# Ethics-Focused User Metrics: How User Experience Metrics Can Trigger Ethical Design Sensibility

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 5843


Fig. 1. Multi-stage study design and results showing stepwise increase in UX designers' rejection of dark patterns when provided with autonomy-focused evaluation data alongside standard UX metrics. Mixed effects analysis reveals large condition effects ($\chi^2(2) = 32.65$, $p < 0.001$) with substantial practical differences in design decisions.

How do evaluation frameworks influence designers' ethical decisions? This study investigates how different evaluation data affects designers' willingness to implement design patterns that infringe user autonomy. Through an experiment with 141 UX professionals, we demonstrate that evaluation frameworks systematically guide ethical judgments.

Participants evaluated 15 interfaces under three conditions: no user data, standard UX metrics, or autonomy-focused measures. The evaluation data was real feedback from 126 social media users, and participants were unaware interfaces contained dark patterns. Results showed clear progression: any user feedback increased rejection of problematic designs, but autonomy-focused evaluation nearly doubled rejection rates. Analysis of 1,313 explanations revealed frameworks also changed reasoning patterns, shifting focus from business justifications toward user well-being.

These findings reveal evaluation tools actively shape design ethics. To foster responsible technology, we must equip designers with frameworks that make ethical risks visible.

CCS Concepts: • **Human-centered computing** → **Usability testing**; **User interface design**; • **Social and professional topics** → *Codes of ethics.*

Additional Key Words and Phrases: Dark Patterns, Ethics, Manipulation, Social Media, Measuring, User Experience

**ACM Reference Format:**
Anonymous Author(s). 2025. Ethics-Focused User Metrics: How User Experience Metrics Can Trigger Ethical Design Sensibility. In *Proceedings of CHI '26*. ACM, New York, NY, USA, 31 pages.

## 1 INTRODUCTION

Most technology companies today rely on metrics to organize corporate success and guide design decisions. Nearly all organizations with larger UX departments—from Google's HEART framework [68] to Duolingo's focus on engagement


Manuscript submitted to ACM

and learning metrics [63]—have established quantitative measures that their design teams optimize against. These metrics are effective and useful, promising more objective decision-making that enables flat hierarchies and management at scale [72, 82].

Concurrently, the design community has shown sustained interest in ethical aspects of interactive design over recent decades. This is evidenced through value-sensitive design approaches [32], participatory design methods [71], professional ethics tools like conversation cards [35], and privacy-by-design frameworks [17]. One instance of unethical design that is increasingly acknowledged in the UX community is dark patterns—interface elements designed to manipulate user behavior—represent unethical design practices that undermine user autonomy, privacy, and wellbeing [36, 49].

However, a critical gap exists between ethical commitment and standard UX evaluation metrics that lack explicit ethical dimensions. While companies rely heavily on established metrics like the User Experience Questionnaire (UEQ) [45] or System Usability Scale [14], these tools focus primarily on usability and satisfaction without addressing ethical concerns. Research shows that UX professionals often employ "soft resistance" tactics when advocating for ethical design choices that conflict with business goals [84], and product managers face challenges negotiating social values in digital product design [47]. This creates a misalignment where designers committed to ethical practice may not have salient ethical considerations in their evaluation processes.

Our work demonstrates that *metrics make ethical values salient*—and ethically-focused metrics can *trigger ethical sensibility* in design decision-making. To test this hypothesis, we conducted a two-stage study. First, we developed an autonomy-focused evaluation scale by adapting the UEQ with supplementary items specifically designed to capture user reactions to unethical designs. We used a comprehensive taxonomy of dark patterns in social media, derived from academic literature and regulatory guidelines, as ground truth for unethical design practice. In our UX+Autonomy Evaluation Study, 126 social media users evaluated dark pattern interfaces using both standard UEQ metrics and our autonomy-focused scale.

The evaluation data from this UX+Autonomy Evaluation Study was then systematically fed into our main experiment with 141 UX professionals. When these designers were provided with autonomy-focused evaluation data alongside standard UX metrics, they rejected interfaces containing dark patterns at nearly double the rate (56.2% vs. 30.0%). This demonstrates that evaluation metrics are not neutral measurement tools but active design interventions that shape ethical judgment.

This finding has important implications as metrics become increasingly central to design practice. With the rise of automated A/B testing, computational design tools, and AI-driven interface generation, evaluation metrics will likely become even more influential in shaping design outcomes [79]. Our research suggests that incorporating ethical dimensions into these metrics can effectively guide designers toward more ethical decisions—not through external constraint, but by making ethical considerations salient during evaluation.

The paper makes three key contributions: (1) empirical evidence that evaluation metrics actively shape ethical design decisions rather than merely measuring them, (2) demonstration that autonomy-focused metrics can effectively identify manipulative design patterns and influence professional judgment, and (3) practical insights for integrating ethical considerations into existing UX evaluation workflows to support ethical design at scale.

In the subsequent sections, we first explore existing UX metrics and examine how dark patterns are assessed in the literature. We then describe our three-stage methodology: (1) developing an autonomy-focused evaluation scale and a social media dark pattern taxonomy, (2) collecting a perception dataset from social media users, and (3) testing how evaluation data shapes professional design decisions. Finally, we present our findings and discuss their implications for ethical design practice.

## 2 LITERATURE REVIEW

### 2.1 Ethics and Values in Design Practice

*2.1.1 Value-Sensitive Design Approaches.* The design community has demonstrated sustained interest in incorporating ethical considerations into design practice. Value-sensitive design [32] provides frameworks for considering human values throughout the design process. Participatory design approaches [71] emphasize involving stakeholders in design decisions that affect them. Reflective design [73] encourages designers to examine the assumptions and values embedded in their work.

*2.1.2 Professional Tools for Ethical Design.* Multiple tools have emerged to support ethical design awareness. The Ethical Design Toolkit [35] provides conversation cards to surface ethical considerations during design processes. The Privacy by Design framework [17] offers principles for embedding privacy protection into system architecture. Design ethics games and workshops [25, 52] help teams explore ethical implications of design decisions. Additionally, values in games research [59] provides foundational work for understanding how values are embedded in interactive systems.

*2.1.3 Challenges in Ethical Design Practice.* Research shows UX professionals often employ 'soft resistance' tactics when advocating for ethical design choices that conflict with business goals [84]. Chivukula et al. [24]'s interview studies with UX practitioners identified five dimensions of design complexity influencing ethical outcomes spanning individual, collaborative, and methodological framing. Karakus et al. [41] explores the ethical dimensions of accessible UX design, highlighting the challenge of balancing stakeholder interests with user intentions, particularly when dark UX patterns are intentionally inserted to manipulate user behavior.

UX practitioners must develop specific rhetorical strategies to argue for user-centered design decisions [69], demonstrating the systematic organizational pressures that challenge ethical design implementation.

*2.1.4 The Challenge of Implementation.* However, these tools often operate as separate processes rather than integrated evaluation approaches. They function as add-on considerations rather than core metrics that shape ongoing design decisions and business KPIs. This separation creates a disconnect between ethical awareness tools and the measurement metrics that most directly influence design decisions.

### 2.2 Foundations of User Experience Evaluation

*2.2.1 Traditional UX Evaluation Methods.* UX evaluation methods aim to quantify good vs. bad design through diverse approaches. According to ISO 9241-210, user experience encompasses "a person's perceptions and responses that result from the use or anticipated use of a product, system or service," including "all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use"—explicitly including potential for harm [39]. This comprehensive definition recognizes that user experience extends beyond immediate satisfaction to encompass broader impacts on user wellbeing and autonomy.

UX evaluation encompasses distinct methodological approaches with different strengths and applications. Jeffries et al. [40] described four major evaluation techniques: heuristic evaluation (expert inspection using systematic criteria), software guidelines (compliance checking against design rules), cognitive walkthroughs (step-by-step task analysis), and usability testing (empirical measurement with real users). These approaches differ fundamentally in their focus: *heuristics* provide systematic inspection criteria for expert evaluation of interface compliance, while *metrics* offer quantitative performance measures, and *qualitative methods* reveal user mental models and subjective experiences [1].

Methods range from *IsoMetrics* [34], which focuses on standard compliance, to the *Microsoft Desirability Toolkit* that captures subjective aspects through *Product Reaction Cards* [6]. The *System Usability Scale (SUS)* offers quantitative assessment by scoring system usability from 0 to 100 based on ten statements [14]. Human-Computer Interaction and Human-Robot Interaction fields have many scales available for evaluating subjective human experiences, such as the NASA Task Load Index for workload assessment, but only few address ethical value relevance, as an overview of measurement scales demonstrates [70].

*2.2.2 User Experience Questionnaire (UEQ).* The UEQ assesses users' subjective perceptions using 26 items across six scales: *Attractiveness* (overall impression), *Efficiency*, *Perspicuity*, and *Dependability* (pragmatic qualities), plus *Stimulation* and *Novelty* (hedonic qualities) [45]. Users rate paired adjectives on a -3 to +3 scale.

The UEQ has achieved widespread adoption, particularly in Europe, with validation in over 30 languages. Since 2017, it has surpassed AttrakDiff in usage [26]. Its success stems from free availability, detailed documentation, and business-ready analysis tools. Many professional UX platforms incorporate UEQ as a key metric, and benchmarks allow comparison with common websites and apps [54].

*2.2.3 Business Goals and Measurement-Driven Design.* UX designs goal is to align business objectives with user goals [43]. Key Performance Indicators (KPIs) measure progress toward business goals, with UX measures serving as KPIs for holistic design performance [72]. However, as measurements easily become targets [79], it's crucial that business measurements incorporate ethical considerations.

Research shows UX professionals often employ 'soft resistance' tactics when advocating for ethical design choices that conflict with business goals [84]. Product managers face challenges negotiating social values in digital product design [47], highlighting the tension between business metrics and ethical design. Product managers face challenges negotiating social values in digital product design [47], highlighting the tension between business metrics and ethical design.

### 2.3 Dark Patterns and Unethical Design

*2.3.1 Defining Dark Patterns.* Dark patterns represent interface design choices that benefit an online service by prioritizing business objectives over user interests, systematically undermining user autonomy, privacy, and wellbeing [36]. These design practices are characterized by their intentional manipulation of user decision-making through exploiting psychological vulnerabilities and cognitive biases. As Mathur [50] demonstrates, dark patterns engineer users' choice architectures by either modifying the information available to users or by modifying the set of available choices—eliminating and suppressing options that disadvantage the manipulator.

Their relationship to traditional user experience metrics remains complex and counterintuitive. Calawen [16] found that dark pattern application had no significant effect on overall user experience measurements, with no significant differences observed in UEQ scales of Perspicuity, Efficiency, Dependability, Stimulation, and Novelty, nor in System Usability Scale scores between manipulative and non-manipulative interface versions. This finding suggests that traditional UX evaluation metrics may be fundamentally unable to detect ethical problems in interface design, as manipulative patterns can maintain or even improve conventional usability metrics while undermining user autonomy.

*2.3.2 User Recognition and Response.* Research shows users can identify dark patterns when asked explicitly. Mildner et al. [56] found that users could distinguish between screenshots containing dark patterns and those without. Bongard-Blanchy et al. [9] found 59% of participants successfully recognized five or more dark patterns among nine interfaces, though some patterns (like confirmshaming) are easier to detect than others (like hidden information).

*2.3.3 Behavioral Effects of Manipulative Design.* Monge Roffarello and De Russis [58] demonstrated behavioral effects through a 3-week experiment removing social investment displays from Facebook feeds, showing reduced time spent on pages. This aligns with research by Baughan et al. [5] demonstrating that social media design features influence dissociation, affecting users' ability to meaningfully engage with content.

### 2.4 The Philosophy and Psychology of Design Influence

*2.4.1 Theoretical Frameworks for Design Influence.* In technology design, ethical considerations present complex challenges. Designers develop choice architecture that inevitably influences user behaviors. Drawing a line often, such as between manipulative and persuasive influence, can be hard [77]. This challenge is complicated by the inherent spectrum of behavioral influence that design can exert on users.

Tromp et al. [80] provide a foundational framework for understanding different types of design influence based on user experience, proposing a classification system built on two critical dimensions: salience (whether the influence is apparent or hidden to the user) and force (whether the influence is strong or weak). Their framework identifies four distinct types of influence: *coercive* (strong and apparent), *persuasive* (weak and apparent), *seductive* (strong and hidden), and *decisive* (weak and hidden). This classification reveals that the ethical implications of design influence depend not only on the intent but also on how users experience and perceive that influence.

*2.4.2 Manipulation vs. Persuasion.* This theoretical framework illuminates why distinguishing between manipulation and persuasion proves so challenging in practice. As Sánchez Chamorro et al. [77] observe, the boundaries between these concepts are often unclear to practitioners, particularly when design influence operates along gradients of salience and force rather than discrete categories. Alavi [3] argue that by categorizing design patterns into binary "dark" versus "light" categories, we oversimplify the user experience as evil versus moral, when the reality is often more nuanced.

