need.

So the matrix establishes something fairly clear. Ethical exposure in this illustration is not primarily a property of what is being advertised or how much is being spent. It is a property of how the audience was assembled, on what consent basis, with how much inference, and under how much automation, every one of which the advertiser decides and can revise. The rest of the section depends on that.

4.3 Predictive Modelling

Two model families were trained: a random forest (Breiman, 2001) and a gradient boosting machine using the XGBoost implementation (Chen and Guestrin, 2016). Categorical variables were one-hot encoded, the data was split 75:25 into training and held-out test sets, and no tuning was performed against the test set. Tree ensembles were chosen for the reasons Hastie et al. (2009) give: they handle mixed categorical and continuous predictors without transformation, capture interactions of the kind evidently present here without their being specified in advance, are robust to monotone transformations of the inputs, and admit exact, computationally efficient SHAP attribution. The framework requires that last property rather than merely benefiting from it.

For the regression task, performance was compared against two baselines drawn from actual governance practice. The first is the pre-launch checklist expectation at face value. The second is that expectation corrected by the mean excess observed in training, which represents an experienced compliance manager who knows the checklist runs optimistic and adjusts it by a constant.

Table 1. Complaint-Intensity Prediction Performance Against Two Governance Baselines (480 held-out campaigns)

Estimation method	MAE	RMSE	R ²
Compliance checklist (as-is)	5.77	7.35	0.348
Checklist + mean correction	5.65	6.81	0.440
Random forest	3.54	4.41	0.765
XGBoost	3.18	4.07	0.800

XGBoost achieved a mean absolute error of 3.18 complaints per million impressions against 5.77 for the raw checklist, a reduction of 44.8 per cent, and 43.6 per cent against the mean-corrected checklist at 5.65. The random forest followed at 3.54. Variance explained rose from 0.348 for the raw checklist to 0.800 for XGBoost.

The second comparison is the one that carries weight. Beating a raw checklist is easy and not very interesting, since any experienced compliance manager already applies a mental correction to it. The mean-corrected baseline represents that manager. It improves the raw checklist by about two per cent on mean absolute error; the model improves it by forty-four. What that says is that the error in the as-is process is not a level error a constant can fix. It is a variance error. The checklist is wrong by different amounts for different campaigns, systematically, in ways predictable from configuration variables it does not record.

For the classification task the positive class represented 30.7 per cent of records. The random forest achieved accuracy of 0.846, precision of 0.783, recall of 0.647, F1 of 0.709 and ROC-AUC of 0.904. XGBoost achieved 0.854, 0.767, 0.712, 0.739 and 0.915 respectively.

These numbers are good without being decisive, and I report them as such. A recall of 0.647 means more than a third of the campaigns that went on to cause a material incident were never flagged, a serious limitation for a control whose whole purpose is to catch them before launch. Precision of 0.783 means roughly one flagged campaign in five would have been fine without intervention, consuming review capacity for nothing. Neither figure supports replacing human judgement, and I make no such claim.

It is a fixable limitation, though, and the fix is a governance decision rather than a modelling one. A default threshold of 0.5 assumes a false positive and a false negative cost the same. In ethical governance they plainly do not. A false positive costs a few hours of unnecessary review. A false negative costs a regulatory escalation, a consumer harmed, and a reputational event whose cost is bounded only by how far it travels. Lowering the random forest threshold to 0.36 raised recall from 0.647 to 0.856 while precision fell from 0.783 to 0.630 and

overall accuracy slipped from 0.846 to 0.813. In plain terms, the organization now catches roughly six in seven of the campaigns that would have caused a material incident, and pays for it by reviewing a modest number that did not need it. That is not a better model. It is a decision about what the organization is willing to pay, made explicit and written down, which is what a CSR posture ought to require.

4.4 Explaining the Predictions with SHAP

Accuracy establishes that the model is useful; explanation establishes that it is usable. SHAP values were computed for the gradient boosting classifier using the TreeExplainer algorithm (Lundberg and Lee, 2017) on a sample of 400 held-out campaigns, producing a global ranking of feature influence and a local decomposition for every individual campaign.

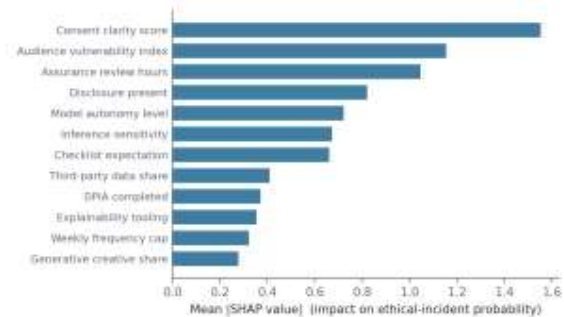


Figure 4 SHAP Global Feature Importance

Five variables dominate the global attribution. Consent clarity carried the largest mean absolute SHAP value at 1.55, followed by audience vulnerability at 1.15, allocated assurance review hours at 1.05, presence of a disclosure at 0.82 and model autonomy at 0.72. Inference sensitivity and the checklist expectation followed at 0.67 and 0.66, then third-party data share at 0.41, completion of a data protection impact assessment at 0.37, explainability tooling at 0.35, weekly frequency cap at 0.32 and generative creative share at 0.28.

Start with what is absent. Spend, impressions and click-through rate together account for 3.7 per cent of total attribution, and none appears in the top twelve. Product category, the variable most real compliance triage is built on, accounts for 2.7 per cent once the checklist expectation is included as a predictor in its own right. Category is not unimportant in absolute terms. It is unimportant conditional on knowing how the campaign was actually configured, which is a stronger and more useful statement.

Then notice that four of the top five are governance choices, not properties of the market, the product or the consumer. Consent clarity is determined by the interface the firm ships. Assurance hours are the firm's own allocation. Disclosure is a policy decision. Autonomy is a configuration setting. Only audience vulnerability describes something about the people being addressed, and even that is partly endogenous, since the firm chose the audience. The consequence is that the model's output is almost entirely actionable. A campaign flagged as high risk is flagged for reasons the team can change before launch. That is the difference between a risk score and a warning.

