

Original Research

Received 23 June 2026

Revised 27 July 2026

Accepted 09 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (16s): 58–73

KEYWORDS:

forensic consumer psychology; digital deception; fake online reviews; dark patterns; deceptive advertising; consumer vulnerability

Forensic Consumer Psychology: An Evidence-Based Framework for Detecting Deception and Digital Market Manipulation

¹Raiyan Haider, ²Prof. Dr. Tosendra Dwivedi, ³Dr Charu Chugh, ⁴Robin Augustine, ⁵Dr. Rati Khurana, ⁶Dr Manish Dubey, ⁷Sandeep Kumar Srivastava

¹Graduate, University of Lancashire, Preston, UK.

Orchid id: 0009-0002-1153-9987

Email: raiyanhaider6@gmail.com (Corresponding author)

²Professor of Psychology, IILM University,

Orcid Id: 0000-0002-5766-0363,

Email: tosendra.dwivedi@iilm.edu

³Associate Professor, IILM University, Greater Noida.

Orcid Id : 0009-0007-7444-6952

charu.chugh@iilm.edu

⁴Education Secretary of Department of Education & Secretary Satna Diocesan Society, Madhya Pradesh, India

Email Address: kachappillyrobin@gmail.com

⁵Assistant Professor, IILM University, Gurugram

Orcid Id : 0000-0002-0986-3617

Khurana.rati29@gmail.com

⁶Assistant Professor Amity school of communication Amity university Madhya Pradesh gwalior

mdubey1@gwa.amity.edu, manish.dubeymd01@gmail.com

Orchid-<https://orcid.org/0009-0005-1240-690>

⁷Govt Assistant Teacher, Uttar Pradesh, India

Email: sandeepau29@gmail.com Orcid <https://orcid.org/0009-0007-8078-7251>

ABSTRACT: Digital markets expose consumers to deception through misleading advertising, fabricated reviews, concealed costs, and interface designs that steer choices while obscuring relevant information. This paper proposes a forensic consumer psychology framework for identifying such practices by integrating deceptive cues, choice-architecture features, and consumers' cognitive and psychographic characteristics. It situates detection within the interaction between persuasive content and interface design, recognizing that consumer suspicion, vulnerability, and defensive processing can shape responses to potentially deceptive encounters. The framework distinguishes observable indicators of manipulation from the processes through which they may influence judgment and consumer outcomes, providing a structured basis for examining digital deception across advertising, reviews, pricing, and consent interfaces. By connecting individual-level psychology with market practices and digital choice environments, it offers an evidence-based foundation for consumer research, detection strategies, and the assessment of consumer harm.

INTRODUCTION:

Consumer deception now operates across connected commercial settings rather than within advertising alone. Promotional claims can misstate product performance, reviews can manufacture apparent peer endorsement, and interfaces can arrange choices so that a consumer accepts costs or data practices that were difficult to perceive. These forms differ in their immediate mechanism, but each can alter the information, attention, or control available at the point of decision. Research on Internet deception identifies tactic selection as dependent on the target and the identity presented by the deceiver, with consequences for trust in electronic commerce. A contemporary forensic approach therefore examines the communication, the consumer's inference, and the market context together.

The evidence also shows that deception is not restricted to demonstrably false sentences. Truthful statements may imply unsupported conclusions, promotional formats may conceal sponsorship, and low headline prices may suppress later charges [1], [2]. Fake reviews create a related problem because their wording can resemble ordinary experience while distorting the distribution of observed opinion [3]. Digital fraud extends the pathway from influence to loss through impersonation, urgency, personalization, and payment requests. Generative artificial intelligence intensifies these risks through realistic content, targeted personalization, and automated attack infrastructure [4].

This analysis treats forensic consumer psychology as an applied evidence system. It links deceptive stimuli to cognitive processing, vulnerability, behavioral response, and institutional accountability. The sections that follow first define the relevant constructs, then synthesize methods across advertising, reviews, dark patterns, and fraud. Findings are organized around the evidentiary signals that can support detection while preserving uncertainty where automated or consumer judgments remain fallible.

Conceptual Foundations: Deception, Consumer Cognition, Vulnerability, and Autonomy

From False Claims to Misleading Implications and Shrouded Prices

A narrow definition of deception would identify an objectively false claim. That definition misses cases in which the literal statement is accurate but the communication induces a materially misleading inference. Hastak and Mazis classify truthful-but-misleading advertising through legal cases and psychological theories, showing that implication depends on how consumers process omitted, qualified, or contextually framed information [1]. Advertising deception therefore concerns the relationship between claim, inference, and decision, not sentence-level falsity alone. A forensic assessment should ask what a reasonable audience is likely to understand, which qualification is salient, and whether the seller benefits from the resulting misunderstanding.

Native advertising illustrates the importance of source recognition. When a commercial message resembles legitimate news, consumers may treat its content as independent reporting rather than persuasion. A field experiment on mobile search examined sponsorship disclosure in native advertising, directly linking disclosure design to consumer deception [5]. Fake news advertising also generated measurable market traffic: after a Federal Trade Commission shutdown, the probability that a product site received no new visits increased by 22 percent, despite substitution toward regular advertisements [6]. The result indicates that deceptive format can recruit consumers who would not otherwise reach the product site.

Shrouded pricing operates through omission rather than a false headline. Firms may advertise an initial product while obscuring expensive add-ons, such as complementary supplies or required services. Gabaix and Laibson model this practice as profitable when some consumers are myopic or unaware, even in competitive markets and where advertising is costless [2]. Drip pricing similarly introduces add-on components after a seemingly attractive price has established a reference point. Yet consumer response is conditional: expectations and familiarity can make detailed pricing appear fairer and can reduce negative evaluations of the practice [7]. Detection must therefore record both the price sequence and the consumer's prior awareness.

Trust, Persuasion Knowledge, Cognitive Processing, and Consumer Vulnerability

Consumers do not process every commercial signal with the same level of attention. **Persuasion knowledge** allows people to recognize an attempt to influence them, but recognition can be incomplete, delayed, or directed toward the wrong feature. Training consumers to detect one tactic, such as an illegitimate endorser, can increase attention to that tactic while leaving them more exposed to a restrictive footnote; general skepticism produced broader detection without that reported goal-shielding effect [8]. A forensic system should therefore distinguish tactic recognition from comprehensive scrutiny.

Trust is both a decision aid and a consequence of deception. Deceptive advertising can induce defensive stereotyping, reducing credibility assigned to later advertising from the same source and from unrelated sources [9]. Perceived deception also affects product and retailer satisfaction, with downstream effects on word-of-mouth that differ between online and offline channels [10]. These findings make distrust an outcome variable, not simply a consumer trait. A consumer who rejects one message may also withdraw from legitimate information, creating costs beyond the original transaction.