Contemporary philosophical analysis reveals that manipulation encompasses three primary forms: *bypassing reason* (influences that operate outside conscious rational deliberation), *trickery* (inducing faulty mental states through deception), and *pressure* (exerting non-coercive force through tactics like emotional blackmail or peer pressure) [60]. Building on this foundation, Susser et al. [76] define online manipulation as "the use of information technology to covertly influence another person's decision-making, by targeting and exploiting their decision-making vulnerabilities." Contemporary design research has developed frameworks for operationalizing these concepts, with Ahuja and Kumar [2] proposing methods for evaluating dark patterns from an autonomy perspective.

*2.4.3 Cognitive and Philosophical Perspectives.* Perhaps most significantly for design professionals, philosophical analysis reveals that manipulation can occur even when the manipulator lacks conscious intent to manipulate—people can "behave manipulatively despite consciously intending not to" [60]. This observation aligns with Sánchez Chamorro et al. [77]'s finding that "designers are not moral philosophers, and, although they have an intuition, we cannot always expect them to make a complete assessment of the ethical aspects of designs."

The challenge of defining ethical boundaries in design is further complicated by the concept of "interpersonal justification" [42], which suggests that design choices must be justifiable to those affected by them. A foundational

challenge emerges from the lack of consensus among UX professionals about core concepts in the field itself. Lallemand et al. [44] conducted an international survey of 758 practitioners, revealing significant differences in how UX is understood and practiced, with findings indicating that "despite many attempts to understand, define and scope UX, one may still wonder whether a consensus has been reached on this concept."

The manipulation of user behavior through dark patterns is well-documented, with studies showing how certain design choices influence user consent practices against their preferences [8].

## 2.5 Professional Practice and Ethical Decision-Making

*2.5.1 Ethical Reasoning Patterns in UX.* Research reveals systematic patterns in how UX professionals approach ethical decision-making that illuminate fundamental challenges in evaluation metrics. Sánchez Chamorro et al. [77] conducted an extensive investigation of ethical tensions in UX design practice, finding that professionals consistently invoke principles of "trust, transparency, and user autonomy" as guiding values, yet often fail to implement these principles due to contextual pressures and organizational constraints.

This research reveals a critical phenomenon of "ethical blindness" where designers create problematic designs unintentionally, not through malicious intent but through systematic limitations in their decision-making processes. The concept of ethical blindness, as defined by Palazzo et al. [62], suggests that "people might behave unethically without being aware of it" due to "a complex interplay between individual sense making activities and context factors." Research on design student behavior provides empirical evidence for this phenomenon, with Chivukula et al. [20] finding that while design students demonstrated sensitivity toward user values, they often contradicted these values through tacit intentions to persuade users in order to achieve stakeholder goals.

*2.5.2 Organizational Challenges and Responsibility.* UX professionals frequently "rely on other roles which hold the 'ethics ownership': business, managerial or legal departments" [77], raising fundamental questions about responsibility attribution in design processes. This tendency toward responsibility diffusion becomes particularly problematic in light of legal scholarship demonstrating that dark patterns constitute "disloyal design"—a form of wrongful self-dealing that takes advantage of design affordances to the detriment of vulnerable users [38].

The research acknowledges that "designers are not moral philosophers, and, although they have an intuition, we cannot always expect them to make a complete assessment of the ethical aspects of designs" [77].

*2.5.3 Cognitive Limitations and Evaluation Tools.* While empirical evidence in Human Computer Interaction research is scarce, Farzandipour et al. [29]'s controlled comparison of Heuristic Evaluation versus Cognitive Walkthrough found that evaluation method choice fundamentally altered evaluator reasoning patterns: HE identified 104 unique problems versus CW's 24, with only 33.3% overlap and completely different severity patterns. HE evaluators focused on effectiveness and satisfaction while CW evaluators emphasized learnability and efficiency—clear evidence that evaluation methods shape professional attention and judgment.

Evidence from adjacent professional domains demonstrates that assessment metrics fundamentally alter reasoning patterns and judgment priorities. Meta-analyses across medicine, education, and clinical psychology consistently demonstrate that structured assessment metrics outperform holistic professional judgment, with studies showing statistical methods superior to clinical intuition across 136 studies [7, 53]. Lens Model Theory explains how assessment metrics act as cognitive filters, systematically weighting available information to create predictable judgment patterns [15]. The psychological foundation for these effects lies in bounded rationality [74], which recognizes that professionals operate

under cognitive limitations and rely on heuristics for decision-making. Recent HCI research has begun examining how cognitive biases affect human-computer interaction [10].

### 2.6 Current Approaches to Ethical Evaluation

*2.6.1 Existing Ethical Assessment Tools.* Although instruments like the Ethical Climate Questionnaire exist to assess organizational ethics [64] and tools like the Facebook Addiction Questionnaire gauge individual platform dependency [28], these only yield data on business organization and individual human long-term effects. More concretely, some researchers have proposed 'moral cards', designed for designers to use as a reflective tool on adhering to ethics by design [25].

The proliferation of ethical design methodologies has become extensive, with comprehensive guides offering hundreds of techniques for becoming more ethically aware and responsible in design practice [21], and systematic surveys documenting 63 existing methods intentionally designed for ethical impact [22]. In artificial intelligence, considerable efforts have been made to incorporate ethics, though a review indicates that these tools investigate ethics *narrowly*, focusing on a few dimensions rather than examining all aspects holistically [65].

A notable attempt to quantify dark patterns specifically is the System Darkness Scale, which provides a targeted 25-item checklist to evaluate whether software employs specific dark pattern elements [81]. While this tool offers valuable insights into dark pattern detection, it focuses specifically on those design elements rather than broader ethical considerations in user interfaces.

*2.6.2 Methodological Challenges.* Efforts to develop tools for evaluating persuasive potential face considerable obstacles. Meschtscherjakov et al. [55] developed the Persuasive Potential Questionnaire (PPQ) to measure persuasive influence across five dimensions, but acknowledge significant challenges and drawbacks in creating generic, context-independent measures of persuasive effects. Their work highlights the fundamental difficulty of standardizing ethical evaluation measures across diverse interface contexts.

Chamorro and Lallemand [18] identify the need for new research approaches to study manipulative design effects, noting limitations in current methods for measuring the impact of deceptive interfaces on users. These methodological challenges are compounded by the inherent ethical implications of evaluation methods themselves, as Williamson and Sundén [83] demonstrate that research methodologies in HCI carry ethical dimensions that may influence design practices in unintended ways.

Some researchers have explored alternative evaluation paradigms that move beyond traditional efficiency-focused metrics. Lu et al. [48] propose focusing on user empowerment measures rather than engagement metrics when evaluating responses to dark patterns, emphasizing awareness-raising and enabling user action against manipulative designs. Similarly, Odom et al. [61] extend slow technology theory to propose evaluation criteria based on meaningful interaction and reflection rather than speed and efficiency.

*2.6.3 The Research Gap.* Despite extensive research on ethical design tools and widespread recognition of dark patterns as problematic, unethical design practices continue to proliferate across digital platforms [13]. This persistence occurs not due to lack of ethical awareness or available tools, but suggests a more fundamental issue: the evaluation metrics that drive design decisions may themselves promote unethical outcomes.

The evidence points to a critical gap in understanding how measurement shapes design practice. While Brignull [11, 12] warned early that metric-driven design could inadvertently encourage manipulative practices, and Thomas and Uminsky [79] demonstrated how "measurements become targets," no empirical research has examined whether

evaluation metrics themselves influence ethical reasoning in design teams. This represents a significant oversight, given extensive evidence from adjacent professional domains that assessment tools fundamentally alter judgment patterns and decision priorities [7, 15, 53].

The research void is particularly striking given that HCI has produced robust empirical research on professional moral reasoning in UX practice [23, 33, 77], yet none of this work examines how the evaluation methodologies themselves shape ethical decision-making. If evaluation metrics act as cognitive filters that systematically weight different considerations [15], then the choice of measurement approach may fundamentally determine whether ethical concerns become salient during design processes.

Our work addresses this gap by providing the first empirical investigation into how different UX evaluation metrics influence ethical reasoning patterns in design practice. Rather than developing new ethical tools, we examine whether the measurement approaches that most directly drive design decisions can be configured to promote more ethical outcomes.

## 3 METHOD

This study employed a multi-stage research design to investigate how autonomy-focused evaluation data influences UX designers' decisions regarding interfaces containing dark patterns. The research comprised three sequential stages: (1) taxonomy development and interface creation, (2) UX+Autonomy Evaluation Study with social media users to establish evaluation benchmarks, and (3) main experiment with professional UX designers.

**Ethics Approval:** The main experiment received IRB approval as exempt research and followed informed consent protocols, participant debriefing, and voluntary participation procedures. The UX+Autonomy Evaluation Study followed university ethics guidance including informed consent and did not require formal IRB approval.

### 3.1 Stage 1: Dark Pattern Taxonomy Development

We developed a comprehensive taxonomy of social media dark patterns through systematic literature review and integration of existing taxonomies. Following established approaches for taxonomy development [36], we combined findings from two peer-reviewed academic papers [57, 58] and the European Data Protection Board regulatory guidelines [27].

Our consolidation process involved: (1) gathering dark pattern types from the three source taxonomies, (2) removing duplicates, (3) merging similar patterns with consistent definitions, and (4) organizing patterns into strategic categories based on manipulation mechanisms. This process yielded 15 distinct low-level dark patterns organized under six high-level strategic categories aligned with Gray et al.'s authoritative framework [37]: *Nagging*, *Obstruction*, *Sneaking*, *Interface Interference*, *Forced Action*, and *Social Engineering*.

The complete taxonomy comprises: **Nagging** (Nagging), **Obstruction** (Overcomplicated Process, Hindering Account Deletion), **Sneaking** (Sneaking Bad Default, Expectation Result Mismatch), **Interface Interference** (False Hierarchy, Trick Wording, Toying with Emotion), **Forced Action** (Forced Access), and **Social Engineering** (Gamification, Social Pressure, Social Connector, Content Customization, Endlessness, Pull to Refresh).

High-fidelity interface mockups were created for each pattern using a consistent minimalistic social media application design framework. Each mockup presented a specific dark pattern within realistic social media contexts, recreated based on examples provided in the source taxonomies where available, ensuring ecological validity while maintaining experimental control across stimuli.

Table 1. Complete Taxonomy of 15 Dark Patterns in Social Media

| High-Level Strategy | Low-Level Pattern | Source Patterns | Definition |
|---|---|---|---|
| Nagging | Nagging | Continuous Prompting [27] | Situations where users are annoyed by repeated attempts to get them to do something they did not intend to |
| Obstruction | Overcomplicated Process | Longer Than Necessary, Lacking Hierarchy, Too Many Options, Privacy Maze, Labyrinth [27, 57] | Processes intentionally designed to be complicated by adding unnecessary steps, too many options, or lack of clear hierarchy |
| | Hindering Account Deletion | Clinging to accounts, Forced Grace Period [57] | Difficulties encountered by users while attempting to (immediately) delete their social media accounts |
| Sneaking | Sneaking Bad Default | Auto Accept Third Party Terms, Deceptive Snugness [27, 57] | Instances where default selections are made by platforms to serve their own interests, possibly without users' awareness |
| | Expectation Result Mismatch | Misleading action, Plain Evil, Dead end, Decontextualising, Inconsistent Interface [27, 57] | Situations where users expect a specific outcome but receive a different result |
| Interface Interference | False Hierarchy | Hidden in Plain Sight, Look over there, Reduced Friction [27, 57] | Designs that form users' choice by enticing attention away from undesired options to platform's preferred ones |
| | Trick Wording | Language discontinuity, Ambiguous wording, Conflicting information [27] | Inappropriate language used to confuse users, leaving them in misunderstanding |
| | Toying with Emotion | Emotional Steering, Persuasive Language [27, 57] | Situations where emotions are used to steer users' decisions in a particular direction |
| Forced Action | Forced Access | Forced Access Granting [57] | Instances where users are forced to give access to something to continue getting to their initial goal |
| Social Engineering | Gamification | Gamification, Social Investment [57, 58] | Situations where rewards are used, pressuring users to continue using the platform to not lose achieved progress |
| | Social Pressure | Fear Of Missing Out, Regression Toward The Mean [57] | Instances where users are pressured to visit the platform to avoid missing out on important topics |
| | Social Connector | Social Connector [57] | Situations where users are prompted to provide information about friends or invite them to expand their network |
| | Content Customization | False Content Customisation, Recommendations [57, 58] | Customizing content to manipulate users through falsified or overly precise customization |
| | Endlessness | Infinite Scrolling, Autoplay [57, 58] | Endlessly and automatically loading content, eliminating need for user decision-making |
| | Pull To Refresh | Pull To Refresh [57, 58] | Functionality that requires users to pull on the interface to reload the page |

### 3.2 Stage 2: UX+Autonomy Evaluation Study - User Evaluation of Dark Patterns

*3.2.1 Participants.* A total of 126 social media users participated in the UX+Autonomy Evaluation Study to establish baseline user perceptions of the dark pattern interfaces. Participants were recruited through convenience sampling via social media platforms (Facebook groups), survey exchange platforms (SurveyCircle), and snowball sampling within research networks. The sample comprised 53 females (42%) and 73 males (58%), with ages ranging from 19 to 80 years. Seventy-five percent resided in Austria, with additional participants from the United States (9%), Switzerland and Germany (5%), and other locations (11%). Eighty-seven percent considered themselves at least somewhat technically savvy (56% responding "Yes" and 31% responding "A bit" to technical savviness).