The prominence of assurance review hours needs care in interpretation. Assurance effort forms part of the incident

definition, since an incident occurs where excess exposure exceeds the effort allocated to absorb it, so its high attribution partly reflects the target definition and not an independent causal discovery. Presenting it as evidence that review effort works would be dishonest. What the variable does contribute legitimately is the mechanism Section 4.5 exploits: allocation is the organization's own lever, the model quantifies its relationship to outcome, and so allocation can be optimized instead of defaulted.

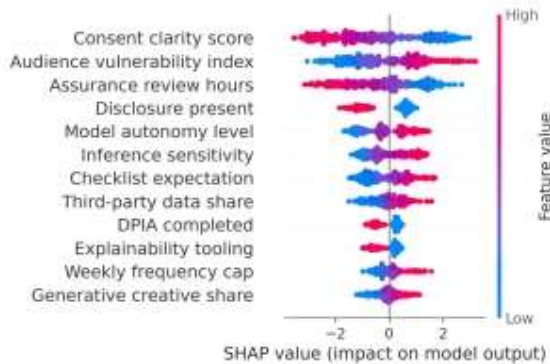


Figure 5 SHAP Summary Distribution Across Test Campaigns The beeswarm in Figure 5 shows direction as well as magnitude. High consent clarity values, in red, push predictions firmly to the low-risk side; low values, in blue, push the other way, with a wide asymmetric tail. Poor consent clarity can move a prediction much further than good consent clarity can move it back. Audience vulnerability shows the expected inverse coloring, and assurance hours a clean inverse relationship. Disclosure and impact-assessment completion appear as tight binary separations with modest but highly consistent protective effects, a useful property for governance, since it means both controls behave predictably across the whole population instead of only in particular segments. Model autonomy shows a marked asymmetry that matches the descriptive statistics. Campaigns at autonomy level one recorded a mean excess of -0.45 and level two 0.59, both essentially benign. Level three recorded 2.53, level four 3.64, level five 6.55. The relationship is convex, not linear, and that has a direct operational implication: the marginal ethical cost of moving from human-approved to partially automated is small, while the marginal cost of moving from partially automated to fully autonomous is large. An organization reading only the linear correlation of 0.30 will miss this and will underprice the last step. The last step is the one that matters. Generative creative share is the clearest case where local explanation earns its keep. Its global attribution is modest, at 0.28, so a global ranking would file it as a minor concern. But the generating process makes its effect conditional on disclosure: machine-generated creative carries essentially no penalty where a disclosure is present and a substantial one where it is absent. Average that across a population in which roughly forty per cent of campaigns carry disclosures and you get a small global effect. For an individual campaign running heavily generated creative with no disclosure, the local attribution is large and the fix is obvious and cheap. A process

built only on global importance would deprioritize the single best cost-benefit intervention available to it.

4.5 From Prediction to Governance Decision

A prediction that changes no decision is worth nothing, so the last step converts model output into a governance action and measures it against existing practice. The action is reallocation of assurance review effort. I chose it because it is the lever most organizations already have, it needs no additional budget, and it can be implemented without renegotiating anything with a platform or a regulator.

The as-is policy was a uniform allocation of assurance hours to every campaign, set at the dataset mean of 6.48. This represents common practice, where review capacity is distributed by campaign count instead of by risk. It is a deliberately fair comparison and not a straw man: total capacity is held identical across both scenarios.

The predictive policy redistributed exactly the same total across exactly the same campaigns, allocating between two and eighteen hours in proportion to predicted incident probability and rescaling so the totals matched to the hour. Across the 480 held-out campaigns, 3,108 assurance hours were allocated under both policies.

Table 2. Effect of Reallocating an Unchanged Assurance Budget by Predicted Risk

Measure	Uniform allocation	Risk-weighted allocation	Change
Total assurance hours allocated	3,108	3,108	0%
Material incident rate	28.8%	21.7%	-24.6%
Unmitigated exposure (complaint-points)	706.8	350.8	-50.4%
Additional governance headcount	-	None	-

The material incident rate fell from 28.8 per cent to 21.7 per cent, a relative reduction of 24.6 per cent. Unmitigated exposure, the summed excess complaint intensity not absorbed by allocated assurance effort, fell from 706.8 to 350.8 complaint-points, down 50.4 per cent. No additional review capacity was purchased, no campaign was cancelled, and no targeting capability was withdrawn.

The gap between those two figures is the most interesting result in the section, and I did not anticipate it. Risk-weighting reduced the depth of the worst exposures roughly twice as effectively as it reduced their frequency. The explanation is structural. Risk-weighting moves hours from campaigns with small excesses to campaigns with large ones, so a campaign that was going to overshoot its allocation by twenty points and now overshoots by four still counts as an incident under a binary definition, while contributing sixteen fewer points of

exposure. Incident counts improve modestly. The severity distribution improves a great deal.

Whether that is the right trade depends on how the costs of ethical failure behave, and a firm has to answer that for itself. Where harm scales with the number of consumers affected and the duration of exposure, a discriminatory delivery pattern running six weeks instead of one, a manipulative flow encountered by two hundred thousand users instead of twenty thousand, the exposure measure is the right one and the result is strong. Where the dominant cost is fixed, because a single regulatory notice triggers the same investigation regardless of scale, the incident rate is the right measure and a 24.6 per cent improvement, while real, is more modest. Most organizations face a mixture, weighted differently by category. The honest recommendation is to choose the objective explicitly and state it, instead of letting it default to whichever metric the tooling happens to report.

5. CUMULATIVE AND INTEGRATED TECHNOLOGY IMPACTS

Section 4 presented the components in sequence. Their value is not additive. Prediction without explanation produces recommendations a marketing team will override, because a pause instruction backed by an unexplained score cannot survive a conversation with a commercial stakeholder who can quantify what the pause costs. Explanation without a decision layer produces insight that changes nothing: knowing consent clarity is the dominant driver does not tell you which of six hundred live campaigns to review this week. And a decision layer without prediction is the uniform allocation the framework set out to replace. A partial implementation will underperform its business case by more than the fraction it omits.