Consumer vulnerability is best defined as a state in which restricted access to or control over resources inhibits marketplace functioning and exposes the consumer to harm [11]. The definition is dynamic: vulnerability can arise from digital literacy, time pressure, prior victimization, impulsive buying, disability, age, or the design of the environment. A Korean survey study found that advertising falsity affected harm through credibility, distorted risk assessment, and purchase pressure, while low digital literacy was associated with an adverse effect approximately 1.7 times that of high literacy; the conditional indirect effect for a high-vulnerability group was approximately 2.3 times that of a low-vulnerability group [12]. These estimates are context-specific, but they support a mechanism in which the same stimulus produces unequal consequences.

Autonomy adds a normative dimension. A choice may be formally available while being shaped through concealment, obstruction, or exploitative timing. The autonomy account of consumer law treats dark patterns as problematic when choice architecture exploits behavioural biases rather than merely when it lowers welfare [13]. This distinction matters for forensic classification: a system should document how a design affects comprehension, alternatives, timing, and control before inferring an autonomy violation.

RESEARCH DESIGN: EVIDENCE SYNTHESIS AND APPLIED FORENSIC SYSTEM DEVELOPMENT

Corpus Mapping, Thematic Coding, and Methodological Clusters

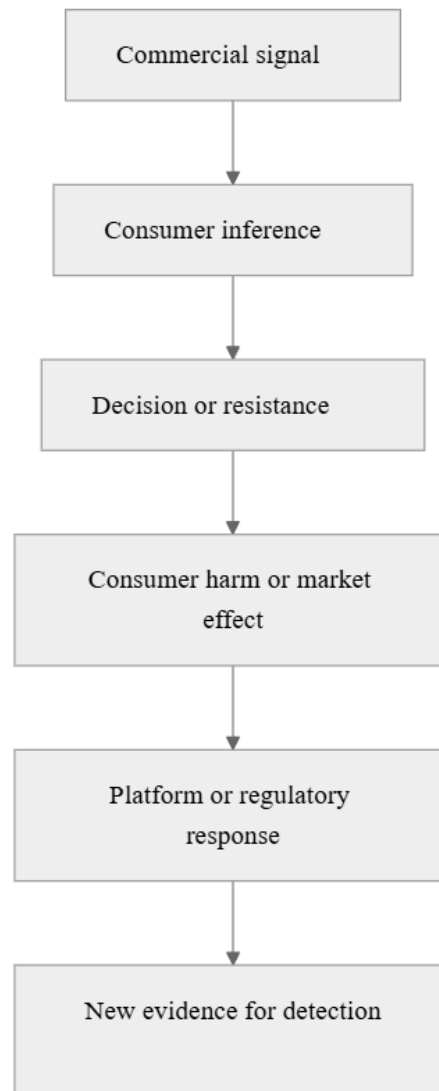
The evidence base spans conceptual articles, systematic reviews, experiments, surveys, field observations, computational classifiers, legal analyses, and economic models. A defensible synthesis should code each study for threat type, commercial setting, target population, stimulus, cognitive mechanism, outcome, method, and evidentiary limitation. Reviews of deceptive marketing identify broad construct diversity and changing definitions across advertising, ethics, covert communication, and policy research [14]. A separate review of 193 advertising-deception publications reports strong reliance on Persuasion Knowledge Theory, self-reported measures, and Western contexts, with limited use of neuroscientific methods [15].

The coding scheme separates three analytical layers:

Layer	Questions for coding	Examples of evidence
Stimulus	What is presented, omitted, or arranged?	False claim, native format, fake review, hidden fee, consent interface
Processing	What inference or pressure follows?	Credibility, risk distortion, suspicion, reactance, urgency
Consequence	What changes for consumer or market?	Purchase, satisfaction, distrust, financial loss, word-of-mouth, autonomy

Methodological clusters should not be treated as interchangeable. Experiments identify effects under controlled exposure, field studies observe behavior in market settings, surveys estimate associations and perceived mechanisms, and classifiers assess feature-performance relations. Legal and economic analyses add normative or equilibrium implications that do not substitute for psychological outcome measures. The synthesis therefore compares convergence in mechanisms rather than forcing all studies into a common effect estimate.

A corpus map can be represented as a sequence from commercial signal to institutional response:



Integrating Experimental, Field, Survey, Computational, Legal, and Economic Evidence

Experimental evidence is strongest for causal claims about exposure. Studies of deceptive claims, consent interfaces, dark patterns, and suspicious reviews can manipulate presentation and compare responses. Field evidence tests whether effects persist in ordinary traffic or purchasing conditions, as shown by the decline in visits after fake news advertisements were removed [6]. Survey and structural models are useful for individual differences and mediating pathways, but self-report can diverge from observed conduct. The Korean study's structural equation and moderated mediation analyses clarify cognitive pathways, while its single-country, online-purchase sample limits transportability [12].

Computational evidence addresses scale. Fake-review research includes linguistic features, supervised classification, natural language processing, and generated training data. A consumer-electronics framework reported an F-score of 82 percent, but performance varied by city and domain [16]. Salminen and colleagues found near-perfect classification in a task involving generated reviews, while human raters showed lower accuracy and agreement; the finding supports machine assistance but does not establish universal performance against adaptive adversaries [17]. Screen-reader advertising detection offers another example, with multimodal features producing F1 scores of 0.86 on a test dataset and 0.88 on random websites [18].

Legal analysis defines protected interests and prohibited design features. The EU-focused mapping of dark patterns distinguishes deception, manipulation, and distortion or impairment as autonomy-violation types [19]. Economic models test

how regulation, competition, consumer sophistication, and seller incentives alter outcomes. These models caution against assuming that every restriction improves welfare: higher penalties can reduce consumer surplus in some markets [20], and limiting deceptive practices can produce counterintuitive buyer effects in experience-good markets [21]. An applied system should therefore report legal risk, psychological risk, and predicted market effect separately.

Scope, Evidence Constraints, and Standards for the Applied Demonstration

The synthesis covers deceptive advertising, reviews, dark patterns, and digital fraud, with generative AI treated as an amplifier rather than a separate harm category. The unit of analysis is a consumer-facing episode: an advertisement, listing, review set, interface sequence, or fraud communication. The demonstration should not infer intent from one linguistic feature. Instead, it should preserve the original content, reconstruct the consumer pathway, identify the relevant comparison condition, and record uncertainty.

Three standards govern interpretation. First, causal language is reserved for experimental or otherwise supported causal designs; observational associations remain associations. Second, model performance is reported with its dataset, domain, and endpoint, rather than converted into general accuracy. Third, vulnerability is treated as context-dependent. Evidence indicates susceptibility across groups to dark patterns, with only weak evidence that common proxies such as income, education, or age materially explain differences [22]. At the same time, fraud studies link victimization to age, impulsivity, addictive tendencies, and risky online routines [23]. These findings are not contradictory because dark-pattern susceptibility and fraud victimization involve different tasks, exposures, and outcomes.