*3.2.2 Instrumentation.* We adapted the User Experience Questionnaire (UEQ) to include ethical evaluation dimensions through an iterative development process. The adaptation involved: (1) reviewing all 26 UEQ adjective pairs for applicability to dark pattern evaluation, (2) identifying items likely to elicit strong participant responses, and (3) adding four autonomy-focused items addressing manipulation techniques.

The autonomy-focused items were developed through a systematic process involving autonomy literature review and expert consultation. Three dark pattern experts with backgrounds in HCI ethics, manipulation theory, and user autonomy research collaboratively developed contrasting autonomy elements based on established frameworks for user agency and freedom from manipulation. The expert team reviewed autonomy literature to identify key dimensions of user control, transparency, and freedom from coercive influence, drawing particularly on Tromp et al.'s framework of design influence salience and force [80], Noggle's philosophical analysis of autonomy-preserving versus autonomy-undermining influence [60], Susser et al.'s definition of online manipulation through covert decision-making exploitation [76], and Ahuja et al.'s methods for evaluating dark patterns from an autonomy perspective [2]. These theoretical constructs were translated into semantic differential pairs suitable for the UEQ format.

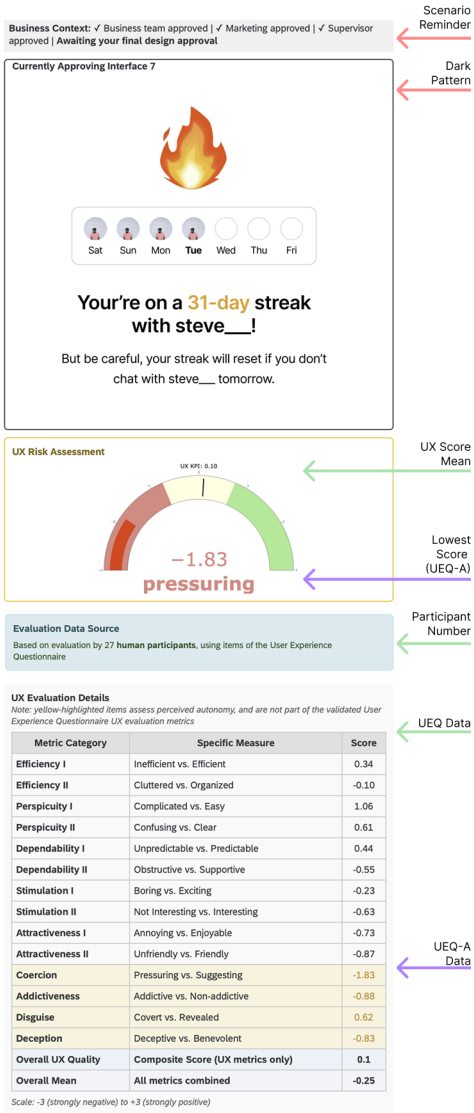



Fig. 2. Example stimuli showing the three experimental conditions. • Baseline condition shows the interface and business context presented to all participants. • UEQ condition shows additional UEQ evaluation data including gauge with UX score mean and detailed UEQ evaluation data with participant numbers. • UEQ-A condition shows additional UEQ-A autonomy data with the lowest autonomy score highlighted.

The final instrument comprised 14 items: 10 items forming 5 two-item scales from the UEQ-S (UEQ short version) excluding Novelty—specifically Perspicuity, Efficiency, Dependability, Stimulation, and a two-item Attractiveness scale—plus 4 standalone manipulation-focused items: *pressuring / suggesting*, *addictive / non-addictive*, *covert / transparent*, and *deceptive / benevolent*.

*3.2.3 Procedure.* Each enduser participant evaluated 5 randomly selected dark pattern interfaces from the complete set of 15 patterns using 7-point semantic differential scales (-3 to +3) consistent with standard UEQ methodology. To prevent survey fatigue, participants saw only 5 of the 15 patterns rather than the complete set. A "don't know/not applicable" option was provided for each item, and item order was randomized with balanced positive/negative term presentation.

This UX+Autonomy Evaluation Study generated the user evaluation data that formed the basis for the experimental manipulations in the main study, ensuring that evaluation summaries presented to designers were grounded in actual user perceptions rather than researcher-generated content. Key findings showed that dark patterns were perceived as pressuring, annoying, and unfriendly, with *Forced Access* rated worst among all patterns.

### 3.3 Stage 3: Main Experiment - Designer Decision-Making

*3.3.1 Participants.* A total of 141 UX/design professionals participated in the main three-condition experiment following a power analysis indicating a target recruitment of 140 participants. Participants were randomly assigned to one of three between-subjects conditions: UEQ evaluation data only ($n$ = 49), UEQ + Autonomy-focused evaluation data ($n$ = 47), or No evaluation data baseline ($n$ = 45).

**Recruitment and Compensation:** Participants were recruited through Prolific Academic with pre-screening requirements including: (1) professional experience in UI/UX design, product design, or design decision-making roles, (2) minimum 1 year of professional design experience, and (3) fluency in English. Additional Prolific pre-screening limited recruitment to participants indicating knowledge of software development techniques, responsive design, UI design, A/B testing, and UX methodologies. Participants were compensated based on Prolific's $12/hour guideline for an estimated 20-minute study, but median completion time was 30 minutes, resulting in an average compensation rate of $8.06/hour.

*3.3.2 Design and Procedure.* Participants were randomly assigned to one of three between-subjects conditions manipulating the type of evaluation information presented alongside interface designs:

- **UI condition (baseline)**: Participants evaluated interfaces without any user feedback information, providing a control condition for intrinsic design assessment.
- **UEQ condition**: Participants viewed standardized user evaluation summaries based on the User Experience Questionnaire [45], focusing on pragmatic and hedonic quality dimensions derived from the Stage 2 UX+Autonomy Evaluation Study.
- **UEQ-A condition**: Participants viewed the same UEQ summaries plus additional autonomy-focused evaluation data emphasizing perceived coercion, perceived addiction, perceived disguise, and perceived deception also derived from the Stage 2 UX+Autonomy Evaluation Study.

**Evaluation Data Presentation:** In the UEQ and UEQ-A conditions, evaluation data was presented as realistic user feedback summaries alongside each interface. The display included: (1) an overall mean UEQ score gauge showing the aggregate rating across all UEQ-S dimensions, (2) a detailed breakdown table showing individual subscale scores

for Perspicuity, Efficiency, Dependability, Stimulation, and Attractiveness, and (3) in the UEQ-A condition only, an additional section highlighting the lowest-ranking autonomy-related item from among the four manipulation-focused measures (*pressuring*, *addictive*, *covert*, *deceptive*). Since the autonomy items were standalone measures rather than forming an aggregate scale, we presented the most problematic individual item to maximize ecological validity while avoiding misleading aggregation across conceptually distinct autonomy dimensions.

The evaluation data was presented in a professional business dashboard format including: (1) a business approval context header to simulate realistic organizational decision-making scenarios for a social media startup, (2) a "UX Risk Assessment" section containing the gauge visualization showing the overall UEQ mean score with color coding (red for negative, yellow for neutral, green for positive ratings), and (3) a detailed data table presenting individual metric scores on the -3 to +3 scale consistent with standard UEQ methodology. All evaluation data was derived from the Stage 2 UX+Autonomy Evaluation Study ensuring ecological validity, with sample sizes clearly indicated (ranging from 23-47 participants per interface depending on the randomization scheme used in the UX+Autonomy Evaluation Study).

Each participant evaluated 10 interfaces randomly selected from the set of 15 dark pattern interfaces developed in Stage 1. The interfaces were presented within a realistic business approval context to simulate authentic decision-making scenarios in a social media startup environment.

For each interface, participants completed two primary dependent measures and two secondary measures providing contextual insight. **Primary dependent measures:** (1) release tendency ("How likely would you be to release this interface design?") rated on an 8-point Likert scale (0 = *definitely would not release* to 7 = *definitely would release*), and (2) a binary release decision ("Would you release this interface design?" yes or no). **Secondary measures:** (3) a written explanation ("Please explain your decision") providing qualitative insight into reasoning patterns, and (4) decision confidence ("How confident are you in the decision you just made?") rated on an 8-point scale (0 = *completely uncertain* to 7 = *completely certain*).

*3.3.3 Participant Characteristics.* All participants reported professional experience in UI/UX design, product design, or design decision-making roles meeting Prolific's pre-screening requirements. Following completion of the main experimental task, participants provided demographic information about their professional backgrounds. Participants represented diverse professional backgrounds, with UX/UI Designers (52.0%) being the most common role, followed by Design Managers and Product Managers (12.0% each), UX Researchers (7.0%), Product Designers (6.0%), and Design Directors (3.0%). The remaining 8.0% reported other design-related roles. Professional experience exceeded minimum requirements with substantial representation across experience levels.

**Design Decision Authority:** Participants demonstrated substantial decision-making authority within their organizations, with 75.0% holding either final decision authority (29.0%) or significant influence (46.0%) over interface design decisions in their current roles. Only 20.0% reported some input, 4.0% reported little input, and 1.0% reported no decision authority.

**Dark Pattern Domain Knowledge:** The sample demonstrated strong familiarity with dark pattern concepts, with 77.0% reporting some level of familiarity (38.0% very familiar, 39.0% somewhat familiar), 19.0% slightly familiar, and only 5.0% unfamiliar. This domain knowledge distribution indicates appropriate professional qualification for the experimental task.

**Data Quality and Final Sample:** Eleven participants were excluded from analysis due to suspicious response patterns identified through both automated screening and manual review of professional explanations. This quality

assurance process resulted in a final analytical sample of 141 participants with verified professional qualifications and engaged participation.

### 3.4 Statistical Analysis

We conducted statistical analysis at two levels to capture both participant-level patterns and per-evaluation effects:

**Participant-level analysis:** For rejection rates, we calculated mean rejection rates for each participant across their evaluated interfaces and conducted one-way ANOVA with planned contrasts and Tukey's HSD post-hoc tests. This approach addresses the nested structure while meeting normality assumptions for binary outcome aggregation.

**Per-evaluation mixed effects analysis:** For release tendency, we employed mixed effects modeling to analyze individual evaluation responses while accounting for participant and interface random effects. This approach treats each interface evaluation as an independent observation while controlling for non-independence due to repeated measures within participants and across interfaces. Mixed effects models used maximum likelihood estimation with condition as fixed effect and participant and interface as random effects.

**Interface-level analysis:** Individual interface effects were examined using planned contrasts between UEQ and UEQ-A conditions with false discovery rate (FDR) correction for multiple comparisons across the 15 interfaces. This focused comparison was theoretically motivated by our specific hypothesis that autonomy-focused evaluation metrics would enhance designers' sensitivity to manipulative design elements compared to standard usability metrics, making the UEQ vs UEQ-A contrast the most theoretically relevant for understanding evaluation framework effects on dark pattern detection.

**Confidence analysis:** Decision confidence was analyzed using mixed effects modeling parallel to the tendency analysis, with participant and interface random effects.

All analyses were conducted in R (version 4.4.2) with significance set at $\alpha = .05$.

### 3.5 Text Analysis of Professional Explanations

To analyze the justification patterns underlying design decisions, we employed computational text analysis combining initial LDA topic modeling with hypothesis-driven theme refinement. We developed eight justification categories: Manipulation Awareness, Responsibility Avoidance, Ethics-Focused Reasoning, Aesthetic-Focused Reasoning, Business-Focused Reasoning, Emotional Intensity, Conformity Justification, and Interface Design Elements. For each justification category, we performed binary classification of explanations (justification present/absent) and conducted chi-square tests with Bonferroni correction ($\alpha = 0.00625$) and Cramér's V effect sizes. Complete methodological details, keyword lists, and validation procedures will be available in the OSF repository [4].

### 3.6 Materials and Stimuli

*3.6.1 Dark Pattern Interface Development.* The study used 15 social media dark patterns identified through systematic literature review combining peer-reviewed papers and regulatory documents. These patterns were categorized under six distinct strategies: nagging, forced action, obstruction, interface interference, sneaking, and social engineering. Each pattern was recreated as a high-fidelity visual mockup in a minimalistic social media app design framework, ensuring consistency while highlighting specific manipulative mechanisms.

Interface mockups were designed to represent realistic social media contexts including user profiles, feeds, settings pages, and notification systems. Each interface prominently featured the target dark pattern while maintaining visual consistency through a unified design system including consistent typography, color palette, iconography, and layout

structure. This approach ensured that differences in participant responses could be attributed to the specific dark pattern mechanisms rather than confounding visual design differences.