The interdependence works like this. The predictive layer turns a category-level judgement into a campaign-level probability distribution, which is the only form in which the question can be posed quantitatively at all. The explanatory layer decomposes that probability into named contributions: so much from the consent basis, so much from inference depth, so much from the autonomy setting, so much from a missing disclosure. The decision layer turns the decomposition into an allocation and, more usefully, into a set of specific pre-launch changes, each attached to the contribution it addresses.

That decomposition does two things at once. It makes the recommendation actionable, since every contributing factor implies an intervention the team can price and execute before launch - replace the consent basis for this segment, drop the two inferred attributes acting as health proxies, lower the autonomy level from five to three for this flight, add a disclosure to the generated assets. And it makes the recommendation contestable, which determines whether the system survives its first serious disagreement. A marketer who thinks the model is wrong can point at the specific attribution they dispute instead of rejecting an opaque score. If they turn out to be right, the disagreement produces a labelled example that improves the next model. A control nobody can argue with does not get followed. It gets circumvented.

The loop closes when realized outcomes flow back into training. Complaint volumes, opt-out spikes, platform enforcement notices, subject access requests and moderation escalations are all recorded somewhere in most organizations already, usually in systems that do not speak to the campaign record. Connecting them is unglamorous integration work, and it is the point at which the framework stops being a project and becomes a capability. Each quarter adds several hundred labelled examples. The model improves. More importantly for the resource-based argument in Section 2, the labelled record becomes an asset a competitor cannot acquire by buying the same software, because it encodes this firm's audiences, this firm's consent interfaces, this firm's consumers.

Table 3. Marketing Ethics Governance Before and After Integration of Prediction, Explanation and Risk-Weighted Assurance

Dimension	Conventional practice	Integrated framework	Outcome
Exposure estimate	Category checklist (MAE 5.77)	Campaign-level model (MAE 3.18)	45% error reduction
Incident visibility	Judgement at creative sign-off	Campaign-level probability, AUC 0.904	Triage becomes evidence based
Assurance allocation	Uniform, 6.5 hours per campaign	Risk-weighted, 2-18 hours	50.4% less unmitigated exposure
Basis of the decision	Undocumented professional judgement	SHAP attribution per campaign	Auditable and contestable
Learning across campaigns	None; each campaign restarts	Outcomes retrain the model	A trust asset that compounds
Locus of responsibility	Legal review at the end	Configuration choices at the start	Prevention rather than defence

Two rows in Table 3 are worth drawing out. The improvement in exposure estimation is substantial but instrumental; it matters because it enables the allocation, not because a more accurate complaint forecast is intrinsically valuable to anyone. The change in the locus of responsibility carries the most organizational significance and the least quantitative support in this chapter. Conventional practice concentrates ethical review at the end, where the only available remedies are approval, rejection and delay - all expensive, all adversarial. The framework moves the operative decisions to the point where the audience and the automation are configured, where remedies are cheap and mostly technical. That is a claim about workflow, not about modelling, and my illustration supports it only indirectly. It is probably still the most consequential change the framework implies.

At the organizational level the system creates value for three groups with genuinely different interests, which is part of why adoption is hard and part of why it is possible at all. For the

performance marketing team, it removes the blanket restrictions compliance functions impose when they cannot discriminate between campaigns, and substitutes targeted constraints on the configurations that actually carry risk; a team that can demonstrate clean consent and moderate inference depth should meet less friction, not the same friction. For legal and privacy, it converts a defensive posture based on categorical rules into a documented process that answers a regulator's question about how decisions were reached with something better than an assertion of diligence. For the executive, it converts an unquantified reputational risk into a managed exposure with a stated tolerance, which is the only form in which risk can be discussed at board level.

There is also a real tension here, and it should be named instead of resolved rhetorically. The performance team's interest is served only while the constraints stay targeted. The moment the model flags a campaign that is performing well, the alignment breaks and the system falls back on governance authority. Any implementation that assumes permanent alignment will fail at precisely the moment it becomes valuable. Section 7 draws the design implication: build the override path and the escalation route before the first disagreement, not during it.

The strategic frameworks from Section 2 provide the outer check. PESTEL confirms that the legal environment is converging on requirements this framework generates as by-products - records of the basis for automated decisions, documented impact assessment, transparency about synthetic content, demonstrable diligence around vulnerable audiences. Porter confirms that the leverage is asymmetric: an advertiser controls its own configuration choices and cannot inspect the platform's delivery algorithm, which puts a hard ceiling on what any advertiser-side framework can achieve and makes collective action a complement, not an alternative. The resource-based view confirms that the durable asset is the labelled outcome record and not the modelling. And complex adaptive systems thinking supplies the warning that all of this happens in an environment that responds to being measured.

6. CHALLENGES AND RISKS IN IMPLEMENTATION

Data quality is the first and largest obstacle. It is less a technical problem than a records problem compounded by a measurement problem. A model of the kind described in Section 4 needs, for every historical campaign, an accurate record of the configuration at launch and an accurate record of what happened afterwards. Most organizations have neither in usable form. Configuration is scattered across the ad platform, the customer data platform, the consent platform and a set of briefing documents, and what was deployed frequently diverges from what was approved. Audience definitions get edited mid-flight without versioning. Autonomy settings are platform defaults nobody recorded, because nobody chose them.

The outcome side is worse, and the difficulty is structural. Complaints are a filtered signal. They come from consumers who noticed the problem, cared enough to act, knew where to

complain, and had the literacy and time to do it. The label is therefore biased against exactly the vulnerable populations the framework is supposed to protect. A campaign that exploited a financially distressed audience with great efficiency may generate fewer complaints than a mistargeted advertisement that irritated a well-resourced one. Better modelling does not fix this, and it is the single most serious objection to the whole approach. There are partial mitigations - supplement complaint data with proactive panel research, treat platform enforcement actions and regulatory queries as additional labels, weight complaints by segment vulnerability, audit a random sample of campaigns instead of only flagged ones - and they reduce the bias without removing it. Any organisation deploying this should state the residual bias out loud, because the model's apparent objectivity will otherwise conceal it.