The applied demonstration consequently functions as a structured case analysis, not a validated diagnostic instrument. It can prioritize review, expose evidence gaps, and generate a transparent record for human judgment. It cannot establish deceptive intent, predict an individual's future loss, or replace legal adjudication. This boundary prepares the findings by separating observable signals from claims about harm.

Findings I: Deceptive Advertising, Disclosure, and Consumer Harm

False and Truthful-but-Misleading Claims, Native Advertising, Humor, and Drip Pricing

Advertising deception ranges from explicit falsity to implication, omission, and presentation. Evidence from ski-resort advertising illustrates how claims can be systematically exaggerated when the commercial return is greater: self-reported snowfall exceeded comparable government measures, with larger discrepancies on weekends and among resorts with expert terrain or no money-back guarantee [24]. This field evidence connects claim inflation to incentives rather than portraying deception as random error. Crowdfunding models likewise show that false advertising interacts with price signals and regulation, and that stricter regulation does not mechanically eliminate low-quality imitation [25].

Humor can reduce scrutiny by changing the interpretive frame. In a content analysis of 238 advertisements, deceptive claims appeared in 73.5 percent of humorous ads, and 74.5 percent of those claims were masked by humor [26]. These figures describe coded advertisements, not consumer misled rates. Their forensic value lies in directing attention to the joint presence of amusement and substantive claim content. A message framed as a joke may still generate a factual impression, especially when the qualification is less memorable than the punchline.

Native advertising creates a different source-monitoring problem. Consumers may process a sponsored message using expectations associated with journalism or search results. The mobile-search field experiment on sponsorship disclosure supplies direct evidence that disclosure design matters in this setting [5]. Fake news advertising also produced incremental visits that disappeared after regulatory intervention, indicating that format affected market entry rather than simply reinforcing existing demand [6].

Drip pricing and shrouded attributes make the temporal order of information part of the deception. A low initial price establishes attraction before mandatory fees, add-ons, or usage costs appear. Yet familiarity modifies response: consumers who expect drip pricing may evaluate detailed prices more positively and perceive the practice as less deceptive [7]. This result warns against using consumer acceptance as proof of transparent conduct. Normalization may reflect adaptation to a market practice rather than informed consent.

Consumer Inference, Anticipated Harm, Distrust, and Channel Differences

The psychological pathway generally begins with an inference about credibility, quality, scarcity, or price. Internet advertising research identifies credibility, distorted risk assessment, and purchase pressure as mediators between falsity and harm [12]. Performance exaggeration showed the strongest direct association with consumer harm among the dimensions examined in that study, while risk assessment distortion was the strongest indirect pathway [12]. These findings support a forensic distinction between the deceptive signal and the cognitive error it induces.

Consumers also evaluate the anticipated consequences of being wrong. Experiments separating perceived deception from anticipated harm found that greater anticipated harmfulness increased the diagnosticity of perceived deception and contributed to negative brand attitudes and purchase intentions [27]. A misleading claim about a trivial attribute and one affecting health, finances, or safety may therefore receive different scrutiny even when their formal structure is similar. Harm assessment should be recorded as an endpoint, not assumed from the presence of deception.

Distrust extends beyond the immediate advertiser. Initial deception can produce defensive beliefs about advertising and marketing generally, reducing responses to subsequent messages from the source and from second parties [9]. Retailer perceptions also travel through satisfaction: perceived deception affects retailer satisfaction through product satisfaction and affects word-of-mouth through retailer satisfaction, with moderation by online versus offline channel [10]. Online settings may heighten the importance of information search and perceived Internet usefulness, whereas shopping enjoyment and materialism have greater explanatory relevance in offline contexts [28].

The evidence supports a layered harm model rather than a single conversion metric. A message may generate a purchase, distort risk assessment, reduce later trust, or produce negative word-of-mouth. A forensic review should therefore preserve both immediate and delayed endpoints. It should also distinguish a consumer's failure to detect a claim from a firm's responsibility for constructing an environment in which a misleading inference was predictable.

Evidence Matrix of Advertising Tactics, Psychological Signals, and Consumer Consequences

The matrix below aligns observable advertising features with the psychological signals and consequences documented across the evidence base.

Advertising tactic	Psychological signal	Consumer consequence	Forensic implication
Performance exaggeration	Inflated credibility and distorted risk assessment [12]	Purchase pressure and harm	Compare claim with independent evidence and qualification
Fake news format	Source confusion and borrowed editorial credibility [6]	New site visits and product uptake	Record visual and textual similarity to news
Native advertising	Weak sponsorship recognition [5]	Commercial message treated as independent information	Test disclosure prominence and timing
Humor masking	Reduced attention to claim scrutiny [26]	Misleading impression may persist after entertainment	Separate humorous frame from factual proposition
Drip pricing	Anchoring on headline price [2]	Acceptance of later add-ons or revised fairness judgment [7]	Capture the complete price sequence
False scarcity or urgency	Compressed deliberation and purchase pressure [12]	Faster choice and possible harm	Verify whether scarcity is substantiated
Broad deceptive claim	Defensive stereotyping after detection [9]	Distrust of later advertising and reduced word-of-mouth	Measure spillover beyond the original message

The matrix shows why disclosure alone cannot be evaluated without considering placement, inference, and consequence. It also establishes the analytic bridge to reviews, where credibility is distributed across many apparently independent voices rather than one sponsored message.

Findings II: Fake Reviews, Credibility Cues, and Detection Challenges

Review Manipulation, Detection Approaches, and the Limits of Automated Classification

Online reviews function as third-party information, which consumers often regard as more credible than promotional communication. That credibility creates an incentive to manipulate positive, negative, or neutral reviews. The synthesis of fake-review research reports damage to the informativeness of genuine reviews, distorted product perceptions, and reduced marketplace credibility [3]. Manipulation may benefit a seller, harm a competitor, or alter the apparent consensus without changing the product itself.

Detection methods span linguistic, behavioural, metadata, network, and machine-learning features. A survey of the field identifies recurring strengths and bottlenecks across these approaches, emphasizing that credible classification remains difficult despite extensive methodological development [29]. A consumer-electronics framework constructed a domain dataset, defined features, and evaluated classifiers across four cities, reaching an F-score of 82 percent, with AdaBoost performing best under the reported statistical comparison [16]. The domain and location dependence of this result caution against treating one classifier as a general-purpose detector.

Generated text changes the adversarial environment. Salminen and colleagues used ULMFiT and GPT-2 to create fake product reviews and then trained a classifier on the resulting data. The classifier performed near perfectly in that task and also detected human-generated fake reviews, whereas human raters had lower accuracy and agreement [17]. Synthetic training data can therefore help identify recurring textual regularities, but it may also encourage models to detect the generator or dataset rather than deception itself. A forensic system should retain provenance, test on human and generated material, and expect performance to change as authors adapt.