*3.6.2 Evaluation Data Generation.* The evaluation data presented to participants in the UEQ and UEQ-A conditions was derived from the Stage 2 UX+Autonomy Evaluation Study with 126 social media users who evaluated these same interfaces using an extended User Experience Questionnaire. Standard UEQ summaries focused on traditional usability and hedonic quality metrics, while UEQ-A summaries additionally highlighted autonomy-related concerns including user control, transparency, and freedom from manipulation. This data formed the basis for both experimental conditions, ensuring ecological validity of the evaluation information.

In the UEQ condition, participants viewed standardized evaluation summaries presenting six core UEQ dimensions (Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty) with mean scores and confidence intervals. The UEQ+Autonomy condition presented identical UEQ data plus four additional autonomy-focused dimensions (pressuring vs. suggesting, addictive vs. non-addictive, covert vs. revealed, deceptive vs. benevolent) highlighted to emphasize user control and freedom from manipulation. All evaluation data was presented in consistent table format with scores normalized on a 7-point scale, ensuring comparable information density across conditions while manipulating the ethical salience of the feedback.

## 4 RESULTS

### 4.1 UX+Autonomy Evaluation Study: Data Generation

The UX+Autonomy Evaluation Study with 126 social media users successfully generated the evaluation dataset used in the main experiment. As detailed in Appendix Table 2, the 15 dark pattern interfaces showed varied user perceptions across both standard UEQ dimensions and autonomy-focused measures. Forced Access received the most negative overall evaluation (Mean = -0.90), while Pull to Refresh was rated most positively (Mean = 0.49). The autonomy-focused items effectively differentiated manipulative patterns, with Nagging, False Hierarchy, and Gamification showing the strongest negative responses on pressuring vs. suggesting measures. This UX+Autonomy Evaluation Study provided ecologically valid evaluation summaries for the subsequent designer decision-making experiment. More detailed analysis of the UX+Autonomy Evaluation Study can be found in our workshop paper (ANONYMIZED). Complete datasets will be made available through the Open Science Framework upon publication, and descriptive statistics are provided in the appendix.

### 4.2 Main Study: Effects of Evaluation Condition on Designer Decisions

Analysis revealed significant differences between evaluation conditions for both dependent variables, confirming our hypotheses about the influence of evaluation frameworks on design decision-making.

*4.2.1 Release Tendency.* Mixed effects analysis accounting for participant and interface random effects revealed significant differences in release tendency between evaluation conditions, $\chi^2(2) = 32.65$, $p < .001$. This analysis treats each interface evaluation as an independent observation (N = 1,410 evaluations from 141 participants across 15 interfaces) while controlling for the nested structure of multiple evaluations per participant and per interface.

Post-hoc comparisons using estimated marginal means with Tukey adjustment confirmed our hypothesized pattern: participants in the UI (no evaluation data) condition showed significantly higher release tendency ($M = 4.66$, $SE = 0.31$) compared to both the UEQ condition ($M = 3.81$, $SE = 0.31$), $p = .006$, $d = 0.37$, and the UEQ-A condition ($M = 3.12$, $SE =$

0.31), $p < .001$, $d = 0.67$. Additionally, the UEQ condition showed significantly higher release tendency than the UEQ-A condition, $p = .026$, $d = 0.29$.

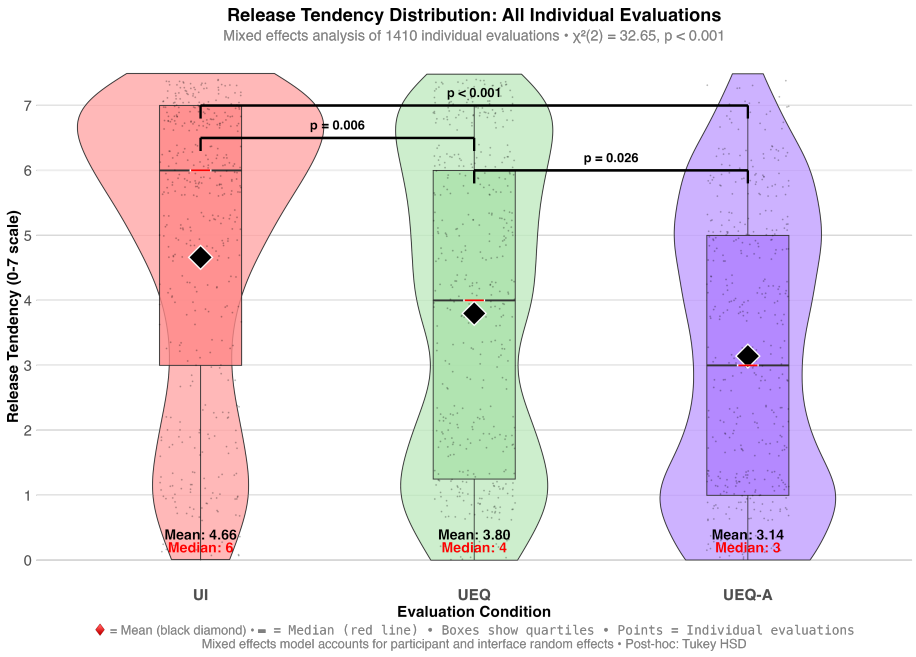



Fig. 3. Release tendency by evaluation condition. Participants rated their tendency to release each interface on a 0-7 scale (0 = *definitely would not release*, 7 = *definitely would release*). Participants showed systematically different judgments depending on the evaluation framework provided, with autonomy-focused evaluation data leading to significantly more critical assessments. Error bars represent 95% confidence intervals.

*4.2.2 Rejection Rates.* Participant-level analysis of rejection rates revealed the hypothesized reversed pattern. One-way ANOVA of participant mean rejection rates showed significant between-group differences, $F(2, 138) = 15.97$, $p < .001$, $\eta^2 = .188$. Post-hoc comparisons using Tukey's HSD confirmed significant differences between all conditions: UEQ-A participants rejected significantly more interfaces ($M = 56.2\%$, $SD = 21.3\%$) compared to UEQ participants ($M = 43.9\%$, $SD = 19.9\%$), $p = .021$, $d = 0.60$, and UI participants ($M = 30.0\%$, $SD = 25.3\%$), $p < .001$, $d = 1.13$. UEQ participants also rejected significantly more interfaces than UI participants, $p = .008$, $d = 0.62$.

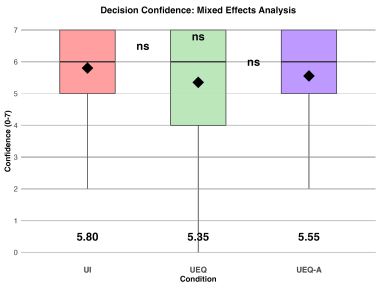



Fig. 4. Decision confidence remained consistently high across all conditions (0 = *not at all confident*, 7 = *extremely confident*), despite significant behavioral differences in release decisions.

*4.2.3 Decision Confidence.* Analysis of participants' confidence in their release decisions revealed a striking pattern: despite the significant behavioral differences observed in release tendency and rejection rates, confidence ratings remained consistently high across all conditions with no significant differences. Mixed effects analysis accounting for participant and interface random effects found no significant condition effect, $\chi^2(2) = 4.26$, $p = .119$.

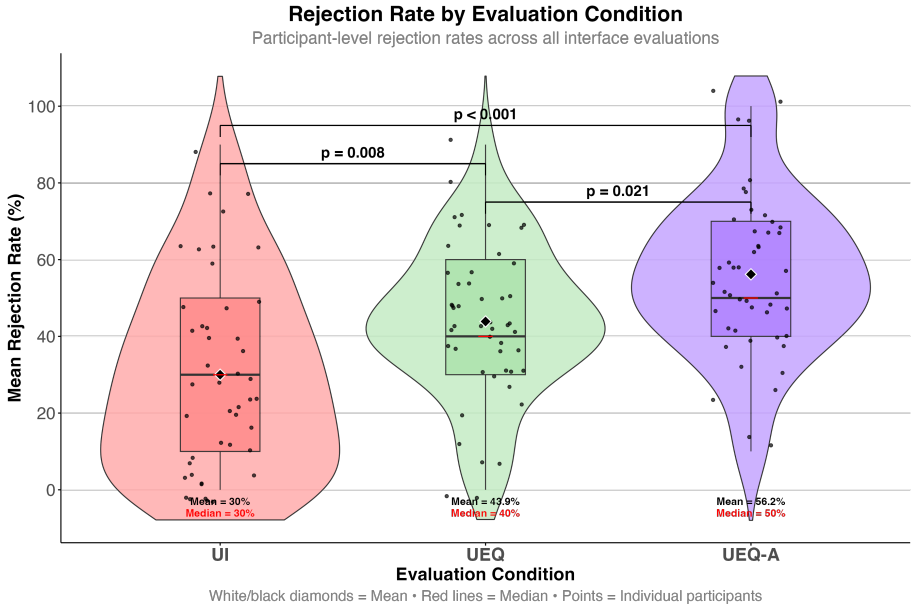



Fig. 5. Rejection rates by evaluation condition. These rates represent the percentage of interfaces each participant rejected across their 10 evaluated interfaces, based on their binary yes/no release decisions. Participants who received autonomy-focused evaluation data rejected significantly more interfaces containing dark patterns. Error bars represent 95% confidence intervals.

Mean confidence ratings on the 0-7 scale (0 = *completely uncertain*, 7 = *completely certain*) regarding their binary release decisions were: No Evaluation Data condition ($M$ = 5.80, $SD$ = 0.16), UEQ condition ($M$ = 5.35, $SD$ = 0.15), and UEQ+Autonomy condition ($M$ = 5.55, $SD$ = 0.15). All post-hoc comparisons were non-significant (all $p > .10$), with median confidence ratings of 6 across all conditions, indicating high confidence regardless of the evaluation information provided.

This pattern suggests systematic overconfidence in design decision-making: participants remained equally confident in judgments that were demonstrably different when provided with additional evaluation information, revealing a disconnect between decision confidence and decision quality.

### 4.3 Justification Pattern Analysis

Computational text analysis of 1,313 professional explanations revealed differences in justification patterns across evaluation frameworks. Seven of eight justification categories showed condition differences after Bonferroni correction ($p < 0.00625$), with effect sizes ranging from small to medium (Cramér's V: 0.107 to 0.299). These justification categories were informed by initial LDA topic modeling, which identified foundational categories including business approval processes, visual design aesthetics, and usability considerations, with subsequent hypothesis-driven refinement to capture nuanced ethical justification patterns.

*4.3.1 Manipulation Awareness.* The UEQ-A condition increased manipulation awareness ($\chi^2$ = 117.1, $p < 0.001$, V = 0.299). Manipulation awareness was present across all conditions: UEQ-A (77.7%), UI (20.7%), and UEQ (1.7%).

Examples from UI condition included direct statements like "too pushy" and "it is very manipulative," with participants identifying "emotionally coercive language" and designs that "feels off" with a "warning tone."

UEQ-A participants frequently referenced specific metrics: "coercion score is extremely poor (-1.83), suggesting users feel pressured" and identifying designs as "pressuring or manipulative, potentially harming users."

*4.3.2 Responsibility Avoidance.* The UI condition promoted responsibility avoidance patterns ($\chi^2$ = 42.6, p < 0.001, V = 0.180), with 70.4% of responsibility avoidance instances occurring in UI compared to 18.5% in UEQ-A and 11.1% in UEQ.

UI participants frequently cited organizational approval: "business team, marketing department, and my supervisor have all approved," "already approved by all other parties," and "all key teams" approval. They expressed concerns about "professional and business risk," "competitive market" with "limited runway," and emphasized "momentum," "resources and growth," and "development integration" already in progress.

*4.3.3 Ethics-Focused Reasoning.* The UEQ-A condition showed increased ethics-focused reasoning ($\chi^2$ = 23.3, p < 0.001, V = 0.133), with 54.9% occurring in UEQ-A compared to 30.8% in UI and 14.3% in UEQ.

UEQ-A participants emphasized "user well-being and safety outweigh business considerations," "violating fundamental principles of ethical user interface design," "undermining autonomy," and "harmful to user trust." They referenced "ethical design principles," "user autonomy," "protecting users," and concerns about "manipulative" designs that would "destroy user trust."

*4.3.4 Aesthetic-Focused Reasoning.* Aesthetic-focused reasoning showed condition differences ($\chi^2$ = 27.2, p < 0.001, V = 0.144), with highest prevalence in UI (42.1%) compared to UEQ-A (29.7%) and UEQ (28.2%). UI participants referenced judgments like "clean design," "visually appealing," "functional," "polished," "attractive," "beautiful interface," "minimalistic," "vibrant," and "appealing colors."

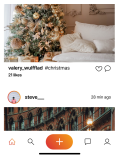
*4.3.5 Business-Focused Reasoning.* Business-focused reasoning showed condition differences ($\chi^2$ = 20.8, p < 0.001, V = 0.126), with highest prevalence in UI (45.3% vs. UEQ 29.1% vs. UEQ-A 25.6%). UI participants emphasized "competitive market" pressures, "revenue potential," "strategic goals," "business goals," "monetization," "user retention," "growth," "market capture," and "business objectives."