The second technical risk is distributional drift, which in this domain moves fast. A model trained during a period of stable identifier availability will mislead once identifiers change. One trained before a platform revised its delivery algorithm will mislead afterwards. One trained before a regulation took effect will mislead once consent interfaces are redesigned in response. The mitigations are conventional and belong in the operating cost rather than the enhancement backlog: monitor prediction error by month, retrain on a rolling window, hold out the most recent quarter as a drift check, and set an explicit trigger at which the model is withdrawn from decision use pending investigation instead of being allowed to degrade quietly.

A third risk follows from the complex adaptive systems argument, and it is the one most likely to hollow the framework out from inside. Teams respond to measurement. A team that learns consent clarity carries the largest attribution will improve its consent clarity score. The question is whether they do that by redesigning the consent interface, which is the intended behavior and a genuine improvement, or by reclassifying existing consents into a more favorable category, which improves the score and changes nothing for a single consumer. The second path is cheaper, faster, and hard to detect from the data the model sees. Any measure that becomes a target degrades as a measure, and this measure is being deliberately made into a target. The only reliable defense is periodic independent verification of the inputs themselves - actually inspecting the consent flow, not the field that describes it. It is expensive, and it is invariably the first thing cut when the programme comes under budget pressure.

Organizational resistance deserves to be taken seriously, because a good deal of it is well founded. Experienced marketers hold knowledge that is not in the data: that a particular audience has historically been sensitive, that a category of creative reads differently in one market than another, that a client's brand carries a specific vulnerability. A system that overrides this because it cannot see it will make worse decisions and will be resented, correctly. There is also a legitimate professional objection to the quantification itself. Reducing the question of whether it is right to target someone on inferred financial distress to a probability score can be read as a category error. The reply that the score allocates review effort and does not settle the ethical question is accurate, and it does not fully dispose of the objection, because scores acquire authority beyond their stated purpose simply by existing.

A further concern is rarely stated openly and strongly affects adoption. A system that records every configuration decision and its ethical outcome is also, whatever its stated purpose, a performance measurement system for the people who made those decisions. Teams work this out quickly. Introduce the framework without an explicit and credible commitment about how the data will and will not be used in performance evaluation, and the rational response is to degrade the input data. The framework then fails for reasons that will be misdiagnosed as technical.

Rudin's (2019) objection has to be confronted rather than noted. Her argument is that an explanation of a black box approximates the model's behaviour and not the underlying reality, and that in high-stakes decisions this breeds unwarranted confidence in a flawed system. My position is narrower than a rejection. Because the decision being supported is where to send a human reviewer, an approximate account of why a campaign was flagged is adequate; the human review that follows is the actual control. That defence holds only while the human review genuinely happens. An organisation that starts by using the score to allocate review and drifts into using it to approve campaigns without review has moved into exactly the territory where Rudin's objection is decisive. Under cost pressure, that drift is the natural direction of travel.

The skills gap is a related constraint, and it is usually misidentified. The scarce role is not the data scientist. It is the person who understands marketing operations well enough to know that a particular consent basis is an artefact of a legacy partner integration and not a deliberate choice, and who understands the modelling well enough to know why that matters for how the variable should be treated. Without that person the model will encode operational accidents as though they were decisions, and the recommendations will be technically correct and practically useless.

Legal exposure adds a layer that points in an uncomfortable direction. A documented, quantified prediction that a campaign carried elevated risk of ethical harm is evidence. If the campaign ran anyway and the harm occurred, the organisation has created a contemporaneous record of its own foreknowledge, which in several jurisdictions turns a question about negligence into a question about knowing conduct. That is a real disincentive to measurement and pretending otherwise would be dishonest. My considered response is that the alternative, deliberately not knowing, is indefensible under any CSR framing worth the name and increasingly untenable legally, since the impact assessment obligations in the General Data Protection Regulation (2016) and the risk management obligations in the EU Artificial Intelligence Act (2024) both presume a diligent operator would have assessed the risk. The exposure is nonetheless real. Discuss it with counsel before deployment, and set the retention policy for risk scores deliberately.

The limitations of the illustration itself need restating plainly. The dataset is synthetic. It came from a specified process, and the models recover relationships I put there by construction. The error magnitudes, the incident rate, the attribution rankings and the 50.4 per cent reduction in unmitigated exposure are properties of that construction and should not be quoted as findings about advertising. Real campaign data contains

missing values, mislabelled categories, structural breaks at platform policy changes, and idiosyncratic brand and market effects that no generating process reproduces. The clean convexity in the autonomy relationship, the tidy separation between consent bases and the near-zero contribution of the commercial variables would all be noisier in practice. The last of those might not survive at all, since in real data spend is entangled with reach, prominence, and the probability that anyone notices a campaign enough to complain about it.

What the illustration does support is the comparative and structural claims. Campaign-level prediction beats both a category checklist and a mean-corrected checklist by a wide margin whenever configuration variables carry information category does not, which is an assertion about where information lives and not about a particular dataset. Reallocating a fixed assurance budget by predicted risk reduces exposure more than it reduces incident counts, which follows from the arithmetic of moving effort toward the tail and holds for any risk distribution with a long tail. Explanations expressed in configuration variables suggest interventions in a way explanations expressed in category or spend do not. Those are the claims the recommendations rest on.

Finally, context matters more than the framework literature usually admits. A large advertiser or agency group running several hundred campaigns a year through a consolidated stack, with a privacy function and a data team already in place, can implement all of this. A firm running forty campaigns a year through an external agency, with campaign records living in the agency's systems and complaint handling spread across a shared inbox, cannot, and should not be told otherwise. For that firm the relevant recommendations are the low-technology ones in Section 7. Record the consent basis and the autonomy setting for every campaign in one consistent place. Keep a clean log of complaints and escalations with the campaign identifier attached. Apply a simple rules-based triage that concentrates review on autonomous campaigns targeting vulnerable audiences on weak consent. That is perhaps a fifth of the benefit for a twentieth of the cost, and for most organizations it is the honest recommendation.

7. BEST PRACTICES AND STRATEGIC RECOMMENDATIONS

Start with the record, not the model. This is the least exciting recommendation in the chapter and it determines whether the rest is achievable at all. Establish a single campaign register recording the consent basis, targeting method, autonomy level, inferred-attribute list, disclosure status and allocated review effort for every campaign, and link it by identifier to complaints, opt-outs, escalations and enforcement notices. With two clean years of that, everything in Section 4 can be built in a matter of weeks. Without it, none of it can be built at any level of analytical capability, and a good deal of money will be spent discovering this.