Automation is particularly valuable for triage, not final attribution. A prediction that a review is suspicious should trigger examination of timing, reviewer history, product relation, language repetition, and campaign patterns. It should not automatically remove a review, because unusual style can reflect translation, disability, expertise, emotional experience, or a genuine outlier. False positives can suppress legitimate criticism and further distort the marketplace.

Suspicion, Cognitive Dissonance, Poster Motivation, and Marketplace Effects

Consumers use cues to judge whether a review is authentic, but suspicion itself changes evaluation. Research on suspicious reviews examines effects on opinion valuation, brand and website attitudes, and purchasing, while identifying the cues consumers use to assess credibility [30]. Suspicion can reduce the weight assigned to the review, yet it may also generalize to the brand or platform. The relevant endpoint is therefore not simply whether the consumer labels a review fake, but how that judgment changes evidence integration.

Cognitive dissonance provides a mechanism for interpreting these judgments. A conceptual model validated with responses from 342 participants found that cognitive processes affected attitudes toward detecting fake reviews, while product knowledge moderated the relationship between review attitudes and perceived intention to post deceptive reviews [31]. Product knowledge may help consumers identify implausible claims, but it can also make them confident in a mistaken interpretation. Detection systems should record confidence separately from correctness.

Poster motivation is heterogeneous. A study using natural language processing and logistic regression linked Machiavellian traits to calculated use of fake reviews, while narcissistic and psychopathic traits showed patterns associated with avoiding or being deterred from the demanding task of producing plausible deception [32]. These findings are preliminary and should not be translated into personality screening of reviewers. Their practical value is narrower: deceptive activity may reflect strategic campaigns, identity concerns, or short-term opportunism, so temporal and relational features may be more informative than text alone.

At marketplace level, manipulation reduces the signal value of all reviews. A platform with many suspicious entries imposes higher verification costs on consumers and may cause them to abandon peer information. Genuine businesses can suffer from negative competitor campaigns, while consumers may purchase inferior products because apparent consensus

substitutes for inspection. The appropriate response combines automated flagging, reviewer-level safeguards, transparent appeal, and platform-level monitoring rather than reliance on a single probability score.

Comparative Table of Review Signals, Detection Methods, and Likely False Positives

Review signal	Detection method	Evidence or endpoint	Likely false positive
Recurrent linguistic patterns	Language model or supervised classifier	Near-perfect performance in a generated-review task [17]	Template use by legitimate sellers or translated text
Domain-specific feature combinations	AdaBoost and feature framework	82% F-score in consumer electronics [16]	Genuine technical reviews with unusual metadata
Suspicious credibility cues	Consumer judgment and attitude measures	Changes in opinion valuation, brand, website, and purchase responses [30]	Negative but authentic experience
Product-knowledge mismatch	Cognitive and survey modelling	Product knowledge moderates detection-related relationships [31]	Expert minority opinion
Reviewer personality indicators	NLP plus logistic regression	Trait-specific associations with deceptive posting [32]	Personality inference from sparse text
Review-set distortion	Multidisciplinary synthesis	Reduced informativeness and credibility of genuine reviews [3]	Sudden legitimate demand or coordinated real events

The table separates a signal from the conclusion drawn from it. No listed method establishes deception in every case, and the absence of a signal does not establish authenticity. Review evidence therefore leads directly to the broader problem of dark patterns and fraud, where manipulation may occur before a consumer ever reaches a review page.

Findings III: Dark Patterns, Digital Fraud, and Unequal Consumer Exposure

Manipulative Choice Architecture, Consent Interfaces, and Autonomy Violations

Dark patterns alter the conditions under which users select, refuse, purchase, subscribe, disclose data, or exit. Their force lies in the interaction between interface structure and ordinary limits on attention. A systematic review of dark patterns and digital nudging identifies recurring contexts, characteristics, theories, and methods across the literature [33]. Consumer complaints from a marketplace also indicate that several patterns can appear together rather than as isolated design choices [34]. Forensic analysis must therefore reconstruct the full sequence, including defaults, visual emphasis, wording, friction, and the availability of an equivalent alternative.

Consent interfaces show how manipulation can concern data as well as money. Ambiguous language and limited choice increase perceived privacy threat and reduce perceived control, activating persuasion knowledge and psychological reactance; transparent designs improve comprehension and autonomy perceptions [35]. The design may thus produce resistance after exposure, but resistance does not erase the prior impairment of informed choice. Targeted advertising can reactivate memories of earlier consent experiences, making interface history part of the evidentiary record [35].

An autonomy-based account distinguishes deception, manipulation, and distortion or impairment. The mapping of 59 dark patterns to EU Digital Services Act categories supplies design-oriented factors for deciding when an interface crosses those boundaries [19]. This approach avoids reducing every unwanted outcome to low welfare. A user might receive a discount and still experience an autonomy violation if the interface obscures refusal or exploits a biased choice process. Conversely, a persuasive but comprehensible recommendation is not automatically deceptive.

Empirical evidence complicates assumptions about who is susceptible. A study found strong susceptibility across groups and only weak evidence that income, education, or age materially explained responsiveness to dark patterns [22]. Added payment friction reduced effectiveness, suggesting greater exposure when stored payment details allow a single-click purchase [22]. Research on quick-commerce interfaces similarly found that false scarcity, basket sneaking, and confirm-shaming increased perceived betrayal, punitive intent, and negative word-of-mouth, although brand trust did not change across the tested patterns and loyalty declined only for confirm-shaming [36]. Immediate retaliation and longer-term trust loss should therefore be measured separately.

Online gambling illustrates how design can combine inducement and obstruction. Sludge consists of detrimental friction, such as difficult withdrawal, while dark patterns include deceptive interface features and dark nudges include broader framing that presents gambling as fun [37]. The classification is useful because harm can arise from both excessive ease of entry and excessive difficulty of exit. Consent and payment investigations should consequently inspect the entire transaction lifecycle.

Cyber-Fraud Pathways, Social Engineering, and Generative-AI-Enabled Deception

Digital fraud differs from ordinary persuasion because the communication seeks unauthorized transfer, disclosure, access, or impersonation. Routine Activity Theory directs attention to motivated offenders, suitable targets, absent guardianship, market conditions, and unwitting collaborators [38]. Social presence and media affordances also matter. Communication technologies that mask cues of deceit and reduce social presence can increase willingness to rationalize fraud, while anonymity, rehearsability, and parallelism can support fraudulent conduct [39]. The same media can expose cues and increase social presence, which may deter rationalization.