*4.3.6 Emotional Intensity.* Emotional intensity showed condition differences ($\chi^2$ = 15.1, p < 0.001, V = 0.107), with highest prevalence in UEQ-A (47.9%) compared to UEQ (29.0%) and UI (23.1%). UEQ-A participants used stronger emotional language including "extremely poor," "awful," "terrible," "highly unsuitable," and "definitely." UI participants used terms like "I HATE," "unacceptable," "amazing," and "love it."

*4.3.7 Interface Design Elements.* Interface design elements showed condition differences ($\chi^2$ = 17.3, p < 0.001, V = 0.115), with highest prevalence in UI (39.1%) compared to UEQ-A (32.0%) and UEQ (28.9%). Participants across conditions referenced technical components including "efficient," "organized," "ease of use," "usability," "clear," "confusing," and "navigation."

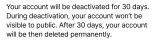
*4.3.8 Conformity Justification.* Conformity justification showed differences across conditions ($\chi^2$ = 7.8, p = 0.020, V = 0.077) but did not reach significance after Bonferroni correction. UI participants (51.6%) most frequently referenced "standard," "familiar," "user expectations," "already in place," "other social media," and "platform-level" conventions.

**Content Customization**
UEQ Mean: 0.09
Lowest UEQ-A: pressuring
Score: -0.71
**d = 1.02**

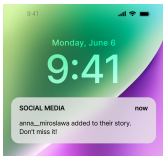


**Endlessness**
UEQ Mean: 0.01
Lowest UEQ-A: addictive
Score: -1.47
**d = 0.73**


**Trick Wording**
UEQ Mean: -0.70
Lowest UEQ-A: annoying
Score: -1.85
**d = 0.70**




**Account Deletion**
UEQ Mean: 0.02
Lowest UEQ-A: deceptive
Score: -0.76
**d = 0.64**




**Social Pressure**
UEQ Mean: 0.03
Lowest UEQ-A: pressuring
Score: -1.02
**d = 0.61**


**Pull to Refresh**
UEQ Mean: 0.49
Lowest UEQ-A: addictive
Score: -1.00
**d = 0.55**

Fig. 6. The six interfaces showing the strongest sensitivity to autonomy-focused evaluation frameworks (UEQ vs. UEQ+Autonomy differences, FDR corrected, ordered by effect size). Each interface shows the pattern name, overall UEQ mean score, the lowest-ranking autonomy attribute from the UEQ-A condition, and the Cohen's d effect size for the UEQ vs. UEQ-A comparison. Full evaluation data from the foundation study (n=126) is detailed in Appendix Table 1 and was available to participants in their respective experimental conditions. Note: Content Customization, Endlessness, and Pull to Refresh interfaces were presented as short animations in the foundation study; static images are shown here for publication.

### 4.4 Individual Interface Effects

To examine which specific dark patterns were most sensitive to evaluation condition effects, we conducted interface-by-interface analyses using planned contrasts between UEQ and UEQ+Autonomy conditions. This focused comparison was theoretically motivated by our hypothesis that autonomy-focused evaluation metrics would specifically enhance designers' sensitivity to manipulative design elements compared to standard usability metrics, rather than comparing all possible condition pairs.

Individual interface analysis revealed that 6 of 15 interfaces showed significant differences between UEQ and UEQ+Autonomy conditions after FDR correction ($q < .05$). The largest effects were observed for:

- **Content Customization**: $d = 1.02$, $p_{FDR} = .002$
- **Endlessness**: $d = 0.73$, $p_{FDR} = .009$
- **Trick Wording**: $d = 0.70$, $p_{FDR} = .017$
- **Hindering Account Deletion**: $d = 0.64$, $p_{FDR} = .018$
- **Pull to Refresh**: $d = 0.55$, $p_{FDR} = .040$
- **Social Pressure**: $d = 0.61$, $p_{FDR} = .040$

Omnibus ANOVA across all three conditions identified 9 interfaces with differences. Detailed violin plots showing the distribution of release tendency ratings for each interface are provided in Appendix Figure A1.

### 4.5 Autonomy-Focused Evaluation Metrics Increase Designer Sensitivity to Dark Patterns

Substantial effects were observed with a 26.2 percentage point difference in rejection rates between No Evaluation Data and UEQ+Autonomy conditions (30.0% vs. 56.2%) and the 1.52-point difference in release tendency on the 0-7 scale. Effect sizes for the primary comparisons were large according to Cohen's conventions: No Evaluation vs. UEQ+Autonomy showed $d = 1.20$ for tendency and $d = 1.13$ for rejection rates.

The results demonstrate systematic effects of evaluation frameworks on designer decision-making. The UX+Autonomy Evaluation Study successfully generated differentiated evaluation data across 15 dark pattern interfaces. In the main experiment, participants showed significantly different release tendencies and rejection rates depending on the evaluation

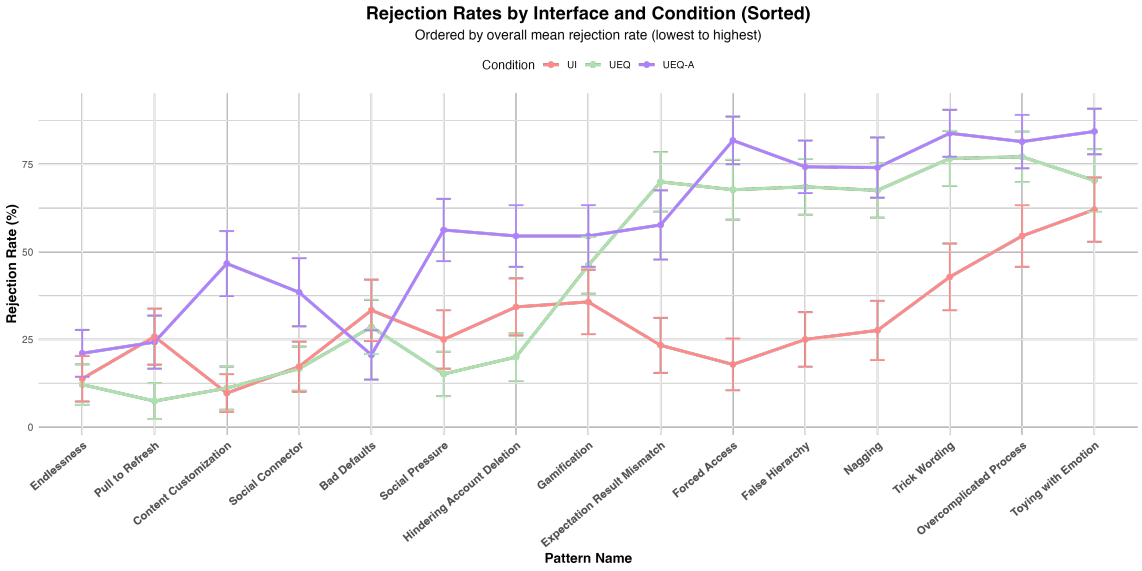



Fig. 7. Rejection rate trends by interface, sorted by effect magnitude. Interfaces are ordered from largest to smallest difference between No Evaluation Data and UEQ+Autonomy conditions, showing the percentage of participants who answered "no" to releasing each interface in their binary release decisions.

information provided, with autonomy-focused metrics leading to more critical assessments. Despite these behavioral differences, decision confidence remained consistently high across conditions, suggesting systematic overconfidence in design decision-making.

Computational text analysis revealed seven of eight justification categories showed condition differences after Bonferroni correction, with effect sizes ranging from small to medium (Cramér's V: 0.107 to 0.299). The progression from no evaluation data to standard usability evaluation to autonomy-focused evaluation corresponded to increasingly critical assessment of interfaces containing dark patterns. Individual interface analysis identified six interfaces with significant sensitivity to autonomy-focused evaluation metrics, with effect sizes ranging from medium to large ($d$ = 0.55 to 1.02).

## 5 DISCUSSION

This study provides evidence that evaluation metrics actively shape design, by triggering ethical sensibility. Through a controlled experiment with 141 UX professionals, we demonstrated that the type of evaluation data provided to designers systematically influences their willingness to approve interfaces containing dark patterns, with autonomy-focused metrics nearly doubling rejection rates compared to no evaluation data (56.2% vs. 30.0%).

### 5.1 Evaluation Metrics as Design Interventions

Our results challenge the assumption that evaluation tools are neutral measurement instruments. Instead, they reveal evaluation metrics as active interventions that shape what designers notice, consider, and ultimately approve. The

stepwise progression in rejection rates—from 30.0% (no data) to 43.9% (standard UX metrics) to 56.2% (autonomy-focused metrics)—demonstrates that providing any user feedback increases ethical scrutiny, but the specific dimensions emphasized in that feedback fundamentally alter design decisions.

This finding aligns with literature on measurement's transformative effects [79], but extends it to show how evaluation metrics specifically trigger ethical sensibility. Standard UX metrics like the UEQ, while not explicitly ethical, still increased critical assessment of dark patterns compared to pure UI evaluation. However, autonomy-focused metrics produced the strongest effects, suggesting that making ethical considerations salient through evaluation metrics can effectively guide designers toward more ethical decisions.

The practical implications are significant: organizations cannot treat evaluation metrics as passive tools that simply capture designer preferences. Instead, the choice of evaluation metrics becomes a design intervention itself—one that either supports or undermines ethical design goals.

### 5.2 Triggering Latent Ethical Sensitivity

The qualitative analysis of 1,313 professional explanations reveals that evaluation frameworks don't just change what designers decide, but fundamentally reshape how they think about design problems. When provided with autonomy-focused evaluation data, 77.7% of participants demonstrated manipulation awareness compared to only 20.7% in the no-data condition. This represents a nearly 4-fold increase in explicit ethical reasoning.

This transformation suggests that UX professionals possess latent ethical sensitivity that can be activated by appropriate evaluation tools. Rather than requiring extensive ethics training or external oversight, our findings indicate that designers are receptive to ethical considerations when those considerations are made salient through their existing evaluation workflows.

The shift from responsibility avoidance (most common in no-data conditions) to manipulation awareness (dominant with autonomy metrics) indicates that evaluation frameworks influence not just ethical conclusions but the entire ethical reasoning process. This suggests that incorporating ethical dimensions into evaluation tools could be a scalable approach to promoting ethical design practice.

### 5.3 Designer Overconfidence and Bias Blindness

Our confidence findings reveal a notable disconnect between decision behavior and decision certainty that has important implications for UX evaluation practice. Despite significant behavioral differences in release decisions across evaluation conditions—with participants being considerably more conservative when provided with UEQ metrics and autonomy risk information—participants maintained consistently high confidence in their judgments regardless of the evaluation framework provided.

This pattern suggests that **designers systematically overestimate their rational decision-making capabilities**. When presented with interface screenshots alone, participants were highly confident in decisions that were demonstrably different from those they would make with more comprehensive evaluation information. This confidence did not decrease when additional UEQ metrics revealed potential usability issues, nor when autonomy risk assessments highlighted ethical concerns that further influenced their decisions.

The text analysis reveals that evaluation frameworks do not merely influence what designers decide, but fundamentally reshape how they think about ethical considerations. The shift from responsibility avoidance to manipulation awareness when autonomy-focused data is present suggests that designers possess latent ethical sensitivity that can be activated

by appropriate evaluation tools. This implies that designers may be unaware of how evaluation frameworks shape their moral reasoning and blind to their own biases.

The overconfidence effect has several critical implications for UX practice. First, designers may develop false certainty about release decisions based on limited information, potentially leading to premature product releases or inadequate evaluation processes. Second, high confidence in initial assessments may create resistance to conducting more comprehensive UX evaluations, as designers may feel their initial judgments are sufficient.

### 5.4 Interface-Specific Vulnerability to Ethical Evaluation

The interface-by-interface analysis revealed that different dark patterns show varying sensitivity to evaluation framework effects. Content customization, endlessness, and trick wording showed the largest differences between standard UX and autonomy-focused evaluation, suggesting these patterns are particularly vulnerable to ethical scrutiny when autonomy considerations are made salient.

This heterogeneity in pattern responses has practical implications for evaluation strategy. Some dark patterns may be effectively identified through standard UX metrics alone, while others require explicit ethical evaluation dimensions to trigger appropriate designer concern. Understanding which patterns require which types of evaluation could help organizations prioritize their evaluation investments and develop targeted assessment protocols.

The finding that 6 of 15 interfaces showed significant autonomy-focused effects after correction suggests that roughly 40% of dark patterns may be specifically vulnerable to autonomy-based evaluation approaches. This specificity indicates that different ethical evaluation approaches may be needed for different categories of manipulative design. Beyond manipulation concerns, future work could explore other user-centric ethical dimensions (e.g., privacy, fairness, security) to see if different patterns respond to different ethical lenses.

### 5.5 Practical Implications for Design Organizations

These findings have immediate practical implications for design organizations seeking to improve ethical outcomes. Rather than relying on designer training or external oversight, organizations can incorporate ethical considerations directly into their existing evaluation workflows through autonomy-focused metrics.

The stepwise progression in critical assessment (30% $\rightarrow$ 44% $\rightarrow$ 56% rejection rates) suggests that even standard UX evaluation provides some ethical benefit compared to pure aesthetic judgment, but autonomy-focused evaluation provides the strongest ethical guidance. Organizations could implement tiered evaluation approaches, using standard UX metrics as an initial screen and autonomy-focused evaluation for interfaces that raise initial concerns.