Define the target relative to expectation. A model that predicts which campaigns generate the most complaints will predict the sensitive categories, which the organization already knows. A model that predicts which campaigns exceed what was

expected of them identifies the campaigns that will surprise the organization, and absorbing surprise is what governance capacity is for. This one definitional choice separates a model that restates the compliance manual from a model that adds to it.

Set the decision threshold from harm cost, not from convention. Section 4.3 showed a move from 0.5 to 0.36 raising recall from 0.647 to 0.856, paid for in unnecessary reviews. The default threshold encodes an assumption of symmetric costs that nobody in an ethics function actually holds. Make the asymmetry explicit, choose deliberately, record the reasoning, and revisit it whenever the harm cost changes, which it does the moment the firm enters a regulated category or a stricter regime.

Deploy explanations to the campaign manager, not the analyst. SHAP output belongs in the campaign setup interface, next to the audience definition, in marketing language: this audience rests on bundled consent for sixty per cent of its records, three of its inferred attributes are health-adjacent, autonomy is at maximum, and the generated creative carries no disclosure. Put that way, the model is a colleague raising four specific and fixable objections before launch. Put as a risk score of 0.71, it is an obstacle, and it will be treated as one.

Fix the configuration defaults the model exposes. Four of the five highest-attribution variables in Section 4.4 are settings, not circumstances. Where consent clarity dominates the ranking, redesigning the consent interface returns more than any amount of downstream review, because it removes the exposure instead of absorbing it. The general principle: a governance model earns most of its value by identifying which upstream defaults generate the downstream work. An organization that responds to the model only by reviewing more campaigns has captured the smaller half of the benefit.

Treat autonomy as a decision requiring justification, not a performance setting. The convexity in Section 4.4 means the last increment of automation carries a disproportionate share of the ethical cost. A defensible policy makes the highest autonomy levels available by default only in low-vulnerability categories on first-party consented data, and requires explicit sign-off elsewhere. It is cheap, it is easy to audit, and it maps onto the risk-tiering logic the EU Artificial Intelligence Act (2024) applies to AI systems generally.

Design the override path before deployment. Marketers will disagree with the model and will sometimes be right, particularly where they hold context the data does not contain. The system should permit an override, require a recorded reason, route it to a named accountable person above a defined risk level, and feed the outcome back into training. Overrides are not a failure of the system. They are how it learns the boundaries of its own competence, and an implementation with no overrides is one where people have quietly stopped engaging with it.

Do not communicate the programme until it has been running for a year. Du et al. (2010) applies with full force here: announcing a responsible AI marketing framework before it has changed a single campaign invites exactly the scepticism the announcement was meant to avoid, and turns an operational programme into a reputational liability if it later underperforms.

Build it, run it, accumulate the record of what it changed, and let the specifics carry the communication when it comes.

Sequence the rollout by configuration risk, not by category. The first deployment should cover campaigns combining high autonomy, weak consent and vulnerable audiences. In the illustration those concentrated in real-money gaming, financial services and health, but they appeared in every category. Sequencing by category alone reproduces the blind spot the framework exists to remove, leaving the badly configured retail campaign unreviewed while adding a fourth review layer to a well-configured credit campaign.

Assign ownership for the operating phase and fund the verification. The commonest failure mode is an implementation that succeeds and then decays, because nobody owns retraining, drift monitoring, or the independent verification of inputs that Section 6 identified as the only real defense against gaming. That ownership belongs with a named person in privacy or marketing operations, with a defined review cadence and a budget line. Not with the project team that built it, and not with a data science group that has moved on.

Scale the ambition to the organization. For a large advertiser or agency group the full framework is achievable within a year. For a mid-sized firm, a defensible first version is a consistent campaign register, a simple logistic model on six configuration variables, and a two-tier review triage, which captures a meaningful share of the benefit. For a small firm, record the consent basis and autonomy setting consistently, keep a clean complaint log, and apply the rules-based triage from the end of Section 6. The framework should never be presented as all-or-nothing, because presented that way most organizations will correctly conclude it is nothing.

8. CASE STUDIES FROM INDUSTRY

8.1 Meta - Algorithmic Delivery Skew and the Variance Reduction System

Meta's advertising system is the most fully documented industrial example of the delivery-skew problem Ali et al. (2019) identified academically. In June 2022 the United States Department of Justice announced a settlement resolving allegations that Meta's housing advertisement delivery system violated the Fair Housing Act. Meta agreed to discontinue its Special Ad Audience tool and to build a new system addressing disparities between the audience an advertiser selected and the audience that actually received the advertisement. The resulting Variance Reduction System was deployed for housing advertisements in the United States and later extended to employment and credit.

Three features of the case bear on the framework proposed here. The harm arose from optimization and not from targeting, so it was invisible to the advertiser and could not have been caught by any advertiser-side creative or targeting review, however diligent. The remedy operates on delivery outcomes and not on inputs, measuring variance between eligible and reached audiences and adjusting delivery to reduce it - an outcome-monitoring control of exactly the type Section 5 argues advertisers cannot build for themselves, because they cannot see inside the auction. And the change was produced by

litigation, not by the internal ethics processes that existed throughout the preceding period. Whatever else it shows, the case shows the limits of voluntary self-governance at the platform layer, which supports the institutional reading in Section 2 that coercive pressure is the reliable mechanism.

8.2 Google - Privacy Sandbox and the Fragility of Voluntary Commitment

Google's Privacy Sandbox, announced in 2019, proposed replacing cross-site tracking in Chrome with a set of privacy-preserving APIs, including the Topics API that succeeded the abandoned Federated Learning of Cohorts proposal after criticism of its fingerprinting and sensitive-inference properties. Alongside it, Google's personalized advertising policies restrict targeting on sensitive categories such as health conditions and financial hardship, and a limited-ads mode serves non-personalized advertising where consent is absent. Taken together, it was the largest voluntary re-architecture of advertising infrastructure ever attempted.