Generative AI expands three attack capacities: realistic content creation, advanced targeting and personalization, and automated attack infrastructure [4]. These capacities can make phishing messages grammatically credible, align them with a target's social identity, and scale individualized contact. A forensic system should therefore treat polished language as weak evidence of legitimacy. Identity, channel provenance, request consistency, payment destination, and independent confirmation become more informative than style.

Fraud communications often combine role legitimacy with pressure. An impersonated bank, employer, government body, or family member can make the request appear socially obligatory. A dual-regulation model distinguishes profit-induced fraud,

which appeals to gain or loss avoidance, from role-embedded fraud, which relies on recognized identities and relational legitimacy [40]. Self-regulatory and social-regulatory efficacy reduced vulnerability among digitally active young adults, but information literacy moderated these protections differently by fraud type [40]. Education that teaches factual checking may not protect adequately against a message that exploits relational obligation.

The forensic boundary remains important. Suspicious content is not proof of fraud, and a failed payment request may reflect error rather than intent. Evidence should preserve message headers where lawful, links, timestamps, account identity, requested action, and the availability of independent verification. Human review is necessary when classification may trigger account restriction, public accusation, or referral to authorities.

Vulnerability Across Age, Digital Literacy, Routine Activities, and Self-Regulation

Vulnerability is distributed unevenly, but no single demographic profile explains exposure. Older adults in EU data were more likely to experience fraud in general, though not online fraud specifically; identity theft involving impersonation of a reputable organization was particularly relevant, and older victims reported more anger, embarrassment, and negative physical-health impacts despite being less likely to report financial impact [41]. These results reject a simple claim that older age universally increases online victimization. They support targeted attention to impersonation and non-financial consequences.

A large survey framework links cyber-fraud victimhood to age, urgency and sensation-seeking impulsivity, addictive tendencies, and frequent risky online routines [23]. One-off and repeat victims showed little distinction in the reported model, suggesting that prevention should address routine exposure rather than assume a single accidental lapse. Routine activities can place users repeatedly in contact with high-risk messages, platforms, or transactions. Guardianship therefore includes platform verification, social support, account controls, and interface protections, not only individual vigilance.

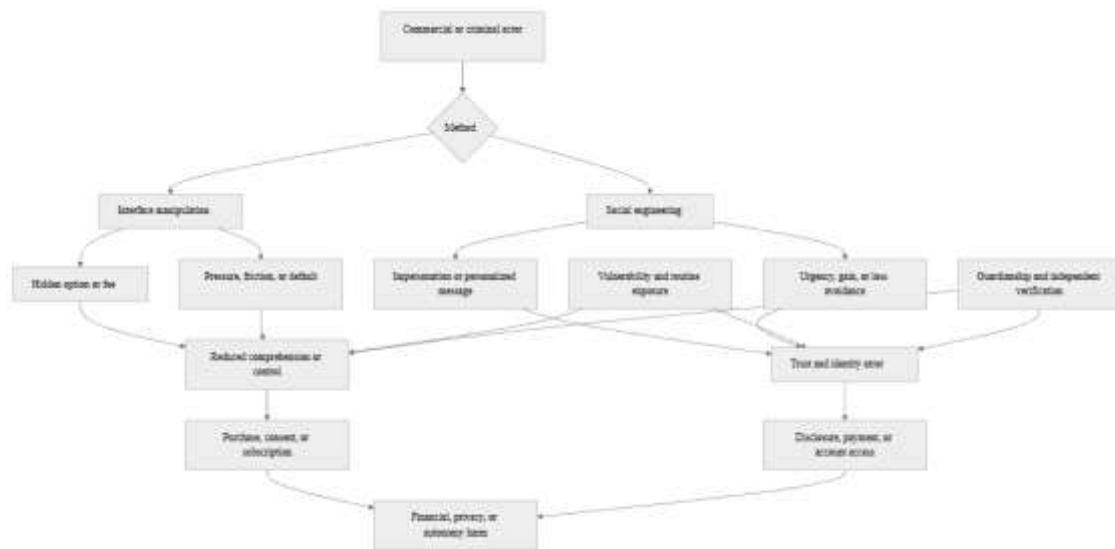
Digital literacy remains protective in some settings but is not a complete solution. In Internet advertising, lower digital literacy intensified the relationship between falsity and harm [12]. Research on AI-driven marketing among young consumers likewise reports that perceived manipulation mediated unethical consumer behaviour and that higher digital literacy moderated this relationship [42]. Yet general training can produce narrow vigilance, and users may still struggle against dark patterns because of limited awareness, partial control over cognitive biases, and preference for short-term benefits [43]. The burden should not be shifted entirely to consumers.

Accessibility creates additional exposure pathways. Blind screen-reader users reported deception when advertisements blended contextually into surrounding page content, while ad blockers did not reliably remove internal advertisements or could be used to deny access [18]. This case shows that a design can be deceptive through semantic and navigational context, even when visual distinctions appear obvious to sighted users. Detection should therefore test multiple modalities and preserve the user's actual interaction path.

The relationship between vulnerability and design is reciprocal. Dark patterns may affect broad populations, while fraud campaigns exploit more specific routines, identities, and capacities. Algorithmic personalization can intensify harm by identifying people who lack information or display behavioural biases, with distributional effects across identifiable groups [44]. A fair forensic system must examine exposure conditions without labeling the consumer as the cause of the harm.

Threat-and-Vulnerability Diagram for Comparing Interface Manipulation and Fraud

The comparison below separates the point of intervention from the endpoint of harm. Dark patterns primarily alter choice architecture, whereas fraud adds a deceptive actor and an unauthorized objective.



This diagram supports a comparative triage. For interface manipulation, investigators should capture the screen sequence, defaults, alternative paths, price changes, consent language, and exit friction. For fraud, they should add identity verification, channel provenance, request legitimacy, and payment destination. For both threats, the evidence should identify whether the consumer could understand the choice, whether refusal was materially harder, and whether the design or message exploited a known constraint.

The combined model also explains why consumer training alone is insufficient. A user may recognize a suspicious pattern but lack time, control, accessibility, or an independent channel for verification. Platforms can reduce exposure through transparent interfaces, friction at high-risk payment points, review integrity controls, and escalation routes. Regulators can require design accountability where the architecture predictably impairs autonomy. These interventions address the source and environment of harm rather than treating every adverse outcome as an individual failure.