The high confidence across conditions also suggests that designers may not recognize when they need additional evaluation information. Structured evaluation protocols that require consideration of multiple evaluation dimensions before release decisions could help mitigate overconfidence effects and ensure more comprehensive ethical assessment.

*5.5.1 Metrics vs. Guidelines: The Power of Empirical Weight.* Our approach differs fundamentally from traditional design guidelines or ethical frameworks. While design guidelines typically offer prescriptive rules (e.g., 'avoid deceptive language' or 'provide clear navigation'), metrics carry empirical weight derived from actual user responses to specific design features. This distinction is critical: when evaluation data shows that users rate an interface as 'pressuring' (-1.4 on autonomy scales) or 'covert' rather than 'transparent,' it provides concrete evidence rather than abstract principles.

Our approach also differs from existing dark pattern measurement tools such as the System Darkness Scale (SDS) [81], which focuses on directly assessing the 'darkness' of interfaces through user perceptions of manipulative design

elements. While such approaches provide valuable insights into dark pattern detection, our framework takes an alternative approach by emphasizing positive autonomy values rather than focusing on negative assessments of manipulation. This distinction reflects a different methodological choice: rather than asking users to identify what is problematic with an interface, autonomy-focused metrics assess dimensions that support user agency and freedom of choice.

Additionally, it integrates with existing UX evaluation workflows that typically emphasize positive user experience outcomes. The approach also provides guidance on what to optimize for (autonomy, transparency, user control), rather than identifying what to avoid. Our autonomy-focused approach can be integrated with standard evaluation methods like the UEQ, allowing deployment within existing organizational assessment processes.

Metrics translate user experiences into quantified insights that organizations can systematically track, compare, and optimize against. Unlike guidelines that require interpretation and can be dismissed as subjective preferences, autonomy-focused metrics provide the same objective, data-driven foundation that organizations already value in business metrics. This empirical grounding makes ethical considerations harder to ignore or rationalize away during design decision-making.

*5.5.2 Organizational Constraints and Implementation Challenges.* We acknowledge that implementing ethics-focused evaluation frameworks faces significant organizational constraints. Companies may resist incorporating autonomy-focused metrics when such considerations conflict with business objectives focused on engagement, retention, or conversion. Some organizations may view comprehensive ethical evaluation as unnecessary overhead, believing that general ethical sensibility or existing compliance frameworks provide sufficient protection.

Our recommendation for achieving ethical design through evaluation metrics recognizes these realities. We do not suggest that metrics alone can solve all ethical design challenges, nor that organizations will uniformly embrace approaches that potentially constrain profitable but manipulative design practices. Rather, our work provides evidence that when organizations do commit to ethical evaluation, metrics-based approaches can be highly effective at triggering ethical sensibility among design practitioners.

The conditional nature of this approach—requiring organizational willingness to prioritize ethical considerations—reflects broader challenges in corporate responsibility and stakeholder capitalism. However, for organizations genuinely committed to ethical design, our findings demonstrate that evaluation frameworks offer a practical, scalable intervention point for translating ethical intentions into design practice.

*5.5.3 Applications in Auditing and Regulatory Contexts.* Beyond organizational design practice, autonomy-focused evaluation frameworks could serve as valuable tools for external auditing and regulatory oversight. Current regulatory efforts to address dark patterns face significant challenges in establishing consistent, measurable standards for identifying manipulative design [37, 46]. Traditional regulatory approaches often rely on legal definitions, binary classifications, pattern strategy descriptons, or automated pattern matching that may not capture the nuanced ways manipulation manifests in interface design. Further, 'dark' designers may intentionally obscure manipulative elements to evade detection [46], and develop new patterns faster than regulations can adapt [11].

Our metrics-based approach could provide regulators with empirical tools for assessing interface manipulation that complement existing legal frameworks. Rather than relying solely on expert judgment or user complaints, regulatory bodies could employ standardized user centered autonomy-focused evaluations to systematically assess whether interfaces undermine user agency. When evaluation data demonstrates that users experience an interface as 'pressuring'

or 'covert,' this provides concrete, quantifiable evidence of potential manipulation that can inform regulatory decision-making.

Policy efforts have highlighted the need for technical standards that can bridge the gap between legal definitions and design implementation [75]. Autonomy-focused metrics could serve this bridging function by translating abstract legal concepts like 'informed consent' or 'user autonomy' into measurable user experience outcomes. This empirical grounding could help address what Gray et al. [37] identify as the lack of 'universally accepted definitions across the academic, legislative and regulatory space' that has limited the impact of dark pattern scholarship on regulatory action.

Additionally, third-party auditing organizations could employ these evaluation frameworks to provide independent assessments of interface ethics, similar to how accessibility audits currently assess compliance with disability access standards. Such approaches would enable more systematic, evidence-based evaluation of design ethics while providing organizations with clear, actionable feedback for improvement.

### 5.6 Redefining 'Good' Design

While there appears to be a resurgence of interest in what constitutes 'good' interface design [67, 78], discussions within the design community about when to break usability for reflection [66] and what is meant by unethical design [51] relate back to questions of values in design [30, 31]. While values in design work indicate that, depending on context, different values are emphasized by designers, our work highlights that it is not just designers, but also the incentive structures and evaluation frameworks they are assessed by, that guide what constitutes 'good design.'

Hedonic values, with the goal of 'delighting users,' have become a prevailing focus within the UX design community, reflecting a long-standing alignment with business objectives [19]. Designers are rarely free to choose the values they design for and instead must work with 'tactics of soft resistance'[84] and use complex rhetorical arguments[69] to push for values outside of immediate corporate interest, such as ethical design. This analysis of UX evaluation frameworks shows how the most common tools with business proximity fail to meaningfully advocate for user and broader societal values. When all that matters is a vague target of user experience, autonomy-focused evaluation frameworks might help translate ethical values into business risks—the only values management truly cares about [47].

### 5.7 Limitations and Future Work

This study has limitations that should be considered when interpreting the findings. First, our experiment focused specifically on dark patterns in social media interfaces. While these represent well-documented unethical design practices, future research should examine how evaluation frameworks influence decisions about other types of ethically questionable designs across different domains.

Second, participants evaluated static interface screenshots rather than interactive prototypes. While this approach ensured experimental control and allowed us to isolate the effects of evaluation data, it may not fully capture how designers would respond to evaluation frameworks when working with functional interfaces. Future studies could explore how our findings translate to real-world design processes with interactive prototypes and iterative design cycles.

Third, our sample consisted of professional designers recruited through Prolific Academic. While we employed rigorous screening criteria to ensure professional qualifications, the generalizability to other design contexts and organizational cultures remains to be established. Future research could examine how evaluation framework effects vary across different organizational contexts, team structures, and design methodologies.

Fourth, our experiment presented evaluation data in a controlled format that may not reflect the complexity of real-world user feedback. In practice, designers often encounter conflicting or ambiguous user data from multiple sources. Research examining how evaluation frameworks perform with more realistic, complex data presentations would be valuable.

Finally, while our study demonstrates that evaluation frameworks influence design decisions, we did not measure the long-term outcomes of these decisions on actual user experiences. Future research could examine whether designs approved under different evaluation frameworks lead to different user outcomes in terms of autonomy, well-being, and satisfaction.

Despite these limitations, this study provides important evidence that evaluation frameworks are not neutral tools but active shapers of ethical design decisions, opening new avenues for improving design practice through more thoughtful evaluation system design.

### 5.8 Future-Proofing Ethical Design for Computational Systems

Our findings have particular relevance as design practice increasingly incorporates computational and AI-driven approaches. Companies are beginning to utilize automated A/B testing, algorithmic design optimization, and AI-generated interface elements where metrics serve as key guides for automated design decisions. In these computational design contexts, evaluation frameworks become even more influential as they directly drive algorithmic optimization processes.

When aspects of interaction design become automated, the metrics used to guide these systems will fundamentally shape the resulting interfaces. Our research suggests that incorporating ethical dimensions into these evaluation frameworks could trigger ethical considerations in computational design systems, ensuring that automated design processes optimize for user autonomy and wellbeing rather than purely engagement or conversion metrics.

Future work should explore how ethics-focused evaluation frameworks can be integrated into automated design pipelines, how AI systems respond to multi-dimensional evaluation criteria that include ethical considerations, and what new challenges emerge when computational systems must balance potentially competing ethical and business objectives through quantified metrics.

### 5.9 Alternative Approaches and Scope of Application

Some critics might argue that instead of revealing ethics through metrics, organizations should focus on developing designers' ethical sensibility directly through training, education, or organizational culture change. We agree with research suggesting that not everyone possesses innate ethical foresight or expertise [78], making universal ethical training challenging to implement effectively.

Our metrics-based approach is not intended to replace ethical expertise or comprehensive ethics education. Rather, it provides a scalable intervention that can complement other ethical design initiatives. Evaluation frameworks can trigger ethical considerations in the moment of design decision-making, even when designers have not received extensive ethics training or when organizational pressures make ethical considerations less salient.

This approach may be particularly valuable in large organizations where ensuring consistent ethical training across all design personnel is challenging, or in contexts where external contractors and consultants participate in design decisions without deep organizational ethical context. Future research should examine how metrics-based ethical evaluation performs in combination with other ethical design interventions, and whether different combinations of approaches provide synergistic benefits.

Additionally, our focus on dark patterns represents one specific domain of ethical design. Future work should explore how evaluation framework effects extend to other ethical considerations such as accessibility, privacy, environmental impact, and social justice. Different ethical dimensions may require different evaluation approaches, and understanding these specifics will be important for developing comprehensive ethical evaluation systems.

### 5.10 Conclusion

This study demonstrates that evaluation metrics function as active design interventions that systematically shape how designers approach ethical decisions. By showing professional UX designers different types of user evaluation data, we revealed that the lens through which we evaluate user feedback fundamentally shapes what we consider acceptable design practice.

These findings have immediate practical implications: organizations seeking more ethical design outcomes should carefully consider what evaluation frameworks they employ and how those frameworks highlight or obscure ethical concerns. The path toward more responsible technology design may lie not just in training designers to be more ethical, but in designing evaluation systems that make ethical considerations visible and actionable.

## 6 DATA AVAILABILITY

The datasets from both the UX+Autonomy Evaluation Study with social media users and the UI Release Study with UX professionals will be made available upon publication through the Open Science Framework [4].

## REFERENCES


[1] Ritu Agarwal and Viswanath Venkatesh. 2002. Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability. *Information Systems Research* 13, 2 (2002), 168–186. https://doi.org/10.1287/isre.13.2.168.84

[2] Sanju Ahuja and Jyoti Kumar. 2024. Layered Analysis of Persuasive Designs: A Framework for Identification and Autonomy Evaluation of Dark Patterns. Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices Workshop at CHI conference on Human Factors in Computing Systems. https://ceur-ws.org/Vol-3720/paper1.pdf

[3] Tashina Alavi. 2020. Gray Patterns in UX: where do we draw the line between helpful vs. harmful design? https://uxdesign.cc/gray-patterns-in-ux-where-do-we-draw-the-line-between-helpful-vs-harmful-design-ced7fbaa8ad5. Accessed: 2023-2-15.

[4] Anonymous. 2025. Data and Materials for: Ethics-Focused User Metrics: How User Experience Metrics Can Trigger Ethical Design Sensibility. https://osf.io/nw2tj/?view_only=9f62bbab141841808118ca6339802bc4 Open Science Framework. Anonymous view-only link for peer review..

[5] Amanda Baughan, Mingrui Ray Zhang, Raveena Rao, Kai Lukoff, Anastasia Schaadhardt, Lisa D Butler, and Alexis Hiniker. 2022. "I don't even remember what I read": How design influences dissociation on social media. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3501899

[6] Joey Benedek and Trish Miner. 2002. Measuring Desirability: New methods for evaluating desirability in a usability lab setting. (Jan. 2002).

[7] Vincent Berthet. 2022. The Impact of Cognitive Biases on Professionals' Decision-Making: A Review of Four Occupational Areas. *Frontiers in Psychology* 12 (4 Jan. 2022), 802439. https://doi.org/10.3389/fpsyg.2021.802439

[8] Nataliia Bielova, Laura Litvine, Anysia Nguyen, Mariam Chammat, Vincent Toubiana, and Estelle Harry. 2024. The Effect of Design Patterns on (Present and Future) Cookie Consent Decisions. *USENIX Security Symposium. USENIX Association. Accepted for publication* (2024).

[9] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!" - Dark Patterns from the End-User Perspective. In *Designing Interactive Systems Conference 2021*. ACM, New York, NY, USA. https://doi.org/10.1145/3461778.3462086

[10] Nattapat Boonprakong, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. How do HCI researchers study cognitive biases? A scoping review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–20. https://doi.org/10.1145/3706598.3713450

[11] Harry Brignull. 2010. Dark Patterns. https://www.youtube.com/watch?v=zaubGV2OG5U

[12] Harry Brignull. 2010. Dark Patterns: User Interfaces Designed to Trick People. https://old.deceptive.design/main_page/index.html. Accessed: 2025-9-8.