What makes it valuable as a case is its trajectory. The deprecation timeline slipped repeatedly and the commitment was ultimately revised: instead of removing third-party cookies from Chrome, Google moved first to a user-choice mechanism and then, as reported at the time of writing, to retaining third-party cookies while continuing to develop the Sandbox APIs. The stated reasons involved advertiser and publisher readiness and regulatory scrutiny of the competitive effects. Both are legitimate. A unilateral change to shared infrastructure genuinely does raise competition questions, and the same regulators worried about privacy were worried about market power. The lesson for a CSR framework is not that the commitment was insincere. It is that a voluntary commitment imposing concentrated costs on a firm's own revenue base, and on the ecosystem it depends on, is structurally fragile however genuine the intent. An advertiser building a responsible marketing programmer should therefore not treat platform-level privacy commitments as a stable planning assumption, and should keep the ability to operate on first-party consented data whatever any platform eventually does.

8.3 Unilever - Responsible Marketing Data as a Brand Position

Unilever offers the clearest advertiser-side example. The company published a Responsibility Framework for Marketing Data setting out commitments on transparency about data collection, meaningful consumer control, and restraint in the use of sensitive inferences, positioning responsible data use as an extension of existing brand equity instead of as a compliance function. It has also acted at the media-buying level, notably in its 2020 decision to suspend paid advertising on major social platforms in the United States over concerns about polarizing and hateful content, and at the content level through the Unstereotype Alliance convened with UN Women, which addresses representational harm in creative rather than data harm.

The case is instructive in three ways, and should be read with caution, since the available information is largely self-reported and the effects on consumer behavior are not independently established. It demonstrates the buyer-power mechanism from the Porter analysis: a very large advertiser withdrawing spend produced platform responses that individual consumer

complaints had not. It illustrates the congruence condition from Sen and Bhattacharya (2001), since data responsibility is closely matched to a business whose core asset is consumer trust in packaged goods brands, which is the condition under which CSR initiatives help instead of inviting cynicism. And, most relevantly here, the published framework operates at the level of principles and not of campaign-level mechanism. That is exactly the gap Section 3 identified. It states what the organization believes without specifying which campaign gets reviewed on a given Tuesday. The framework in Section 4 is an attempt to supply that missing layer, not an alternative to commitments of this kind.

Read together, the three cases describe an ecosystem where the components discussed in this chapter exist but are distributed across parties with divergent incentives. The platform holds the delivery data and changes behavior mainly under legal compulsion. The advertiser holds the configuration decisions and the brand incentive but cannot see the delivery mechanism. The regulator supplies the power the individual consumer lacks, but arrives late and by category. No party holds the whole problem. That is the structural reason an advertiser-side framework is necessary, and equally the reason it is insufficient on its own.

9. FUTURE OUTLOOK AND RESEARCH IMPLICATIONS

The most urgent requirement is validation on real campaign data. Everything in Section 4 rests on an illustrative dataset, and the honest position is that the framework has been shown to compose, not shown to work. The study that would settle it is a retrospective analysis inside a single advertiser or agency group holding two or more years of linked campaign configuration and complaint records, testing whether configuration variables predict realized complaint intensity beyond what category and spend already explain. That study needs only data several organizations already hold. The obstacle is commercial sensitivity, not method.

A second direction concerns the label bias from Section 6, which is the most serious threat to the validity of the whole approach and is a research problem, not an engineering one. Complaints under-represent harms suffered by the least resourced consumers, so a model trained on complaints will systematically under-protect them, inverting the framework's stated purpose. Work is needed on composite outcome measures combining complaint data with proactively solicited panel evidence, regulatory and platform enforcement signals, and audit findings from randomly sampled campaigns, and on characterizing how much bias survives. Until there is progress here, any deployment should be regarded as protecting the firm more reliably than it protects the consumer.

Third, the generative dimension needs empirical attention it has not had. The volume of machine-generated creative already exceeds what human review can inspect, invalidating the assumption underlying every creative approval process now operating. Research is needed on automated pre-screening of generated assets for representational harm and misleading claims, on what constitutes effective disclosure of synthetic content in short formats, and on how consumers actually

respond to disclosed versus undisclosed synthetic endorsement. The regulatory position is moving faster than the evidence base, and marketing practice faster still.

Fourth, the framework as described stops at allocating review effort. The natural extension is joint optimization across configuration choices, treating consent basis, inference depth, autonomy level and disclosure as decision variables with both commercial and ethical consequences, and identifying configurations that are Pareto-efficient instead of merely compliant. It is a much harder problem, since it needs a credible model of the commercial cost of each ethical improvement. It is also the version that would let a firm answer the question its executives actually ask, which is what responsibility costs.

Fifth, the behavioral questions raised in Sections 5 and 6 need study. How do marketers actually respond to recommendations that conflict with performance targets, and does exposing the attribution improve compliance with the recommendation, as the trust literature suggests it should? Does making a variable salient in a governance score cause teams to improve the underlying practice or the recorded value? These are experimental questions, answerable inside a single organization, and their answers determine whether frameworks of this kind work as designed or become elaborate documentation exercises.

Finally, the distributive dimension deserves attention. If predictive governance capability accrues only to the largest advertisers, the effect is to raise the effective cost of responsible marketing and concentrate it among firms that can afford the apparatus, while smaller firms carry on with category checklists. Shared industry infrastructure would reduce the asymmetry: pooled and anonymized outcome data held by a trade body, common configuration taxonomies, open reference implementations of the triage logic. Whether firms can be brought to pool data about their own ethical failures is doubtful on current evidence. The analogous problem in aviation safety reporting was solved by separating the learning system from the enforcement system, and that precedent deserves serious examination.

10. CONCLUSION

This chapter set out to examine the ethical problems artificial intelligence has created in digital marketing, and to be specific about how a firm might address them instead of restating the problems at greater length. The mechanism proposed has three parts: predict ethical exposure at the level of the individual campaign from its actual configuration, explain each prediction in terms the campaign team can act on, and use the explained prediction to reallocate assurance effort the organization already spends.