Real-World Forensic Analysis: An Integrated Detection and Consumer-Response System

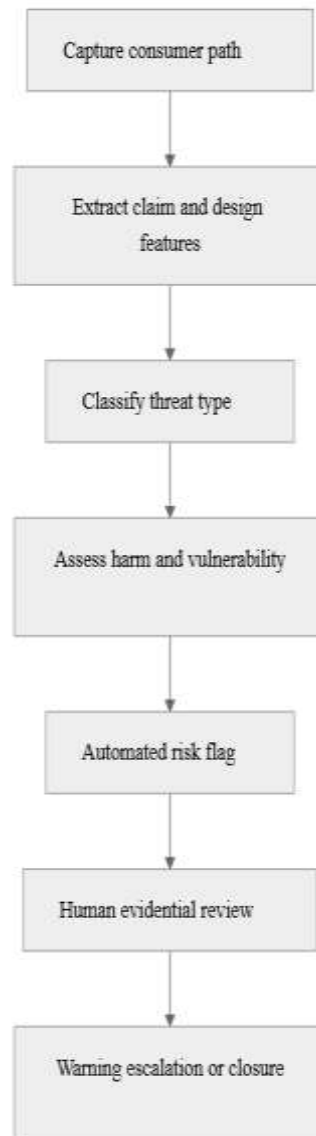
System Architecture: Threat Taxonomy, Evidence Inputs, and Risk-Triage Rules

The system classifies cases into deceptive advertising, manipulated reviews, dark patterns, and digital fraud. It treats deception as a relation between representation, consumer inference, and material consequence, rather than as a property of wording alone [1], [45]. Evidence includes screenshots, interface recordings, pricing states, disclosure position, review histories, message metadata, payment destinations, and the consumer's completed path. Triage prioritizes apparent financial loss, unauthorized access, impaired consent, repeated exposure, and exploitation of restricted consumer resources, consistent with the vulnerability framework of access and control [11].

A **risk register** should distinguish factual falsity, misleading implication, concealed cost, coerced choice, fabricated social proof, impersonation, and automated personalization. This distinction matters because a truthful statement can still induce a materially misleading inference [1], while algorithmic targeting can exploit information deficits or behavioural biases without changing the literal claim [44]. Accessibility conditions must be recorded because contextually blended advertisements can deceive screen-reader users even when visual users recognize an advertisement [18].

Operational Workflow: Capture, Feature Analysis, Risk Scoring, and Human Review

The workflow begins with evidence preservation before content changes: investigators capture the page, consent sequence, price progression, sender identity, and relevant transaction records. Feature analysis then examines linguistic claims, qualification prominence, urgency cues, default settings, exit friction, reviewer-network patterns, and requests for credentials or payment. Automated classification can assist this stage, but performance is domain-dependent: a consumer-electronics review classifier reached an F-score of 82% [16], whereas the broader literature describes automatic review detection as only partially successful [29].



Human review tests whether the flagged feature changed comprehension, control, or transaction outcome. Reviewers should compare alternative interfaces, independent product information, and the evidential basis for the alleged claim. A high score should trigger investigation, not establish liability, because machine-generated text and atypical but genuine consumer experiences can resemble deception [17].

Functional Outputs: Case File, Consumer Warning, Platform Escalation, and Audit Trail

The **case file** records the allegation, preserved evidence, analytical features, confidence level, suspected actor, consumer impact, and unresolved alternatives. A consumer warning should state the observable risk, the verification action, and the available remedy without asserting fraud where evidence establishes only suspicion. This distinction protects consumers while limiting reputational harm to legitimate sellers and reviewers.

Platform escalation should identify the violated design or content condition, affected transaction pathway, and requested corrective action. The audit trail preserves model version, reviewer decisions, evidence hashes, amendments, and reasons for closure. Such records permit later comparison between automated flags and confirmed cases, supporting accountability rather than opaque moderation.

Scenario-Based Walkthroughs, Evaluation Measures, and Safeguards Against Misclassification

In a concealed-fee case, the investigator compares the advertised price with the final price and documents when the add-on first appeared. In a fake-review case, the system combines language, timing, account relations, and product context, while recognizing that suspiciousness can alter brand, website, and purchase responses even without proof of fabrication [30]. In an impersonation case, investigators verify the sender through an independent channel and preserve the payment destination, since older adults are particularly exposed to impersonation-based identity theft and may experience substantial non-financial harm [41].

Evaluation should report precision, recall, false-positive rates, reviewer agreement, warning comprehension, prevented loss, and time to resolution. Safeguards include calibrated confidence labels, appeal routes, subgroup error testing, accessibility review, and mandatory human confirmation before account sanctions. Training should activate general skepticism rather than a single cue, since tactic-specific training can increase vulnerability to an unexamined tactic in the same message [8]. These outputs connect forensic detection with the institutional questions addressed next.

DISCUSSION:

Reconciling Detection, Consumer Protection, and Market-Level Trade-Offs

Contested Effects of Consumer Training, Skepticism, Disclosure, and Penalties

Consumer education can improve detection, but narrow instruction may redirect attention toward one cue and leave other deceptions unnoticed [8]. Broad skepticism can protect against varied tactics, yet generalized distrust may reduce confidence in legitimate advertising and second-party information after an initial deceptive encounter [9]. Disclosure therefore requires testing for comprehension and behavioural effect, not merely its presence; native-advertising experiments and consent research indicate that placement, ambiguity, and perceived control shape consumer interpretation [5], [35].

Penalties also produce non-linear outcomes. Economic models show that stricter limits on deception can reduce consumer welfare under particular competitive conditions, especially where advertising conveys information about quality or market entry [20], [21]. These findings do not justify deception; they indicate that enforcement should assess substitution, disclosure quality, seller heterogeneity, and downstream competition rather than assume that every restriction has identical effects.

From Individual Detection to Platform Accountability and Regulatory Design

Individual detection cannot correct systems that repeatedly generate misleading interfaces, rank manipulated reviews, or target consumers through inferred vulnerability. Platform duties should include traceable design decisions, review-integrity controls, accessible complaint channels, and evidence preservation. The EU Digital Services Act framework distinguishes deception, manipulation, and impairment of choice, giving investigators design-oriented categories for connecting interface evidence to autonomy violations [19].

Regulation should combine outcome evidence with process duties. Platforms need not prove that every consumer was deceived, but they should demonstrate transparent defaults, meaningful refusal paths, and controls against exploitative personalization. Algorithmic transparency remains difficult, yet policy must address how systems use information and behavioural predictions to impose unequal risks [44]. This approach shifts responsibility from isolated consumer vigilance toward accountable market architecture.

Evidence Gaps: Cross-Threat Validation, Longitudinal Harm, and Understudied Populations

The evidence base remains fragmented across advertising, reviews, interfaces, and fraud. Most studies examine one threat, one channel, or short-term attitudes, leaving uncertainty about whether detection skills transfer across contexts and whether distrust, financial loss, or privacy harm persists over time [14]. Automated models also require cross-domain validation because performance reported for generated reviews or one retail category cannot establish reliability elsewhere [16], [17].

Research underrepresents non-Western markets, disabled consumers, older adults in consumer-to-consumer platforms, and people facing restricted access to money, technology, or independent verification. Reviews of advertising research identify limited non-Western evidence and heavy reliance on self-report measures [15]. Future studies should combine behavioural experiments, field records, accessible simulations, platform audits, and longitudinal harm measures.