[13] H Brignull, M Leiser, C Santos, and K Doshi. 2023. Deceptive patterns – user interfaces designed to trick you. https://www.deceptive.design/

[14] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (Nov. 1995).

[15] Egon Brunswik. 2001. Perception and the representative design of psychological experiments [1956]. In *The Essential Brunswik*. Oxford University PressNew York, NY, 260–264. https://doi.org/10.1093/oso/9780195130133.003.0016

[16] Deon Soul Calawen. 2022. Dark Patterns: Effect on Overall User Experience and Site Revisitation. (2022). https://doi.org/10.21427/BRW3-HZ03 Publisher: Technological University Dublin.

[17] A Cavoukian. 2020. Privacy by Design The 7 Foundational Principles Implementation and Mapping of Fair Information Practices. *IEEE consumer electronics magazine* 9 (1 March 2020), 78–82. https://doi.org/10.1109/MCE.2019.2953739

[18] Lorena Sánchez Chamorro and Carine Lallemand. 2024. Towards a second wave of manipulative design research: Methodological challenges of studying the effects of manipulative designs on users. In *CEUR Workshop Proceedings*, Vol. 3720. CEUR-WS.org, Aachen, Germany, 1–6. https://ceur-ws.org/Vol-3720/paper4.pdf

[19] Ravindra Chitturi, Rajagopal Raghunathan, and Vijay Mahajan. 2008. Delight by design: The role of hedonic versus utilitarian benefits. *Journal of marketing* 72, 3 (May 2008), 48–63. https://doi.org/10.1509/jmkg.72.3.048

[20] Shruthi Sai Chivukula, Jason Brier, and Colin M Gray. 2018. Dark intentions or persuasion?: UX designers' activation of stakeholder and user values. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*. ACM, New York, NY, USA, 87–91. https://doi.org/10.1145/3197391.3205417

[21] Sai Shruthi Chivukula and Colin Gray. 2025. *Universal methods of ethical design: 100 ways to become more ethically aware, responsible, and active in your design work*. Rockport, Beverly, MA. https://www.amazon.com/Universal-Methods-Ethical-Design-Responsible/dp/0760393087

[22] Shruthi Sai Chivukula, Colin Gray, Ziqing Li, Anne C Pivonka, and Jingning Chen. 2024. Surveying a landscape of ethics-focused design methods. *ACM Journal on Responsible Computing* (17 July 2024). https://doi.org/10.1145/3678988

[23] Shruthi Sai Chivukula, Aiza Hasib, Ziqing Li, Jingle Chen, and Colin M Gray. 2021. Identity claims that underlie ethical awareness and action. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/3411764.3445375

[24] Shruthi Sai Chivukula, Chris Rhys Watkins, Rhea Manocha, Jingle Chen, and Colin M Gray. 2020. Dimensions of UX Practice that Shape Ethical Awareness. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376459

[25] Lachlan D. Urquhart and Peter J. Craigon. 2021. The Moral-IT Deck: a tool for ethics by design. *Journal of responsible innovation* 8, 1 (Jan. 2021), 94–126. https://doi.org/10.1080/23299460.2021.1880112

[26] Ignacio Díaz-Oreiro, Gustavo López, Luis Quesada, and Luis A Guerrero. 2021. UX Evaluation with Standardized Questionnaires in Ubiquitous Computing and Ambient Intelligence: A Systematic Literature Review. *Advances in Human-Computer Interaction* 2021 (4 May 2021), 1–22. https://doi.org/10.1155/2021/5518722

[27] EDPB. 2023. Guidelines 03/2022 on deceptive design patterns in social media platform interfaces: how to recognise and avoid them. Version 2.0 | European Data Protection Board. https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en

[28] Rachel A Elphinston, Matthew J Gullo, and Jason P Connor. 2022. Validation of the Facebook addiction questionnaire. *Personality and individual differences* 195, 111619 (Sept. 2022), 111619. https://doi.org/10.1016/j.paid.2022.111619

[29] Mehrdad Farzandipour, Ehsan Nabovati, and Monireh Sadeqi Jabali. 2022. Comparison of usability evaluation methods for a health information system: heuristic evaluation versus cognitive walkthrough method. *BMC medical informatics and decision making* 22, 1 (18 June 2022), 157. https://doi.org/10.1186/s12911-022-01905-7

[30] Mary Flanagan and Helen Nissenbaum. 2014. *Values at play in digital games*. The MIT Press. https://doi.org/10.7551/mitpress/9016.001.0001

[31] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (Dec. 1996), 16–23. https://doi.org/10.1145/242485.242493

[32] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory*, Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E Gorman (Eds.). Springer Netherlands, Dordrecht, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

[33] Ajit G. Pillai, A Baki Kocaballi, Tuck Wah Leong, Rafael A. Calvo, Nassim Parvin, Katie Shilton, Jenny Waycott, Casey Fiesler, John C. Havens, and Naseem Ahmadpour. 2021. Co-designing resources for ethics education in HCI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/3411763.3441349

[34] Günther Gediga, Kai-Christoph Hamborg, and Duntsch. 1999. The IsoMetrics Usability Inventory. An operationalisation of ISO 9241-10 supporting summative and formative evaluation of software systems. *Behaviour and Information Technology* 18 (May 1999), 151–164. https://doi.org/10.1080/014492999119057

[35] Colin M Gray, Shruthi Sai Chivukula, and Ahreum Lee. 2020. What kind of work do "asshole designers" create? Describing properties of ethical concern on reddit. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. ACM, New York, NY, USA, 61–73. https://doi.org/10.1145/3357236.3395486

[36] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. https://doi.org/10.1145/3173574.3174108

[37] Colin M. Gray, Cristiana Santos, and Nataliia Bielova. 2023. Towards a Preliminary Ontology of Dark Patterns Knowledge. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–9. https://doi.org/10.1145/3544549.3585676

[38] Johanna Gunawan, Woodrow Hartzog, Neil Richards, David Choffnes, and Christo Wilson. 2025. Dark Patterns as Disloyal Design. *Indiana law journal (Indianapolis, Ind.: 1926)* 100, 4 (2025), 3. https://www.repository.law.indiana.edu/ilj/vol100/iss4/3

[39] International Organization for Standardization. 2010. *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.* Standard. International Organization for Standardization, Geneva, CH.

[40] Robin Jeffries, James R. Miller, Cathleen Wharton, and Ka-Mei Uyeda. 1991. User interface evaluation in the real world: a comparison of four techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, New Orleans, Louisiana, USA, 119–124.

[41] Merve Karakus, Jennifer Smith, and Robert Wilson. 2023. Ethical dimensions of accessible UX design: Balancing stakeholder interests and user intentions. *International Journal of Human-Computer Studies* 178 (Oct. 2023), 103089. https://doi.org/10.1016/j.ijhcs.2023.103089

[42] Maximilian Kiener. 2021. When do nudges undermine voluntary consent? *Philosophical studies* 178, 12 (4 May 2021), 4201–4226. https://doi.org/10.1007/s11098-021-01644-x

[43] Westley Knight. 2019. Business Objectives vs. User Goals. In *UX for Developers: How to Integrate User-Centered Design Principles Into Your Day-to-Day Development Work*, Westley Knight (Ed.). Apress, Berkeley, CA, 29–36. https://doi.org/10.1007/978-1-4842-4227-8_3

[44] Carine Lallemand, Guillaume Gronier, and Vincent Koenig. 2015. User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in human behavior* 43 (1 Feb. 2015), 35–48. https://doi.org/10.1016/j.chb.2014.10.048

[45] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work (Lecture Notes in Computer Science)*, Andreas Holzinger (Ed.). Springer, Berlin, Heidelberg, 63–76. https://doi.org/10.1007/978-3-540-89350-9_6

[46] M. R. Leiser and Cristiana Santos. 2023. Dark Patterns, Enforcement, and the emerging Digital Design Acquis: Manipulation beneath the Interface. https://papers.ssrn.com/abstract=4431048

[47] Eilat Lev Ari, Maayan Roichman, and Eran Toch. 2024. Strategies of product managers: Negotiating social values in digital product design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, USA. https://doi.org/10.1145/3613904.3642409

[48] Yuwen Lu, Chao Zhang, Yuewen Yang, Yaxing Yao, and Toby Jia-Jun Li. 2024. From awareness to action: Exploring end-user empowerment interventions for dark patterns in UX. *Proceedings of the ACM on human-computer interaction* 8, CSCW1 (17 April 2024), 1–41. https://doi.org/10.1145/3637336

[49] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a light on dark patterns. *The journal of legal analysis* 13, 1 (March 2021), 43–109. https://doi.org/10.1093/jla/laaa006

[50] Arunesh Mathur. 2020. Identifying and measuring manipulative user interfaces at scale on the web. (2020). https://dataspace.princeton.edu/handle/88435/dsp012f75rc09f Accepted: 2020-11-20T05:59:39Z Publisher: Princeton, NJ : Princeton University.

[51] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 1–32. https://doi.org/10.1145/3359183

[52] Shikha Mehta, Shruthi Sai Chivukula, Colin M Gray, and Ritika Gairola. 2024. Anti-Heroes: An ethics-focused method for responsible designer intentions. *arXiv [cs.HC]* (6 May 2024). arXiv:2405.03674 [cs.HC] http://arxiv.org/abs/2405.03674

[53] Rob R Meijer, A Susan M Niessen, and Marvin Neumann. 2023. Psychological and Educational Testing and Decision-Making: The Lack of Knowledge Dissemination in Textbooks and Test Guidelines. In *Essays on Contemporary Psychometrics*, L Andries van der Ark, Wilco H M Emons, and Rob R Meijer (Eds.). Springer International Publishing, Cham, 47–67. https://doi.org/10.1007/978-3-031-10370-4_3

[54] Anna-Lena Meiners, Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2024. A benchmark for the UEQ+ framework: Construction of a simple tool to quickly interpret UEQ+ KPIs. *International Journal of Interactive Multimedia and Artificial Intelligence* 9, Regular issue (2024), 104. https://doi.org/10.9781/ijimai.2023.05.003

[55] Alexander Meschtscherjakov, Magdalena Gärtner, Alexander Mirnig, Christina Rödel, and Manfred Tscheligi. 2016. The persuasive potential questionnaire (PPQ): Challenges, drawbacks, and lessons learned. In *Persuasive Technology.* Springer International Publishing, Cham, 162–175. https://doi.org/10.1007/978-3-319-31510-2_14

[56] Thomas Mildner, Merle Freye, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. https://doi.org/10.1145/3563657.3595964 arXiv:2305.13154 [cs].

[57] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3580695

[58] Alberto Monge Roffarello and Luigi De Russis. 2022. Towards Understanding the Dark Patterns That Steal Our Attention. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22).* Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3491101.3519829

[59] Nissenbaum. 2014. Groundwork for Values in Games. In *Values at Play in Digital Games.* The MIT Press. https://doi.org/10.7551/mitpress/9016.003.0004

[60] Robert Noggle. 2022. *The Ethics of Manipulation* (summer 2022 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/

[61] William Odom, Erik Stolterman, and Amy Yo Sue Chen. 2022. Extending a theory of slow technology for design through artifact analysis. *Human-computer interaction* 37, 2 (4 March 2022), 150–179. https://doi.org/10.1080/07370024.2021.1913416

[62] Guido Palazzo, Franciska Krings, and Ulrich Hoffrage. 2013. Ethical blindness. *SSRN Electronic Journal* (6 Feb. 2013). https://doi.org/10.2139/ssrn.2212617

[63] Nilay Patel. 2024. Duolingo CEO Luis Von Ahn wants you addicted to learning. https://app.podscribe.ai/episode/116224228?transcriptVersionReqId=0124c65e-bbdf-4aef-ab39-1692ccc8207e

[64] D K Peterson. 2002. The relationship between unethical behavior and the dimensions of the ethical climate questionnaire. *Journal of business ethics* 41, 4 (2002), 313–326. https://doi.org/10.1023/a:1021243117958

[65] Erich Prem. 2023. From ethical AI frameworks to tools: a review of approaches. *AI and ethics* 3, 3 (Aug. 2023), 699–716. https://doi.org/10.1007/s43681-023-00258-9

[66] A J Quanjer and Maarten H Lamers. 2014. Breaking Usability Rules to Enable Reflection. (2014). https://www.semanticscholar.org/paper/b77e043aacbe6733fb8323eb2b445fb3de6c86f0

[67] Christopher David Quintana. 2024. *Characterizing Digital Design: A Philosophical Approach*. Ph. D. Dissertation. Villanova University. https://www.proquest.com/docview/3066598630?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

[68] Kerry Rodden, Hilary Hutchinson, and Xin Fu. 2010. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/1753326.1753687

[69] Emma Rose and Josh Tenenberg. 2016. Arguing about design: A taxonomy of rhetorical strategies deployed by user experience practitioners. In *Proceedings of the 34th ACM International Conference on the Design of Communication*. ACM, New York, NY, USA. https://doi.org/10.1145/2987592.2987608

[70] Saad, L., Roesler, E., Phillips, B. K., & Trafton, J. G. 2025. Choosing the "perfect" scale: a primer to evaluate existing scales in HRI. *(under review)* (2025). http://hriscaledatabase.psychology.gmu.edu/

[71] E Sanders. 2002. From user-centered to participatory design approaches. In *Design and the Social Sciences*. CRC Press, 18–25. https://doi.org/10.1201/9780203301302-8

[72] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Die UX KPI - Wunsch und Wirklichkeit. In *Mensch und Computer 2017 - Usability Professionals*, Steffen Hess and Holger Fischer (Eds.). Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2017-up-0100

[73] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph 'jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility (CC '05)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/1094562.1094569

[74] Herbert A Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics* 69, 1 (Feb. 1955), 99. https://doi.org/10.2307/1884852

[75] Caroline Sinders. 2021. Designing Against Dark Patterns. (2021).