In the illustration, campaign-level prediction of complaint intensity reduced mean absolute error to 3.18 complaints per million impressions against 5.77 for a compliance-checklist baseline and 5.65 for a mean-corrected version of it. Material ethical incidents were identified with an ROC-AUC of 0.904. Lowering the decision threshold from 0.5 to 0.36 raised recall from 0.647 to 0.856. Reallocating an unchanged budget of assurance hours by predicted risk reduced unmitigated

exposure by 50.4 per cent and material incidents by 24.6 per cent. The dataset was synthetic, and those figures are properties of its construction, not findings about advertising.

Three conclusions do not depend on the magnitudes. The first is that the components have to be assembled, not adopted individually. Prediction alone produces recommendations that get overridden, explanation alone produces insight that changes no decision, and reallocation without prediction is the uniform policy the framework replaces. Partial implementations do not deliver proportionate benefit.

The second is that most of the available benefit sits upstream of where governance currently operates. Four of the five highest-attribution variables were configuration choices the firm makes: the consent interface it ships, the review effort it allocates, the disclosure policy it sets, the autonomy level it permits. Product category contributed 2.7 per cent of attribution and the commercial variables 3.7 per cent. On this evidence, ethical exposure in AI-based marketing is substantially a function of decisions taken when the campaign is configured, not of what is being sold or how much is being spent. That relocates the governance problem from the approval gate at the end to the defaults at the beginning, and it means a firm can reduce exposure considerably without withdrawing a single product or capability.

The third is that the CSR framing does real work here instead of decorating a compliance exercise. A compliance framing asks whether the firm can defend what it did, and documentation satisfies it. A CSR framing asks whether the firm should have done it, and only a changed decision satisfies that. The two diverge precisely where the law has not yet arrived, which in AI-based marketing is most of the interesting territory: inferring sensitive attributes from behavioral proxies, personalizing persuasion to individually discovered vulnerability, generating creative faster than any human process can review it. In this domain Carroll's ethical layer is not a residual category. It is where the substantive decisions get made.

The limitations are equally clear and I have stated them where they arise. The dataset is illustrative. The complaint-based label is biased against the consumers the framework most needs to protect, and no amount of modelling repairs that. The measurement is gameable by the teams it measures, and the only real defense is expensive verification that budget pressure removes first. The post-hoc explanations are approximations, adequate for allocating human attention and inadequate for anything the organization might later be tempted to use them for. And the whole apparatus operates on the advertiser's own configuration choices, which leaves the platform-side delivery mechanisms in Section 8 entirely beyond its reach.

What can be said with more confidence is that the problem is worth the effort and the current alternative is weak. Marketing has spent three decades building an apparatus of unprecedented inferential precision and has governed it with a checklist that reasons at the level of the product category. The gap between the resolution of the machinery and the resolution of its governance is where most of the documented harms have occurred. Closing it does not require new science. The models are public, the explanation methods are mature, and the data mostly sits inside the organizations concerned already. What it

requires is deciding that the question of who is being reached, on what basis, and by a system running at what level of autonomy, is a question the firm is answerable for. Which is, in the end, what corporate social responsibility means once it is applied to a business built on inference.

Reference List

- Acquisti, A., Brandimarte, L. and Loewenstein, G. (2015) Privacy and human behavior in the age of information. *Science*, 347(6221), pp.509-514. Available at: <https://doi.org/10.1126/science.aaa1465>
- Aguirre, E., Mahr, D., Grewal, D., de Ruyter, K. and Wetzels, M. (2015) Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness. *Journal of Retailing*, 91(1), pp.34-49. Available at: <https://doi.org/10.1016/j.jretai.2014.09.005>
- Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A. and Rieke, A. (2019) Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp.1-30. Available at: <https://doi.org/10.1145/3359301>
- Barney, J. (1991) Firm resources and sustained competitive advantage. *Journal of Management*, 17(1), pp.99-120. Available at: <https://doi.org/10.1177/014920639101700108>
- Barocas, S. and Selbst, A.D. (2016) Big data's disparate impact. *California Law Review*, 104(3), pp.671-732.
- Bleier, A. and Eisenbeiss, M. (2015) Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, 34(5), pp.669-688. Available at: <https://doi.org/10.1287/mksc.2015.0930>
- Breiman, L. (2001) Random forests. *Machine Learning*, 45(1), pp.5-32. Available at: <https://doi.org/10.1023/A:1010933404324>
- Calo, R. (2014) Digital market manipulation. *George Washington Law Review*, 82(4), pp.995-1051.
- Carroll, A.B. (1979) A three-dimensional conceptual model of corporate performance. *Academy of Management Review*, 4(4), pp.497-505.
- Carroll, A.B. (1991) The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business Horizons*, 34(4), pp.39-48. Available at: [https://doi.org/10.1016/0007-6813\(91\)90005-G](https://doi.org/10.1016/0007-6813(91)90005-G)
- Central Consumer Protection Authority (2023) Guidelines for Prevention and Regulation of Dark Patterns, 2023. New Delhi: Ministry of Consumer Affairs, Government of India.
- Chen, T. and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785-794. Available at: <https://doi.org/10.1145/2939672.2939785>
- Datta, A., Tschantz, M.C. and Datta, A. (2015) Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015(1), pp.92-112. Available at: <https://doi.org/10.1515/popets-2015-0007>
- Davis, F.D. (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), pp.319-340. Available at: <https://doi.org/10.2307/249008>
- Dietvorst, B.J., Simmons, J.P. and Massey, C. (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), pp.114-126. Available at: <https://doi.org/10.1037/xge0000033>
- DiMaggio, P.J. and Powell, W.W. (1983) The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48(2), pp.147-160. Available at: <https://doi.org/10.2307/2095101>
- Donaldson, T. and Preston, L.E. (1995) The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review*, 20(1), pp.65-91.
- Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Du, S., Bhattacharya, C.B. and Sen, S. (2010) Maximizing business returns to corporate social responsibility: The role of CSR communication. *International Journal of Management Reviews*, 12(1), pp.8-19. Available at: <https://doi.org/10.1111/j.1468-2370.2009.00276.x>
- European Union (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). *Official Journal of the European Union*, L119.
- European Union (2022) Regulation (EU) 2022/2065 of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act). *Official Journal of the European Union*, L277.
- European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
- Federal Trade Commission (2023) Guides Concerning the Use of Endorsements and Testimonials in Advertising. 16 CFR Part 255. Washington, DC: Federal Trade Commission.
- Freeman, R.E. (1984) *Strategic Management: A Stakeholder Approach*. Boston: Pitman.