CONCLUSION:

A usable forensic practice links preserved consumer experience to psychological mechanism, technical feature, institutional responsibility, and remedy. It distinguishes false claims from misleading implications, suspicious signals from confirmed manipulation, and individual vulnerability from conditions created by market design. The integrated system proposed here combines automated screening with human review, accessible evidence capture, calibrated warnings, platform escalation, and auditable decisions.

Its governing principle is proportionality. Detection should reduce preventable harm without converting uncertainty into accusation, while consumer protection should address design and targeting conditions that individual skepticism cannot control. Evidence from deceptive advertising, fake reviews, dark patterns, and digital fraud supports this combined orientation, but cross-threat validation and longitudinal evaluation remain necessary before deployment at scale. Forensic consumer psychology therefore becomes most useful when it treats cognition as evidence, autonomy as an outcome, and accountability as a property of the whole marketplace.

REFERENCES

- [1] M. Hastak and M. B. Mazis, “Deception by implication: A typology of truthful but misleading advertising and labeling claims,” *Journal of Public Policy & Marketing*, vol. 30, no. 2, pp. 157–167, Sep. 2011, doi: 10.1509/jppm.30.2.157.
- [2] I. P. Riquelme, S. Román, and D. Iacobucci, “Consumers’ perceptions of online and offline retailer deception: A moderated mediation analysis,” *Journal of Interactive Marketing*, vol. 35, no. 1, pp. 16–26, Apr. 2016, doi: 10.1016/j.intmar.2016.01.002.
- [3] R. P. Hill and E. Sharma, “Consumer vulnerability,” *Journal of Consumer Psychology*, vol. 30, no. 3, pp. 551–570, Mar. 2020, doi: 10.1002/jcpy.1161.
- [4] 趙東澤, X. Jin, W. Ren, K. Dong, and C.-H. Jin, “Internet advertising falsity and consumer harm: A moderated mediation analysis of consumer cognitive processes and consumer vulnerability,” *Journal of theoretical and applied electronic commerce research*, vol. 21, no. 5, p. 133, Apr. 2026, doi: 10.3390/jtaer21050133.
- [5] M. Brennncke, “A theory of exploitation for consumer law: Online choice architectures, dark patterns, and autonomy violations,” *Journal of Consumer Policy*, vol. 47, no. 1, pp. 127–164, Dec. 2023, doi: 10.1007/s10603-023-09554-7.
- [6] E. Welch and J. M. Galvan, “Fifty years of deceptive marketing research: A systematic review and future research agenda,” *Psychology and Marketing*, vol. 41, no. 11, pp. 2805–2822, Jul. 2024, doi: 10.1002/mar.22085.
- [7] D. Cocks, L. Alvino, H. Luyten, and M. Ehrenhard, “Advertising: Company asset or liability? a systematic literature review on deception,” *International Journal of Consumer Studies*, vol. 50, no. 1, 2026, doi: 10.1111/ijcs.70187.
- [8] R. Barbado, Ó. Araque, and C. Á. Iglesias, “A framework for fake review detection in online consumer electronics retailers,” *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, Mar. 2019, doi: 10.1016/j.ipm.2019.03.002.
- [9] J. Salminen, C. Kandpal, A. Kamel, S. Jung, and B. J. Jansen, “Creating and detecting fake reviews of online products,” *Journal of Retailing and Consumer Services*, vol. 64, p. 102771, Sep. 2021, doi: 10.1016/j.jretconser.2021.102771.
- [10] S. R. Kodandaram, M. Sunkara, S. Jayarathna, and V. Ashok, “Detecting deceptive dark-pattern web advertisements for blind screen-reader users,” *Journal of Imaging*, vol. 9, no. 11, p. 239, Nov. 2023, doi: 10.3390/jimaging9110239.
- [11] S. Ahuja, J. Gunawan, N. Bielova, and C. Santos, “Dark patterns and the EU digital services act: Mapping autonomy violations and design factors,” *SSRN Electronic Journal*. 2025. doi: 10.2139/ssrn.5555765.
- [12] X. Gabaix and D. Laibson, “Shrouded attributes, consumer myopia, and information suppression in competitive markets,” *The Quarterly Journal of Economics*, vol. 121, no. 2, pp. 505–540, May 2006, doi: 10.1162/qjec.2006.121.2.505.
- [13] Y. Wu and T. Geylani, “Regulating deceptive advertising: False claims and skeptical consumers,” *Marketing Science*, vol. 39, no. 4, pp. 788–806, May 2020, doi: 10.1287/mksc.2020.1221.
- [14] S. Piccolo, P. Tedeschi, and G. Ursino, “How limiting deceptive practices harms consumers,” *The RAND Journal of Economics*, vol. 46, no. 3, pp. 611–624, Jul. 2015, doi: 10.1111/1756-2171.12099.
- [15] A. Zac, Y.-C. Huang, A. von Moltke, C. S. Decker, and A. Ezrachi, “Dark patterns and consumer vulnerability,” *Behavioural Public Policy*, pp. 1–50, Feb. 2025, doi: 10.1017/bpp.2024.49.

- [16] M. T. Whitty, “Predicting susceptibility to cyber-fraud victimhood,” *Journal of Financial Crime*, vol. 26, no. 1, pp. 277–292, Jan. 2019, doi: 10.1108/jfc-10-2017-0095.
- [17] J. Zinman and E. Zitzewitz, “Wintertime for deceptive advertising?,” *American Economic Journal Applied Economics*, vol. 8, no. 1, pp. 177–192, Dec. 2015, doi: 10.1257/app.20130346.
- [18] J. Zhang and J. Huang, “Deceptive advertising in a crowdfunding market,” *Transportation Research Part E Logistics and Transportation Review*, vol. 135, p. 101892, Feb. 2020, doi: 10.1016/j.tre.2020.101892.
- [19] H. Shabbir and D. Thwaites, “The use of humor to mask deceptive advertising: It’s no laughing matter,” *Journal of Advertising*, vol. 36, no. 2, pp. 75–85, Jun. 2007, doi: 10.2753/joa0091-3367360205.
- [20] G.-X. Xie, R. Madrigal, and D. M. Boush, “Disentangling the effects of perceived deception and anticipated harm on consumer responses to deceptive advertising,” *Journal of Business Ethics*, vol. 129, no. 2, pp. 281–293, Mar. 2014, doi: 10.1007/s10551-014-2155-2.
- [21] I. P. Riquelme and S. Román, “The influence of consumers’ cognitive and psychographic traits on perceived deception: A comparison between online and offline retailing contexts,” *Journal of Business Ethics*, vol. 119, no. 3, pp. 405–422, Feb. 2013, doi: 10.1007/s10551-013-1628-z.
- [22] D. Vidanagama, T. Silva, and A. S. Karunananda, “Deceptive consumer review detection: a survey,” *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1323–1352, Mar. 2019, doi: 10.1007/s10462-019-09697-5.
- [23] J. Sahut, M. Laroche, and É. Braune, “Antecedents and consequences of fake reviews in a marketing approach: An overview and synthesis,” *Journal of Business Research*, vol. 175, p. 114572, Feb. 2024, doi: 10.1016/j.jbusres.2024.114572.
- [24] L. J. Harrison-Walker and Y. Jiang, “Suspicion of online product reviews as fake: Cues and consequences,” *Journal of Business Research*, vol. 160, p. 113780, Mar. 2023, doi: 10.1016/j.jbusres.2023.113780.
- [25] S. Chatterjee, R. Chaudhuri, A. Kumar, C. L. Wang, and S. Gupta, “Impacts of consumer cognitive process to ascertain online fake review: A cognitive dissonance theory approach,” *Journal of Business Research*, vol. 154, p. 113370, Oct. 2022, doi: 10.1016/j.jbusres.2022.113370.
- [26] M. Borghi and P. Ratcharak, “Deceptive minds in digital spaces: The influence of the dark triad on posting fake online reviews,” *Psychology and Marketing*, vol. 42, no. 7, pp. 1932–1945, Mar. 2025, doi: 10.1002/mar.22212.
- [27] V. Singh, N. K. Vishvakarma, and V. Kumar, “Unveiling digital manipulation and persuasion in e-commerce: A systematic literature review of dark patterns and digital nudging,” *Journal of Internet Commerce*, vol. 23, no. 2, pp. 144–171, Mar. 2024, doi: 10.1080/15332861.2024.2330813.
- [28] T. C. de Oliveira, T. A. Coleti, M. Morandini, R. Balancieri, and A. L. de Oliveira, “Dark patterns nos marketplaces: uma investigação com base nas reclamações dos consumidores,” in *Anais do III Workshop Investigações em Interação Humano-Dados (WIDE 2024)*, Sociedade Brasileira de Computação, Oct. 2024, pp. 39–45. doi: 10.5753/wide.2024.245492.
- [29] I. Ju, C. D. Ham, and E. Yel, “Dark patterns in data-consent disclosures and consumer reactance to online behavioral advertising,” *Journal of Advertising*, vol. 55, no. 3, pp. 307–325, Dec. 2025, doi: 10.1080/00913367.2025.2593666.
- [30] S. Agrawal and S. Narayanan, “When design deceives: consumer reactions to dark patterns in quick commerce,” *Journal of Indian Business Research*, vol. 18, no. 2, pp. 37–54, Jan. 2026, doi: 10.1108/jibr-05-2025-0188.
- [31] P. Newall, “Sludge, dark patterns and dark nudges: A taxonomy of online gambling platforms’ deceptive design features,” *Addiction*, vol. 120, no. 10, pp. 1916–1923, Apr. 2025, doi: 10.1111/add.70085.
- [32] A. Hang Cheuk, C. Chih-Yuan, and H. K.W. Kevin, “Unveiling Formations and Outcomes of Digital Consumer Frauds: An Expanded Routine Activity Theory Perspective,” *Journal of the Association for Information Systems*, 2024.
- [33] A. J. Harrison, “The effects of media capabilities on the rationalization of online consumer fraud,” *Journal of the Association for Information Systems*, vol. 19, no. 5, pp. 408–440, May 2018, doi: 10.17705/1jais.00497.
- [34] M. Schmitt and I. Fléchais, “Digital deception: generative artificial intelligence in social engineering and phishing,” *Artificial Intelligence Review*, vol. 57, no. 12, Oct. 2024, doi: 10.1007/s10462-024-10973-2.
- [35] Y. Guo, M. Yu, X. Zeng, and C.-W. Tan, “When Self-Regulation Meets Social Persuasion: A Dual-System Explanation of Digital Fraud Vulnerability,” *Journal of the Association for Information Systems*, 2026.

- [36] S. Kemp and N. E. Pérez, "Consumer fraud against older adults in digital society: Examining victimization and its impact," *International Journal of Environmental Research and Public Health*, vol. 20, no. 7, p. 5404, Apr. 2023, doi: 10.3390/ijerph20075404.
- [37] U. A. Qadri, A. M. A. Moustafa, and M. A. Ghani, "They misused me! digital literacy's dual role in AI marketing manipulation and unethical young consumer behavior," *Young Consumers Insight and Ideas for Responsible Marketers*, vol. 26, no. 5, pp. 744–764, Jan. 2025, doi: 10.1108/yc-08-2024-2207.
- [38] V. Singh, N. K. Vishvakarma, and V. Kumar, "Investigating consumer challenges against dark patterns using grey influence analysis (GINA)," *Marketing Intelligence & Planning*, vol. 43, no. 5, pp. 1013–1035, Jan. 2025, doi: 10.1108/mip-08-2024-0538.
- [39] O. Bar-Gill, C. R. Sunstein, and I. Talgam-Cohen, "Algorithmic harm in consumer markets," *The Journal of Legal Analysis*, vol. 15, no. 1, pp. 1–47, Aug. 2023, doi: 10.1093/jla/laad003.
- [40] R. N. Aditya, "The psychology of deception in marketing: A conceptual framework for research and practice," *Psychology and Marketing*, vol. 18, no. 7, pp. 735–761, May 2001, doi: 10.1002/mar.1028.
- [41] N. S. Sahni and H. S. Nair, "Sponsorship disclosure and consumer deception: Experimental evidence from native advertising in mobile search," *Marketing Science*, vol. 39, no. 1, pp. 5–32, Jul. 2019, doi: 10.1287/mksc.2018.1125.
- [42] A. Rao, "Deceptive claims using fake news advertising: The impact on consumers," *Journal of Marketing Research*, vol. 59, no. 3, pp. 534–554, Aug. 2021, doi: 10.1177/00222437211039804.
- [43] S. Banerjee, S. Dutta, A. Biswas, and H. Kwak, "The impact of consumer expectations and familiarity on deceptive pricing in advertising: a view from drip pricing practice," *International Journal of Advertising*, vol. 43, no. 2, pp. 254–285, Nov. 2023, doi: 10.1080/02650487.2023.2282868.
- [44] A. E. Wilson, P. R. Darke, and J. Sengupta, "Winning the battle but losing the war: Ironic effects of training consumers to detect deceptive advertising tactics," *Journal of Business Ethics*, vol. 181, no. 4, pp. 997–1013, Sep. 2021, doi: 10.1007/s10551-021-04937-7.
- [45] P. R. Darke and R. Ritchie, "The defensive consumer: Advertising deception, defensive processing, and distrust," *Journal of Marketing Research*, vol. 44, no. 1, pp. 114–127, Feb. 2007, doi: 10.1509/jmkr.44.1.114.