[76] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Technology, autonomy, and manipulation. *Internet policy review* 8, 2 (30 June 2019). https://doi.org/10.14763/2019.2.1410

[77] Lorena Sánchez Chamorro, Kerstin Bongard-Blanchy, and Vincent Koenig. 2023. Ethical tensions in UX design practice: Exploring the fine line between persuasion and manipulation in online interfaces. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. ACM, New York, NY, USA. https://doi.org/10.1145/3563657.3596013

[78] Lorena Sánchez Chamorro, Kerstin Bongard-Blanchy, and Vincent Koenig. 2023. Ethical Tensions in UX Design Practice: Exploring the Fine Line Between Persuasion and Manipulation in Online Interfaces. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2408–2422. https://doi.org/10.1145/3563657.3596013

[79] Rachel L Thomas and David Uminsky. 2022. Reliance on metrics is a fundamental challenge for AI. *Patterns (New York, N.Y.)* 3, 5 (May 2022), 100476. https://doi.org/10.1016/j.patter.2022.100476

[80] Nynke Tromp, Paul Hekkert, and Peter-Paul Verbeek. 2011. Design for socially responsible behavior: A classification of influence based on intended user experience. *Design issues* 27, 3 (1 July 2011), 3–19. https://doi.org/10.1162/desi_a_00087

[81] Christof van Nimwegen, Kristi Bergman, and Almila Akdag. 2022. Shedding light on assessing Dark Patterns: Introducing the System Darkness Scale (SDS). In *35th International BCS Human-Computer Interaction Conference*. BCS Learning & Development, 1–10. https://doi.org/10.14236/ewic/HCI2022.7

[82] Arnold P O S Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. ACM, New York, NY, USA, 521–530. https://doi.org/10.1145/1868914.1868973

[83] Julie R Williamson and Daniel Sundén. 2016. Deep cover HCI: the ethics of covert research. *interactions* 23, 3 (26 April 2016), 45–49. https://doi.org/10.1145/2897941

[84] Richmond Y Wong. 2021. Tactics of soft resistance in user experience professionals' values work. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (13 Oct. 2021), 1–28. https://doi.org/10.1145/3479499

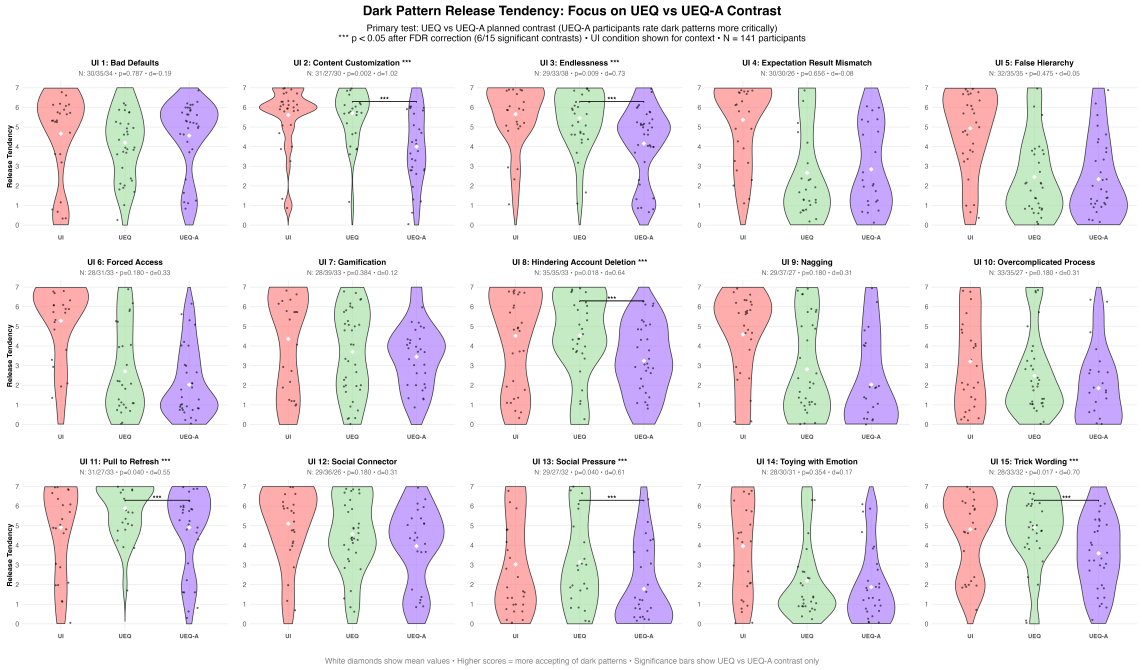
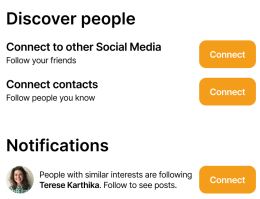
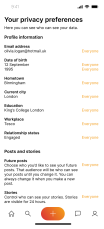



Fig. 8. Interface-specific effects across evaluation conditions. Violin plots show the distribution of release tendency ratings (0-7 scale: 0 = *definitely would not release*, 7 = *definitely would release*) for each interface, with significant differences after FDR correction marked. Dark patterns show varying sensitivity to evaluation framework effects, with content customization, endlessness, and trick wording showing the strongest autonomy-focused evaluation effects.

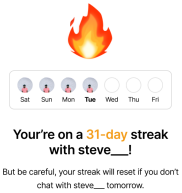
Received 11 September 2025



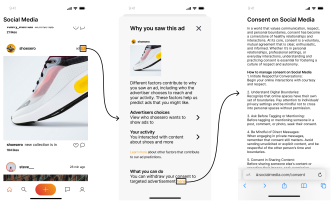

Social Connector



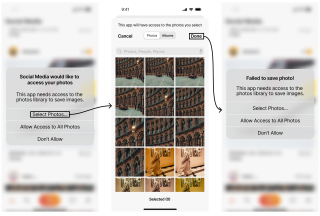

Bad Defaults



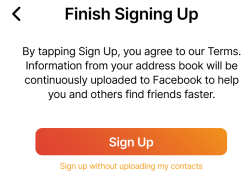

Gamification



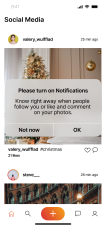

Expectation Result Mismatch




Forced Access

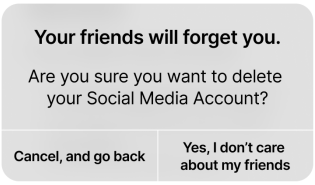



False Hierarchy




Nagging


Overcomplicated Process




Toying with Emotion

Fig. 9. Complete dark pattern interface overview. The nine remaining dark pattern interfaces not shown in Figure 6, ordered by overall mean rejection rate from lowest (top-left) to highest (bottom-right). High resolution versions of all interfaces are available in the OSF repository [4].

| Pattern | Autonomy-Focused Items | | | | Stimulation | | Perspicuity | |
|---|---|---|---|---|---|---|---|---|
| | pressuring vs. suggesting | addictive vs. non-addictive | covert vs. revealed | deceptive vs. benevolent | boring vs. exciting* | not interesting vs. interesting | complicated vs. easy* | confusing vs. clear* |
| NAG | -2.26 | 0.89 | 0.78 | -1.58 | -1.26 | -1.79 | 0.41 | -0.23 |
| OP | -0.39 | 0.63 | 0.18 | -0.61 | -0.48 | -0.20 | -0.83 | -1.12 |
| HAD | -0.46 | -0.30 | 0.56 | -0.76 | -0.15 | 0.03 | 0.43 | 0.53 |
| SBD | -0.33 | 1.26 | 1.49 | -0.56 | -0.09 | 0.28 | 0.98 | 1.30 |
| ERM | -0.65 | 0.88 | 0.06 | -1.28 | -0.19 | -0.24 | -0.59 | -1.29 |
| FH | -2.11 | 0.86 | -0.43 | -2.09 | -0.86 | -0.76 | 0.44 | -0.76 |
| TW | -1.29 | 1.07 | -0.15 | -1.76 | -0.44 | -0.06 | -0.66 | -0.86 |
| TWE | -1.93 | -0.29 | 0.25 | -1.97 | 0.00 | -0.83 | 0.53 | -0.40 |
| FA | -1.59 | 0.29 | -0.44 | -1.53 | -0.58 | -0.75 | -0.79 | -0.90 |
| GAM | -1.83 | -0.88 | 0.62 | -0.83 | -0.23 | -0.63 | 1.06 | 0.61 |
| SP | -1.02 | -0.28 | 0.19 | -0.31 | -0.40 | -0.86 | 1.72 | 1.16 |
| SC | -1.13 | 0.21 | 0.38 | -0.89 | -0.18 | -0.22 | 0.97 | 0.44 |
| CC | -0.71 | 0.17 | 0.87 | -0.61 | -0.35 | -0.16 | 1.09 | 0.58 |
| END | -0.48 | -1.47 | 0.19 | -0.82 | 0.35 | 0.18 | 1.46 | 0.49 |
| PTR | -0.04 | -1.00 | 1.11 | 0.07 | 0.17 | 0.62 | 1.66 | 0.59 |

| Pattern | Efficiency | | Dependability | | Attractiveness | | Mean | UX KPI |
|---|---|---|---|---|---|---|---|---|
| | inefficient vs. efficient* | cluttered vs. organized | unpredictable vs. predictable | obstructive vs. supportive* | annoying vs. enjoyable | unfriendly vs. friendly | (over all surveyed items) | (UEQ-S without novelty) |
| NAG | -0.71 | -0.71 | 0.16 | -2.08 | -1.97 | -1.73 | -0.86 | -0.94 |
| OP | -0.42 | -1.20 | -0.29 | -0.35 | -0.57 | -0.88 | -0.47 | -0.57 |
| HAD | 0.43 | 0.40 | 0.70 | -0.22 | -0.44 | -0.47 | 0.02 | 0.18 |
| SBD | 0.34 | 0.95 | 0.71 | -0.11 | -0.87 | -0.65 | 0.34 | 0.45 |
| ERM | -0.34 | -1.19 | 0.03 | -0.72 | -1.00 | -1.36 | -0.56 | -0.56 |
| FH | -0.25 | 0.38 | 0.46 | -1.17 | -1.97 | -2.13 | -0.81 | -0.56 |
| TW | -0.23 | 0.52 | -0.91 | -1.35 | -1.85 | -1.82 | -0.70 | -0.60 |
| TWE | 0.03 | 0.69 | 0.21 | -1.85 | -2.15 | -2.15 | -0.70 | -0.42 |
| FA | -0.82 | -0.15 | -0.95 | -1.23 | -1.69 | -1.44 | -0.90 | -0.85 |
| GAM | 0.34 | -0.10 | 0.44 | -0.55 | -0.73 | -0.87 | -0.25 | 0.10 |
| SP | 0.41 | 0.72 | 0.61 | -0.29 | -0.67 | -0.60 | 0.03 | 0.29 |
| SC | 0.57 | 0.90 | 0.03 | 0.36 | -0.73 | -0.71 | 0.00 | 0.33 |
| CC | 0.08 | -0.15 | -0.06 | 0.17 | 0.42 | -0.08 | 0.09 | 0.23 |
| END | 0.81 | 0.22 | -0.50 | -0.20 | 0.05 | -0.10 | 0.01 | 0.51 |
| PTR | 0.97 | 0.75 | -0.28 | 1.07 | 0.34 | 0.80 | 0.49 | 0.85 |

Table 2. UX+Autonomy Evaluation Study Results: Overall Mean and Item Means per Dark Pattern. Data from Stage 2 user evaluation study (N=126) showing evaluation scores across UEQ and autonomy-focused dimensions. Item Means ≤ -0.75 are shown in purple, and Item Means ≥ 0.75 are shown in teal. Numbers in colored boxes indicated the lowest scoring attribute per pattern. NAG = Nagging, OP = Overcomplicated Process, HAD = Hindering Account Deletion, SBD = Sneaking Bad Default, ERM = Expectation Result Mismatch, FH = False Hierarchy, TW = Trick Wording, TWE = Toying with Emotion, FA = Forced Access, GAM = Gamification, SP = Social Pressure, SC = Social Connector, CC = Content Customization, END = Endlessness, PTR = Pull To Refresh. Autonomy-focused items designed to assess user manipulation and control, * from UEQ-S, other items and Attractiveness from the full UEQ.