- Gray, C.M., Kou, Y., Battles, B., Hoggatt, J. and Toombs, A.L. (2018) The dark (patterns) side of UX design. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp.1-14. Available at: <https://doi.org/10.1145/3173574.3174108>
- Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edn. New York: Springer.
- Kaminski, M.E. (2019) The right to explanation, explained. Berkeley Technology Law Journal, 34(1), pp.189-218.
- Lambrecht, A. and Tucker, C. (2019) Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Management Science, 65(7), pp.2966-2981. Available at: <https://doi.org/10.1287/mnsc.2018.3093>
- Logg, J.M., Minson, J.A. and Moore, D.A. (2019) Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, pp.90-103. Available at: <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Luguri, J. and Strahilevitz, L.J. (2021) Shining a light on dark patterns. Journal of Legal Analysis, 13(1), pp.43-109. Available at: <https://doi.org/10.1093/jla/laaa006>
- Lundberg, S.M. and Lee, S.-I. (2017) A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, pp.4765-4774.
- Luo, X. and Bhattacharya, C.B. (2006) Corporate social responsibility, customer satisfaction, and market value. Journal of Marketing, 70(4), pp.1-18. Available at: <https://doi.org/10.1509/jmkg.70.4.001>
- Martin, K.D., Borah, A. and Palmatier, R.W. (2017) Data privacy: Effects on customer and firm performance. Journal of Marketing, 81(1), pp.36-58. Available at: <https://doi.org/10.1509/jm.15.0497>
- Martin, K.D. and Murphy, P.E. (2017) The role of data privacy in marketing. Journal of the Academy of Marketing Science, 45(2), pp.135-155. Available at: <https://doi.org/10.1007/s11747-016-0495-4>
- Mathur, A., Acar, G., Friedman, M.J., Lucherini, E., Mayer, J., Chetty, M. and 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), pp.1-32. Available at: <https://doi.org/10.1145/3359183>
- Metaxa, D., Park, J.S., Robertson, R.E., Karahalios, K., Wilson, C., Hancock, J. and Sandvig, C. (2021) Auditing algorithms: Understanding algorithmic systems from the outside in. Foundations and Trends in Human-Computer Interaction, 14(4), pp.272-344. Available at: <https://doi.org/10.1561/11000000083>
- Ministry of Electronics and Information Technology (2023) The Digital Personal Data Protection Act, 2023. Act No. 22 of 2023. Government of India.
- Mitchell, R.K., Agle, B.R. and Wood, D.J. (1997) Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. Academy of Management Review, 22(4), pp.853-886.
- Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016) The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), pp.1-21. Available at: <https://doi.org/10.1177/2053951716679679>
- Nissenbaum, H. (2004) Privacy as contextual integrity. Washington Law Review, 79(1), pp.119-158.
- Nissenbaum, H. (2010) Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford: Stanford University Press.
- Nouwens, M., Liccardi, I., Veale, M., Karger, D. and Kagal, L. (2020) Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp.1-13. Available at: <https://doi.org/10.1145/3313831.3376321>
- Porter, M.E. (1980) Competitive Strategy: Techniques for Analyzing Industries and Competitors. New York: Free Press.
- Raji, I.D. and Buolamwini, J. (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp.429-435. Available at: <https://doi.org/10.1145/3306618.3314244>
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1135-1144. Available at: <https://doi.org/10.1145/2939672.2939778>
- Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), pp.206-215. Available at: <https://doi.org/10.1038/s42256-019-0048-x>
- Sandvig, C., Hamilton, K., Karahalios, K. and Langbort, C. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry, International Communication Association preconference.
- Sen, S. and Bhattacharya, C.B. (2001) Does doing good always lead to doing better? Consumer reactions to corporate social responsibility. Journal of

Marketing Research, 38(2), pp.225-243. Available at: <https://doi.org/10.1509/jmkr.38.2.225.18838>

- Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F.N., Arvanitakis, G., Benevenuto, F., Gummadi, K.P., Loiseau, P. and Mislove, A. (2018) Potential for discrimination in online targeted advertising. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, pp.5-19.
- Suchman, M.C. (1995) Managing legitimacy: Strategic and institutional approaches. Academy of Management Review, 20(3), pp.571-610.
- Susser, D., Roessler, B. and Nissenbaum, H. (2019) Online manipulation: Hidden influences in a digital world. Georgetown Law Technology Review, 4(1), pp.1-45.
- Sweeney, L. (2013) Discrimination in online ad delivery. Communications of the ACM, 56(5), pp.44-54. Available at: <https://doi.org/10.1145/2447976.2447990>
- Teece, D.J., Pisano, G. and Shuen, A. (1997) Dynamic capabilities and strategic management. Strategic Management Journal, 18(7), pp.509-533.
- Tornatzky, L.G. and Fleischer, M. (1990) The Processes of Technological Innovation. Lexington, MA: Lexington Books.
- Tucker, C.E. (2014) Social networks, personalized advertising, and privacy controls. Journal of Marketing Research, 51(5), pp.546-562. Available at: <https://doi.org/10.1509/jmr.10.0355>
- Veale, M. and Zuiderveen Borgesius, F. (2021) Demystifying the draft EU Artificial Intelligence Act. Computer Law Review International, 22(4), pp.97-112.
- Wachter, S., Mittelstadt, B. and Floridi, L. (2017) Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), pp.76-99. Available at: <https://doi.org/10.1093/idpl/ix005>
- Waldman, A.E. (2020) Cognitive biases, dark patterns, and the privacy paradox. Current Opinion in Psychology, 31, pp.105-109. Available at: <https://doi.org/10.1016/j.copsyc.2019.08.025>
- Yeung, K. (2017) 'Hypernudge': Big Data as a mode of regulation by design. Information, Communication & Society, 20(1), pp.118-136. Available at: <https://doi.org/10.1080/1369118X.2016.1186713>
- Zuboff, S. (2019) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs.