

UNIVERSIDAD PANAMERICANA

FACULTAD DE INGENIERÍA

Con estudios incorporados a la
Secretaría de Educación Pública

**“EL PAPEL DE LA TRANSPARENCIA EN LA RELACIÓN ENTRE
CONCIENCIA SOBRE WEB SCRAPING, PREOCUPACIONES DE
PRIVACIDAD Y CONFIANZA DEL USUARIO”**

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRÍA EN INGENIERÍA

P R E S E N T A

CECILIA BETANCOURT BARRITA

ASESOR:

LORENA ALEXANDRA BERUMEN GLINZ

COASESOR:

DAVID ESCOBAR CASTILLEJOS

Índice general

Resumen	1
1 Introducción	2
1.1 Antecedentes	2
1.2 Motivación	2
1.3 Planteamiento del problema	3
1.4 Objetivo general	4
1.5 Objetivos específicos	4
1.6 Hipótesis	4
1.7 Estructura de la tesis	5
2 Manuscrito de investigación	6
3 Discusión	40
3.1 Interpretación de los resultados	40
3.2 Relación con la literatura	41
3.3 Implicaciones prácticas	41
4 Conclusiones y trabajo futuro	42
4.1 Conclusiones	42
4.2 Contribuciones	42
4.3 Limitaciones	43
4.4 Trabajo futuro	43

Resumen

En la sociedad digital actual, marcada por flujos constantes de “información líquida” y por formas de vigilancia cada vez más ubicuas, la recolección y el tratamiento automatizado de datos personales se volvieron elementos estructurales de la economía y de la vida cotidiana en línea. En ese escenario, la transparencia en las prácticas de uso de datos no es solo un requisito normativo: es una condición necesaria para la protección efectiva de la privacidad y para el ejercicio de los derechos vinculados al libre desarrollo de la personalidad. Esta tesis analiza el papel de la transparencia en un escenario simulado de incorporación (*onboarding*) a una red social, y examina su impacto sobre la confianza del usuario, así como su función moderadora en la relación entre la conciencia sobre prácticas de *web scraping* y las preocupaciones de privacidad.

Se diseñó un experimento basado en encuesta en el que los participantes interactuaron con una interfaz que incorporaba dos condiciones contrastantes de transparencia respecto a la recolección automatizada de datos: baja transparencia y alta transparencia. A partir de esa manipulación se midieron tres constructos principales: conciencia sobre *web scraping*, preocupaciones de privacidad y confianza en la plataforma. Los resultados muestran que una mayor conciencia sobre estas prácticas incrementa de forma significativa las preocupaciones de privacidad; sin embargo, cuando la plataforma informa de manera clara y específica qué datos se recogen, con qué finalidad y bajo qué condiciones, la transparencia atenúa la fuerza de esa asociación. Al mismo tiempo, las preocupaciones de privacidad no resultaron un predictor significativo de la confianza, lo que sugiere que la confianza se forma a partir de un conjunto más amplio de factores técnicos, normativos y experienciales. En conjunto, los hallazgos contribuyen a la discusión sobre privacidad en Internet al mostrar cómo la transparencia puede modular los procesos cognitivos y emocionales vinculados al tratamiento de datos personales, y ofrecen orientaciones prácticas para el diseño de interfaces más claras, responsables y centradas en los derechos de las personas usuarias.

Capítulo 1. Introducción

1.1 Antecedentes

Las plataformas digitales contemporáneas operan como infraestructuras básicas de la vida social, económica y comunicativa. Su funcionamiento descansa en la recolección intensiva de datos personales y de uso, que permite personalizar contenidos, recomendar información “relevante” y sostener la atención de los usuarios (Hirsh et al., 2000; Bélanger and Crossler, 2011). Una parte importante de esos datos se obtiene mediante la extracción automatizada de información disponible en la web, práctica conocida como *web scraping* (Lotfi et al., 2022; Pichiyan et al., 2023; Li and Kollnig, 2024). Aunque, en determinadas condiciones, esta técnica puede ubicarse dentro de los márgenes de la licitud (Li and Kollnig, 2024; Gallese, 2023), su carácter silencioso y poco visible para el usuario final plantea interrogantes centrales para la interacción humano-computadora (HCI): qué datos se recogen, con qué finalidad se emplean, cómo se reutilizan y hasta qué punto las personas conservan algún control efectivo sobre dichos procesos (Al-Natour et al., 2020; Bélanger and James, 2020; Kyi et al., 2024). Esa opacidad tensiona principios fundamentales de la experiencia de usuario y de la protección de datos, como la transparencia, la autonomía informativa y la confianza (Culnan and Armstrong, 1999; Bélanger and Crossler, 2011; Lukyanenko et al., 2022).

Desde el diseño de interacción con el usuario, la recolección automatizada de datos no se reduce a un problema de cumplimiento normativo: es también un desafío para la protección de la privacidad y para la configuración de la experiencia de uso (Bélanger and James, 2020; Marengo, 2024). Normas como el Reglamento General de Protección de Datos (General Data Protection Regulation (GDPR), 2018) reforzaron la centralidad de la transparencia y de los derechos de los titulares de los datos. Sin embargo, en la práctica muchas de esas exigencias se traducen en avisos extensos, abstractos y poco accesibles, que rara vez se leen o se comprenden por la mayoría de las personas (Obar, 2022; Kyi et al., 2024). Diversos trabajos mostraron que los usuarios suelen sobreestimar la capacidad protectora de las políticas de privacidad y subestimar el alcance real de la recolección y combinación de sus datos (Xu et al., 2008; Crossler and Bélanger, 2019). Esa brecha entre lo que las personas creen que ocurre con sus datos y lo que efectivamente sucede alimenta las preocupaciones de privacidad y se inscribe en lo que se denominó la *paradoja de la privacidad*: altos niveles de inquietud declarada que no siempre se traducen en conductas de protección consistentes (Awad and Krishnan, 2006; Sutanto et al., 2013; Bélanger and Crossler, 2011).

1.2 Motivación

En este contexto, la transparencia es un concepto multidimensional, que implica apertura, claridad y corrección en la información ofrecida al usuario sobre el tratamiento de sus datos (Obar, 2022; Dehling and Sunyaev, 2024). Los entornos digitales actuales exigen superar las declaraciones legales estáticas y avanzar hacia formas de transparencia interactivas, integradas en el propio flujo de uso de las plataformas y adaptadas a las capacidades cognitivas de las personas (Gierlich-Joas et al., 2022; Esmailzadeh, 2019; Dehling and Sunyaev, 2024). El modo en que se comunican las prácticas de uso de datos, el momento en que se presenta la información, su nivel de detalle, el lenguaje empleado y los controles asociados inciden directamente en los modelos mentales de los usuarios, en el esfuerzo necesario para comprender las implicaciones del tratamiento y en su percepción de justicia, seguridad y control (Culnan and Armstrong, 1999; Xu et al., 2008; Obar, 2022). Cuando las plataformas no explicitan de forma adecuada sus prácticas de recolección, circulación y reutilización de datos, se intensifican las percepciones de riesgo y vulnerabilidad, lo que afecta la evaluación global que los usuarios hacen del sistema y su disposición a confiar en él (Martin, 2018; Mcknight et al., 2011; Gefen et al., 2003).

1.3 Planteamiento del problema

A pesar de la relevancia de estos temas, gran parte de la investigación empírica trató la transparencia como un antecedente directo de las preocupaciones de privacidad o de la confianza, sin explorar suficientemente su papel como variable moderadora en escenarios donde la conciencia sobre prácticas automatizadas, como el *web scraping*, se hace explícita (Bélanger and James, 2020; Gierlich-Joas et al., 2022). Al mismo tiempo, la literatura no ofrece resultados concluyentes sobre si las preocupaciones de privacidad reducen de manera sistemática la confianza en las plataformas digitales; en muchos casos, la confianza se mantiene por factores como la habituación, la percepción de indispensabilidad del servicio o la valoración de beneficios funcionales (Awad and Krishnan, 2006; Martin, 2018; Lukyanenko et al., 2022). Esas lagunas teóricas hacen necesario estudiar cómo la transparencia, entendida como recurso de diseño de interacción y como garantía para el ejercicio de la autodeterminación informativa, puede modular la relación entre conciencia sobre la recolección automatizada, preocupaciones de privacidad y confianza del usuario (Bélanger and James, 2020; Dehling and Sunyaev, 2024).

El presente trabajo se inscribe en este marco y desarrolla un experimento de transparencia interactiva en el contexto de una plataforma de redes sociales. Se diseñó un escenario de *onboarding* que expone a los participantes a dos condiciones contrastantes de transparencia respecto al uso de *web scraping*: una condición de baja transparencia, basada en descripciones mínimas, genéricas y

orientadas al cumplimiento, y una condición de alta transparencia, con explicaciones explícitas y contextualizadas sobre qué datos se recogen, por qué se utilizan y qué opciones de control están disponibles. A partir de ese diseño se analizan tres constructos centrales para HCI: la conciencia sobre la recolección automatizada de datos, las preocupaciones de privacidad y la confianza en la plataforma, tomando como referencia marcos teóricos consolidados sobre privacidad, transparencia y confianza en sistemas digitales (Culnan and Armstrong, 1999; Bélanger and Crossler, 2011; Mcknight et al., 2011; Dehling and Sunyaev, 2024). De manera particular, la tesis explora en qué medida la transparencia modera la relación entre la conciencia de *web scraping* y las preocupaciones de privacidad, y cómo esas preocupaciones se vinculan con la confianza del usuario en un entorno de interacción simulado.

1.4 Objetivo general

Analizar, en el contexto de una plataforma de redes sociales, de qué manera la transparencia en el uso de datos modera la relación entre la conciencia sobre prácticas de *web scraping* y las preocupaciones de privacidad, así como su vínculo con la confianza del usuario en la plataforma.

1.5 Objetivos específicos

1. Diseñar e implementar un experimento de transparencia interactiva durante el proceso de *onboarding* de una red social, contrastando condiciones de baja y alta transparencia sobre prácticas de recolección automatizada de datos.
2. Analizar la asociación entre la conciencia sobre la recolección automatizada de datos (*web scraping*) y las preocupaciones de privacidad de los usuarios.
3. Evaluar si la transparencia percibida atenúa la fuerza de la relación entre conciencia de *web scraping* y preocupaciones de privacidad.
4. Examinar en qué medida las preocupaciones de privacidad influyen en la confianza del usuario en la plataforma digital.

1.6 Hipótesis

A partir del marco teórico sobre privacidad, transparencia y confianza en entornos digitales, esta tesis plantea una relación central entre la conciencia de las personas sobre prácticas de *web scraping*, sus preocupaciones de privacidad y su confianza en la plataforma. Al mismo tiempo, se asume que la transparencia en el uso de datos actúa como recurso de diseño capaz de modificar

la forma en que dicha conciencia se traduce en preocupación. Con base en ello se formulan las siguientes hipótesis de investigación:

- H1.** Un mayor nivel de conciencia sobre las prácticas de *web scraping* se asocia con un aumento en las preocupaciones de privacidad del usuario.
- H2.** La transparencia en la presentación del uso de datos modera la relación entre la conciencia sobre *web scraping* y las preocupaciones de privacidad, de modo que, bajo condiciones de alta transparencia, la asociación positiva entre ambos constructos se debilita.
- H3.** Un mayor nivel de preocupaciones de privacidad se asocia con una menor confianza del usuario en la plataforma digital.

1.7 Estructura de la tesis

La tesis se organiza siguiendo la lógica del proceso de investigación. Tras el resumen, la introducción, la motivación, los objetivos y las hipótesis, el capítulo de marco teórico revisa los conceptos de privacidad, conciencia sobre *web scraping*, transparencia en el uso de datos y confianza del usuario, junto con los principales trabajos relacionados en el ámbito de la interacción humano-computadora.

El capítulo siguiente expone el diseño metodológico del estudio experimental, describe el escenario de *onboarding*, las condiciones de transparencia implementadas, el instrumento de medición, las características de la muestra y los procedimientos de recolección y análisis de datos. A continuación se presentan los resultados empíricos del experimento, con énfasis en la contrastación de las hipótesis planteadas.

La tesis cierra con la discusión de los hallazgos, sus implicaciones teóricas y prácticas, las limitaciones del estudio y las posibles líneas de investigación futura, antes de las conclusiones generales y las referencias bibliográficas utilizadas.

Capítulo 2. Manuscrito de investigación

El presente capítulo incorpora, en modalidad de *compendio europeo*, el manuscrito de investigación que sustenta el cuerpo empírico de esta tesis. El artículo cubre el marco teórico, el diseño metodológico, los procedimientos de recolección de datos y los resultados del experimento de transparencia interactiva sobre prácticas de *web scraping*.

Referencia: Betancourt Barrita, C., Berumen Glinz, L. A., & Escobar Castillejos, D. (2025). *The role of transparency in the relationship between web scraping awareness, privacy concerns and user trust*. Manuscrito sometido a Frontiers.

A continuación se reproduce íntegramente el manuscrito.

Transparency as a Moderator of the Relationship Between Web-Scraping Awareness and Privacy-Protective Behavior: A Quasi-Experimental Onboarding Study

Cecilia Betancourt-Barrita¹, Lorena Alexandra Berumen-Glinz¹, Francisco Cal Y Mayor-Peña^{1,*}, and David Escobar-Castillejos^{1,*}

¹ *Universidad Panamericana. Facultad de Ingeniería. Augusto Rodin 498, Ciudad de México, 03920, México*

Correspondence*:

Corresponding Author

laberumen@up.edu.mx, descobarc@up.edu.mx

2 ABSTRACT

3 **Introduction:** Automated data collection practices, such as web scraping, are widely used for
4 personalization on interactive platforms, and they raise recurring questions about privacy, transparency,
5 and user trust. Prior work has examined these constructs separately. Less is known about whether
6 transparency in onboarding disclosures is associated with a weaker relationship between awareness of
7 automated data collection and privacy-protective behavior. This study examines the relationship between
8 privacy-protective behavior and user trust in a simulated social media environment.

9 **Methods:** A quantitative, between-subjects, scenario-based quasi-experimental survey was conducted
10 with 325 participants. The study employed a simulated Instagram onboarding scenario with two
11 transparency conditions: low and high. Privacy-related outcomes were operationalized as two binary
12 privacy-protective behaviors (adjusting privacy settings and refusing to use a platform), summed into
13 a behavior index that takes integer values of 0, 1, or 2. Additional measures assessed web-scraping
14 awareness, user trust, perceived transparency, and control variables related to social media use and
15 privacy sensitivity. Hypotheses were tested using ordinary least squares regression for the behavior index
16 and continuous outcomes, and logistic regression for each binary behavior, including an interaction term
17 to examine the moderating effect of the transparency condition.

18 **Results:** Greater awareness of web scraping was associated with a higher likelihood of low-barrier
19 privacy-protective behavior, particularly adjusting privacy settings. The transparency condition was
20 associated with a weaker awareness-behavior relationship, consistent with the proposed moderation
21 model. In contrast, the privacy-protective behavior index did not significantly predict user trust in the
22 platform.

23 **Discussion:** The findings are consistent with transparency operating as a contextual buffer of the link
24 between awareness of automated data collection and privacy-protective behavior, rather than as a direct
25 antecedent of trust. Because condition assignment was self-selected and baseline awareness differed
26 across conditions, the results are interpreted as conditional associations within a simulated onboarding

context rather than as causal effects. The results inform the design of onboarding disclosures in AI-driven platforms.

Keywords: Transparency; Web scraping; Privacy concerns; User trust; Human-computer interaction; Social media onboarding

1 INTRODUCTION

Interactive platforms rely on extensive data collection to personalize content, suggest relevant information, and improve user engagement (Hirsh et al., 2000). A substantial amount of this information is obtained from the automated extraction of publicly available content, commonly referred to as web scraping, which raises fundamental human-computer interaction (HCI) questions about what data is collected, why it is used, and how much control users believe they possess over these processes (Lotfi et al., 2022; Weerasinghe et al., 2023). Although web scraping may be lawful under specific circumstances, its lack of transparency undermines key HCI principles, particularly awareness, control, and trust. Regulatory frameworks, including the General Data Protection Regulation (GDPR), prioritize transparency and user rights (General Data Protection Regulation (GDPR), 2018); however, the practical efficacy of these disclosure mechanisms remains ambiguous (Li and Kollnig, 2024; Gallese, 2023).

From an interaction perspective, automated data collection is a matter of regulation and compliance, as well as data protection and user experience. As artificial intelligence (AI) systems rely on increasingly large-scale data, privacy risks extend further, particularly when sensitive information is collected, used, or repurposed without explicit consent (Gomstyn and Jonker, 2025; Sparkes, 2023). The timing, clarity, and available controls in an interface affect users' mental models, cognitive effort, and perceptions of fairness. In the present study, these mechanisms are treated as theoretical explanations for why transparency may matter, rather than as directly measured mediators. When platforms do not articulate their data practices transparently, users report reduced autonomy and greater perceived privacy risks, which in turn shape how they evaluate the system (Al-Natour et al., 2020; Bélanger and James, 2020). These reactions are evident on social media, where inference, personalization, and targeted advertising operate with limited transparency for end users (Marengo, 2024; Kyi et al., 2024).

Generative AI has expanded the scope of automated data extraction (Sengar et al., 2025; Pichiyan et al., 2023) and raised related concerns about intellectual property, competitive intelligence, and the balance between open web principles and large-scale data repurposing (Gallese, 2023). Conveying data practices in ways that align with user expectations and support perceived control, without imposing excessive cognitive load, remains an open HCI design problem (Gallese, 2023).

Empirical research further confirms these experiential challenges, revealing persistent gaps between what users believe and what actually happens to their data. Most individuals do not read privacy policies, overestimate their protective power, and underestimate the extent of data collection (Tsai et al., 2011). This lack of understanding exacerbates privacy concerns and erodes trust in AI-mediated systems (Agozie and Kaya, 2021; Elliott and Soifer, 2022). Recent evidence on AI-based inferences from personal data further shows that user awareness of inferential collection, even when not explicitly disclosed, shapes privacy preferences and willingness to share information (Asthana et al., 2024). At the same time, the well-documented privacy paradox shows that expressed concerns do not always translate into protective behavior, suggesting that perceived benefits, interface design, and effort expectancy mediate these reactions (Jiang et al., 2013; Li, 2012; Barth and de Jong, 2017).

Privacy concerns have been quantified using instruments such as the Concern for Information Privacy (CFIP) and Internet Users' Information Privacy Concerns (IUIPC) measures (Smith et al., 1996; Malhotra et al., 2004), which document their correlation with attitudes and behaviors. Parallel work in information systems has examined assurance cues, including seals, certifications, and trust marks, as visual indicators of security and reliability (Hu et al., 2010; Lins et al., 2024). Static disclosures of this kind are detached from interactive contexts. Contemporary digital environments require disclosures that are interactive, timely, and embedded in user processes such as onboarding or consent.

Transparency is multidimensional, encompassing openness, clarity, and correctness (Schnackenberg and Tomlinson, 2016; Obar, 2022; Dehling and Sunyaev, 2024). It operates through cognitive mechanisms essential to HCI, including perceived control, procedural fairness, and ambiguity reduction (Culnan and Armstrong, 1999; Gierlich-Joas et al., 2022). Two gaps persist. First, most previous research has treated transparency as a direct antecedent of privacy concern or trust (Esmaeilzadeh, 2019). Fewer studies have examined it as a moderator when users become explicitly aware of automated data collection activities, such as web scraping. Second, empirical findings remain inconclusive regarding whether privacy concerns inherently diminish trust, suggesting that compensatory evaluations may emerge under certain interface conditions (Bélanger and James, 2020; Mcknight et al., 2011; Lukyanenko et al., 2022).

This study developed and evaluated a scenario-based onboarding survey that simulated data-use disclosures within a social media environment. The design conceptualizes transparency as an interaction mechanism intended to buffer the effects of heightened awareness of automated data collection. The interface presented participants with either minimal or explicit explanations of web-scraping practices to examine how these disclosures were associated with user responses. The analysis focused on three constructs: awareness of automated data collection, the privacy-protective behavior index, and user trust in the platform.

The objective was to examine whether the transparency vignette was associated with a weaker relationship between awareness of web scraping and reported privacy-protective behavior, and whether that behavior was related to user trust. Accordingly, the following research question was proposed:

- How does data-use transparency moderate the relationship between awareness of web scraping and privacy-protective behavior in an interactive platform context?

To explore this question, three specific objectives guided the study:

- Analyze the association between awareness of automated data collection and the privacy-protective behavior index.
- Test whether the transparency condition is associated with a weaker awareness-behavior relationship.
- Examine how the privacy-protective behavior index is associated with user trust in social-platform environments.

The structure of this paper is as follows. Section 2 reviews related work on web-scraping awareness, privacy-protective behavior, transparency, and trust, situating these constructs within the broader field of HCI. Section 3 describes the methods, detailing the scenario-based onboarding vignette used to simulate the transparency conditions, the participant sampling strategy, and the analytical

procedures. Section 4 reports the results, including the moderating role of the transparency condition in the relationship between awareness and the privacy-protective behavior index and the tested association between the behavior index and trust. Section 5 discusses the contributions to theory and the practical implications. Section 6 summarizes the findings, identifies the limitations of the quasi-experimental design, and outlines directions for future work.

2 RELATED WORK

The increasing prevalence of AI systems that depend on large-scale data collection, particularly through web scraping, has intensified scholarly and public concern regarding privacy, transparency, and ethical data use (Lotfi et al., 2022; Weerasinghe et al., 2023). Web scraping refers to the automated extraction of publicly available information from online sources. Although this practice enables enhanced personalization, targeted advertising, and user engagement (Marengo, 2024), it often occurs without explicit user consent, raising concerns about fairness, informed consent, and data protection (General Data Protection Regulation (GDPR), 2018; Al-Natour et al., 2020; Gallese, 2023). The literature emphasizes that awareness of such automated collection practices elevates privacy concerns by increasing users' perceptions of vulnerability and loss of control (Awad and Krishnan, 2006; Xu et al., 2008). When users perceive limited transparency in data collection, these concerns may intensify, motivating behaviors such as adjusting privacy settings or reducing platform engagement (Crossler and Bélanger, 2019).

2.1 Web-Scraping Awareness and Privacy Perceptions

Web scraping enables automated access to publicly available data, supporting AI-driven functions such as recommendation systems and targeted advertising (Marengo, 2024; Lotfi et al., 2022). Although legally permissible under specific conditions, its covert nature often undermines user autonomy and raises ethical concerns (Al-Natour et al., 2020; General Data Protection Regulation (GDPR), 2018). Empirical research shows that when users become aware of such practices, they report heightened privacy concerns, particularly regarding unauthorized data use and loss of control (Malhotra et al., 2004; Xu et al., 2008; Awad and Krishnan, 2006). Increased awareness can trigger defensive behaviors, including modifying privacy settings or limiting platform engagement (Crossler and Bélanger, 2019).

However, awareness does not always translate into consistent protective behavior. The “privacy paradox” describes this discrepancy between users' expressed concerns and their actual behaviors (Jiang et al., 2013; Barth and de Jong, 2017). While awareness can prompt short-term caution, its long-term influence on attitudes and trust remains uncertain (Bélanger and James, 2020). This underscores the need to examine the mechanisms through which awareness interacts with contextual factors, such as transparency, to shape sustained user responses to automated data collection. The privacy paradox is used here as contextual motivation for studying privacy-protective behavior, but the study does not directly measure inconsistency between stated privacy attitudes and observed behavior.

2.2 Privacy Concerns and Behavioral Outcomes

Privacy concern represents the degree of apprehension individuals experience about the collection and use of personal information (Smith et al., 1996; Malhotra et al., 2004). Studies employing the CFIP and IUIPC frameworks demonstrate that elevated concern predicts privacy-protective

behaviors such as information withholding or withdrawal from digital platforms (Awad and Krishnan, 2006; Crossler and Bélanger, 2019). These actions depend on perceived risk, control, and the clarity of available protective mechanisms (Li, 2012).

Nonetheless, privacy concerns do not uniformly lead to behavioral change. Users often underestimate the extent of data collection or find privacy management cognitively demanding, resulting in limited protective actions (Tsai et al., 2011). The privacy paradox illustrates that users may continue to use services despite privacy apprehensions, particularly when perceived benefits outweigh perceived risks (Culnan and Armstrong, 1999; Sutanto et al., 2013). Moreover, persistent privacy concerns can diminish trust, reduce loyalty, and hinder long-term engagement (Martin, 2018; Mcknight et al., 2011). Yet, this relationship is not strictly linear, suggesting that contextual factors, including transparency and perceived fairness, may mediate or moderate the impact of privacy concerns on trust.

2.3 Transparency as a Mitigating Mechanism

Transparency has been linked to user trust and to perceived control. It reflects the degree to which a system communicates its data practices in an accessible, accurate, and timely manner (Schnackenberg and Tomlinson, 2016; Obar, 2022; Dehling and Sunyaev, 2024). Transparent disclosures support user understanding of data flows and can reduce perceived risks associated with automated data collection (Culnan and Armstrong, 1999; Sutanto et al., 2013; Li, 2012). Under Privacy Calculus Theory, users evaluate the trade-off between perceived risks and benefits, and transparent systems help reduce uncertainty in this evaluation (Li, 2012).

In practice, transparency is implemented through privacy seals, dashboards, or interactive consent mechanisms that visualize how data is collected and used (Hu et al., 2010; Lins et al., 2024). These elements support perceptions of accountability and procedural fairness (Agozie and Kaya, 2021). Their effectiveness depends on usability and comprehension rather than presence alone (Schnackenberg and Tomlinson, 2016; Dehling and Sunyaev, 2024). Empirical studies that examine transparency as a moderating factor, rather than a direct antecedent, remain limited. Examining transparency in this conditional role is informative for understanding how design interventions interact with awareness of automated data collection.

2.4 Trust Formation in Digital Platforms

Trust shapes users' willingness to engage with digital platforms and share personal information. It encompasses beliefs about a system's competence, benevolence, and integrity (Mcknight et al., 2011; Gefen et al., 2003; Lukyanenko et al., 2022). Methodological reviews of trust research in HCI further document that survey-based instruments remain the dominant approach for capturing perceived trust in AI-assisted contexts, while flagging measurement heterogeneity across studies (Vereschak et al., 2021). Transparency directly contributes to these beliefs by demonstrating predictability and openness in data handling (Schnackenberg and Tomlinson, 2016). When users perceive a platform as transparent and consistent, they are more likely to attribute trustworthy intentions and maintain engagement, even in data-intensive contexts (Bélanger and James, 2020; Martin, 2018).

Conversely, opaque data practices can erode trust and trigger disengagement (Awad and Krishnan, 2006; Xu et al., 2008). Still, trust can persist despite privacy concerns when users rely on habitual engagement or perceive a high level of necessity in using the platform (Jiang et al., 2013). Trust and reliance are also conceptually dissociable: users may rely on AI systems that they do not

consider trustworthy, while withholding reliance from systems they nominally trust (Banovic et al., 2023). This distinction informs the interpretation of platform-refusal behavior in the present study as a reliance-related outcome rather than a direct measure of attitudinal privacy concern. These findings highlight the context-dependency of trust formation and the potential role of transparency in stabilizing user confidence amid data opacity.

2.5 Research Gaps

While prior research has extensively examined privacy concerns, transparency, and trust as separate constructs, their interactive dynamics remain underexplored. First, few studies have analyzed whether transparency moderates the relationship between awareness of automated data collection and privacy concern (Bélanger and Crossler, 2011; Crossler and Bélanger, 2019). Second, although transparency is recognized as a key factor in building trust (Schnackenberg and Tomlinson, 2016; Gierlich-Joas et al., 2022), there is limited evidence on how it influences the interpretation of risk in data-intensive environments, such as social media. Third, the direction and strength of the privacy–trust relationship vary across contexts, suggesting the need for models that account for nonlinearity and contextual dependencies (McKnight et al., 2011; Martin, 2018). Finally, most existing studies rely on correlational designs, and scenario-based designs can examine conditional associations under controlled interface descriptions while still requiring caution when assignment is not randomized.

3 METHODS

Building on the literature reviewed above, the following hypotheses connect web-scraping awareness, the transparency condition, the privacy-protective behavior index, and user trust. Prior research links user awareness of data collection practices to higher privacy concern (Malhotra et al., 2004). Greater awareness of automated data extraction, such as web scraping, is associated with higher perceived risk and vulnerability to data misuse, as well as more frequent protective behavior (Crossler and Bélanger, 2019; Al-Natour et al., 2020). Users who understand how automated collection works tend to report stronger concerns about the control and security of their information and to engage in more protective actions (Awad and Krishnan, 2006). The present study captured the behavioral side of this pattern through a 0–2 privacy-protective behavior index. Accordingly, the first hypothesis is stated as follows:

- **H1 (+):** Increased awareness of web scraping is positively associated with privacy-protective behavior (the composite index of adjusting platform privacy settings and refusing to use a platform due to privacy concerns).

Transparency in data management has been identified as a potential moderating factor that may alleviate the adverse outcomes associated with increased awareness of privacy-invasive practices. Transparent communication helps users understand how data is collected, for what purposes, and under what conditions, which can mitigate feelings of uncertainty and reduce privacy-related anxiety (Culnan and Armstrong, 1999; Li, 2012; Dehling and Sunyaev, 2024). When platforms explicitly disclose their data-handling processes and provide meaningful control options, users are more likely to perceive fairness and accountability, which may weaken the link between awareness and concern (Kim et al., 2023). Transparency is expected to moderate the awareness–behavior relationship because disclosure may reduce uncertainty about data practices, reframe perceived risk, and give

users a clearer basis for evaluating whether continued platform use is acceptable. These mechanisms are presented as theoretical motivation for H2 rather than as directly measured mediators in the present study. Based on this reasoning, the second hypothesis is formulated as follows:

- **H2 (-):** The transparency condition moderates the relationship between web-scraping awareness and the privacy-protective behavior index, such that the high-transparency condition is associated with a weaker awareness–behavior relationship.

Existing research also supports the theoretical premise that heightened privacy concerns negatively influence user trust in digital platforms (Martin, 2018; Bélanger and James, 2020). Concerns about data misuse and a lack of clarity in information handling are among the most significant barriers to maintaining trust, particularly in AI-mediated environments where large-scale data collection is routine (Mcknight et al., 2011; Gefen et al., 2003; Lukyanenko et al., 2022). When users perceive uncertainty about how their information is used, confidence in the platform’s ethical integrity and competence diminishes. Reflecting this theoretical foundation, the third hypothesis is stated as follows:

- **H3 (-):** Higher scores on the privacy-protective behavior index are negatively associated with user trust in the platform.

The design follows recommended practices for behavioral research and theory testing in scenario-based HCI studies (Groß, 2021; Mcknight et al., 2011; Peffers et al., 2007). Figure 1 visually depicts the conceptual model and the hypothesized relationships among the constructs. The subsequent section describes the quasi-experimental survey design developed to test these hypotheses.

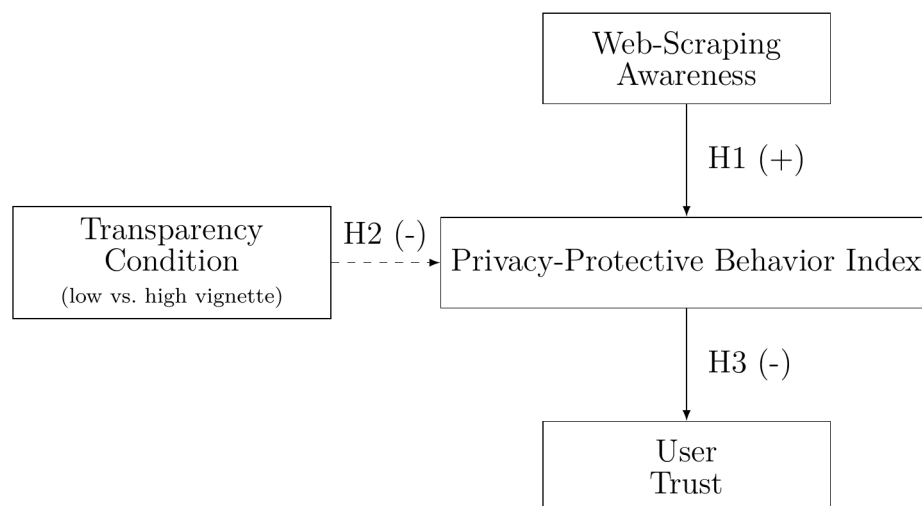


Figure 1. Conceptual model illustrating the hypothesized relationships among web-scraping awareness, the transparency condition, the privacy-protective behavior index, and user trust. The transparency condition is modeled as a moderator of the awareness–behavior relationship.

3.1 Research Design

This study employed a quantitative, scenario-based quasi-experimental survey to examine the relationships among awareness of web scraping, privacy-protective behavior, transparency, and user trust in a social media context. A between-subjects design was implemented, with participants

256 completing one of two survey versions corresponding to distinct transparency conditions during a
 257 simulated onboarding scenario modeled after the Instagram platform.

258 Participants were recruited through university forums, academic mailing lists, and public social
 259 networks. Participation in the survey was entirely voluntary and anonymous. Before beginning,
 260 all participants reviewed an online informed consent statement detailing the study's purpose,
 261 the types of data collected, the intended use of those data, and participants' rights. They
 262 provided explicit electronic informed consent by confirming their agreement on the first page
 263 of the questionnaire. The purpose of the study, data usage, and participants' rights were also
 264 clearly explained. Respondents self-selected into one of the two conditions after neutral instructions
 265 highlighting the equal significance of both options. Consequently, the design should be interpreted
 266 as quasi-experimental rather than randomized, and causal claims are limited.

267 Each survey version began with a brief, realistic vignette describing automated data collection
 268 (i.e., web scraping) during onboarding. In the *low-transparency* condition, practices were described
 269 only minimally, with no justification or user controls (a compliance-oriented baseline). In the *high-*
 270 *transparency* condition, the vignette provided specific explanations of what data was collected,
 271 why it was needed, and which privacy controls were available (user-centered transparency). The
 272 study therefore distinguishes the objective transparency condition, represented by the low- versus
 273 high-transparency vignette, from perceived transparency, measured through participants' responses
 274 after exposure to the vignette. Following the vignette, participants evaluated transparency
 275 during onboarding using items aligned with prior frameworks (Li, 2012; Malhotra et al., 2004;
 276 Schnackenberg and Tomlinson, 2016).

277 No item was explicitly labeled as a manipulation check. Perceived transparency was measured
 278 after exposure to the vignette and was intended to capture participants' subjective evaluation
 279 of disclosure clarity rather than to serve as a standalone manipulation check. Although
 280 no dedicated manipulation-check item was administered, a between-conditions comparison of
 281 perceived transparency serves as a post hoc consistency check for the vignette contrast.
 282 Participants in the high-transparency condition rated perceived transparency as substantially
 283 higher than participants in the low-transparency condition ($M_A = 2.40$, $SD_A = 0.89$, $n_A = 173$;
 284 $M_B = 2.93$, $SD_B = 1.03$, $n_B = 152$; Welch's $t(323) = -4.94$, $p < .001$, Cohen's $d = -0.55$). This
 285 comparison is reported in Supplementary Material (Table S10) and is interpreted as a post hoc
 286 consistency check rather than as a pre-registered manipulation check. The transparency condition
 287 is therefore interpreted in this study as a scenario-based design contrast, and the perceived-
 288 transparency responses are treated as subjective evaluations rather than as independent evidence
 289 that the manipulation was perceived exactly as intended.

290 Unless noted otherwise, items were measured on a 5-point Likert scale (1 = strongly disagree; 5 =
 291 strongly agree). Privacy-related outcomes were operationalized exclusively as two binary (Yes/No)
 292 privacy-protective behaviors: prior adjustment of Instagram privacy settings and prior refusal to use
 293 a platform or deletion of an account due to privacy concerns. These two items were summed to form
 294 a privacy-protective behavior index, with integer values of 0, 1, or 2 (the count of behaviors reported
 295 by each participant). No multi-item attitudinal privacy-concern Likert scale was administered. The
 296 survey assessed (a) awareness of web scraping (with attention checks), (b) the two binary privacy-
 297 protective behaviors and their index, (c) trust in the platform, (d) perceived transparency, and (e)
 298 demographics and controls (e.g., Instagram use frequency, privacy sensitivity). The design follows
 299 recommended practices for testing moderation in¹⁴ behavioral HCI research, with the qualification

that self-selection limits causal inference (Groß [2021] Mcknight et al. [2011] Peffers et al. [2007] Vereschak et al. [2021]).

Table 1 summarizes the constructs, operational variables, definitions, hypotheses, and key sources.

Under the institutional policy applicable at Universidad Panamericana, anonymous, minimal-risk survey research that does not collect identifiable or sensitive personal data is exempt from formal Institutional Review Board (IRB) review. The present study fell within that exempt category, and a formal IRB protocol was therefore not submitted. Before starting the survey, all participants were informed of the study's purpose, the voluntary nature of their participation, the confidentiality of the data collected, and their right to withdraw at any time without adverse repercussions. To proceed with the questionnaire, participants were required to actively confirm their consent on the first page, which was considered as providing electronic informed consent.

3.1.1 Web Scraping Awareness

This construct represents users' familiarity with automated data extraction methods. In this study, automated data collection is used as the broader category, while web-scraping awareness refers to the specific form of data-collection awareness measured. Awareness was treated as a focal independent variable, grounded in IUIPC and Privacy Calculus perspectives. Following Bélanger and Crossler (2011) and Phelps et al. (2000), three items captured understanding of what web scraping is, how firms use it, and its privacy implications.

3.1.2 Privacy-Protective Behavior

Privacy-related outcomes were captured exclusively through two binary self-report indicators: prior adjustment of Instagram privacy settings, and prior refusal to use a platform or deletion of an account due to privacy concerns. These two indicators were summed into a privacy-protective behavior index, taking integer values 0, 1, or 2, corresponding to the number of protective behaviors a participant reported (0 = neither behavior reported, 1 = one of the two behaviors reported, 2 = both behaviors reported). The instrument did not include a multi-item attitudinal privacy-concern Likert scale, and the term "privacy concern" is therefore used in this study to refer to the behavioral manifestations captured by this index, consistent with Smith et al. (1996) and Malhotra et al. (2004) but with the caveat that behavioral indicators do not substitute for an attitudinal measure (Banovic et al. [2023]). The behavior index captures low-barrier privacy-protective actions, meaning self-reported actions that participants can enact with minimal technical effort, such as adjusting platform privacy settings or refusing to use a platform; the indicators are self-reported rather than directly observed.

3.1.3 User Trust

This construct reflects users' beliefs in a platform's integrity, competence, and benevolence. It was measured using four items adapted from Mcknight et al. (2011) and Gefen et al. (2003), focusing on responsibility, honesty, fairness, and reliability in handling personal data. A sensitivity analysis using a broader composite that combines the four trust items with the three perceived-transparency items is reported in the Supplementary Material; both operationalizations yield consistent results in terms of direction and significance.

Table 1. Definitions of Constructs and Supporting Literature

Construct	Operational Variables	Definition	Hypothesis	Supporting Literature
Web Scraping Awareness	Three Likert items measuring knowledge of what web scraping is, how firms use it, and its privacy implications.	Subjective familiarity with automated data extraction methods, such as web scraping.	H1, H2	Awareness as a precursor to privacy behavior (Bélanger and Crossler, 2011); consumer knowledge and disclosure concerns (Phelps et al., 2000).
Privacy-Protective Behavior	Two binary indicators (adjustment of Instagram privacy settings; refusal to use a platform or deletion of an account due to privacy concerns) were summed to form a behavior index that takes integer values of 0, 1, or 2 (0 = neither, 1 = one, 2 = both).	Privacy-related actions reported by participants. The construct is behavioral rather than attitudinal because no multi-item Likert concern scale was administered.	H1, H2, H3	Multidimensional concerns (Smith et al., 1996); concern as a driver of trust and behavior (Malhotra et al., 2004); reliance and refusal as dissociable from trust (Banovic et al., 2023).
User Trust	Four Likert items cover responsible handling, honesty, acting in users' best interest, and willingness to rely.	Beliefs about a platform's competence, integrity, and benevolence in managing personal data.	H3	Trust components (Mcknight et al., 2011); trust and usage intention (Gefen et al., 2003).
Perceived Transparency	Low- versus high-transparency condition; perceived disclosure, policy clarity, and control affordance	Objective scenario conditions and subjective perceptions of how clearly, completely, and credibly the platform communicated its data practices during onboarding.	H2	Multidimensional transparency (Schnackenberg and Tomlinson, 2016).

Note. IUIPC = Internet Users' Information Privacy Concerns; CFIP = Concern for Information Privacy.

3.1.4 Perceived Transparency (Moderator)

Transparency was operationalized at two levels: the scenario condition represented the objective low- versus high-transparency disclosure, while perceived transparency captured

participants' subjective evaluation of that disclosure. Items reflected disclosure (availability), clarity (comprehensibility), and accuracy (truthfulness) per Schnackenberg and Tomlinson (2016). The full set of perceived-transparency items is reported in Appendix A.

Control variables included age, gender, frequency of Instagram use, familiarity with privacy policies, and history of privacy-related disengagement, given their documented influence on privacy and trust (Bansal et al., 2010; Xu et al., 2012).

3.2 Data Collection

The survey (Google Forms) remained open for four weeks. After providing informed consent, participants selected a condition, viewed the onboarding vignette, and completed the questionnaire. Because condition selection was voluntary, group comparisons are interpreted with caution. Responses were anonymous and exported for analysis in Python and SPSS. The final dataset comprised valid, complete cases after excluding failures on attention checks and patterned responses. The sample primarily consisted of university students and early-career professionals aged 18–35, many of whom reported daily Instagram use, which supports contextual relevance to the Instagram-like scenario but limits generalization to broader populations and platform types. A demographic overview appears in Table 2.

Table 2. Participant Demographics by Age and Gender

Age Group	Female	Male	Non-binary	Prefer not to say
<20	4	5	1	0
20–25	44	37	1	3
26–30	60	45	3	3
31–35	42	14	1	4
36–40	20	11	1	1
41–50	7	5	0	2
51–60	6	4	0	0
61+	0	1	0	0
Total	183	122	7	13

The sample composition should be considered when interpreting the findings. Because participants were drawn from a specific population, concentrated among university students and early-career professionals aged 18–35, and self-selected into the study conditions, the results should not be generalized to all social-media users without caution. Generalization to older populations, non-student populations, and platforms with different data practices is limited.

3.3 Data Analysis

Hypotheses were tested using a set of regression models matched to the measurement scale of each dependent variable. Ordinary least squares (OLS) regression was used for the continuous user-trust composite and for the 0/1/2 privacy-protective behavior index, which is treated here as an integer-valued count of the two binary behaviors reported by each participant. OLS was retained for the index so that the slope of awareness represents the change in the expected count of protective behaviors per unit increase in awareness, expressed on the same metric as the index. The known

limits of fitting OLS to a bounded 0/1/2 outcome, including predictions that can fall outside the observed support of the index, are discussed in Section 3.3.3 and listed in the Limitations. Logistic regression was used for each of the two individual binary behaviors. The transparency condition was dummy-coded (0 = low transparency; 1 = high transparency); the interaction term used mean-centered awareness multiplied by the condition indicator. Equations for the four estimated models are reported below.

The bivariate OLS model for H1, in which the privacy-protective behavior index is regressed on web-scraping awareness, is shown in Equation 1

$$BehaviorIndex_i = \beta_0 + \beta_1 Awareness_i + \varepsilon_i. \quad (1)$$

Each of the two binary behaviors composing the index was modeled as a separate logistic regression on web-scraping awareness, as shown in Equations 2 and 3

$$\text{logit}[\text{Pr}(AdjustedSettings_i = 1)] = \beta_0 + \beta_1 Awareness_i. \quad (2)$$

$$\text{logit}[\text{Pr}(RefusedPlatform_i = 1)] = \beta_0 + \beta_1 Awareness_i. \quad (3)$$

The moderation model for H2, used to test whether the transparency condition attenuates the awareness-behavior association, is shown in Equation 4

$$BehaviorIndex_i = \beta_0 + \beta_1 Awareness_i + \beta_2 Transparency_i + \beta_3 (Awareness_i \times Transparency_i) + \gamma^T \mathbf{X}_i + \varepsilon_i, \quad (4)$$

where $\mathbf{X}_i$ denotes the vector of control variables. A significant negative β_3 supports H2, indicating that the high-transparency condition is associated with a weaker positive relationship between awareness of web scraping and the privacy-protective behavior index.

The OLS model for H3, in which user trust is regressed on the privacy-protective behavior index, is shown in Equation 5

$$Trust_i = \beta_0 + \beta_1 BehaviorIndex_i + \varepsilon_i. \quad (5)$$

For the continuous user-trust composite, OLS assumptions were evaluated via standard diagnostics: linearity (residuals vs fitted values), normality (Q-Q plots, Shapiro-Wilk), homoscedasticity (residual dispersion), and multicollinearity (variance inflation factor [VIF] and condition number); no violations of substantive concern were detected. For the 0/1/2 privacy-protective behavior index, residual-normality diagnostics such as Q-Q plots and the Shapiro-Wilk test are not informative because the outcome has only three possible values; the residuals therefore cannot be approximately normal by construction. Inference for the OLS slope on the index relies on the reported standard errors and the large-sample distribution of the coefficient estimator rather than on residual-normality diagnostics, and the modeling caveats associated with applying OLS

to this bounded count are reported in Section 3.3.3. Analyses were conducted in Python (using Statsmodels), with robustness checks performed in SPSS. Two-tailed tests used $\alpha = .05$.

3.3.1 Multiplicity Considerations

The three primary hypothesis tests (H1: OLS on the behavior index; H2: interaction; H3: OLS on trust) target distinct theoretical claims, each with different dependent variables and sets of predictors, and were pre-specified by the conceptual model rather than selected from a larger pool of candidate tests. Consistent with standard practice for a small set of pre-specified confirmatory tests in moderated regression, no familywise correction was applied to these three primary coefficients. For the two logistic models that decompose the behavior index into its components, the awareness coefficient was evaluated against a Bonferroni-corrected threshold within the H1 family (three tests: index OLS, settings logistic, refusal logistic; $\alpha' = .05/3 \approx .0167$). Below that threshold, the associations with adjusting privacy settings ($p = .003$) and the behavior index ($p = .006$) remain significant, whereas the association with platform refusal ($p = .196$) does not. The supplementary group-difference *t*-tests reported in Section 4 (three contrasts on awareness, trust, and the behavior index) were also evaluated at $\alpha' = .05/3 \approx .0167$; the awareness and behavior-index contrasts remain significant ($p < .001$), and the trust contrast remains significant ($p = .006$) under that threshold. The Pearson correlations reported in Section 4 (three pairs) were similarly evaluated at $\alpha' = .05/3 \approx .0167$; the awareness–trust correlation ($r = .36$, $p < .001$) and the awareness–behavior correlation ($r = .18$, $p = .001$) remain significant under the corrected threshold, while the behavior–trust correlation ($r = -.06$, $p = .25$) does not meet the corrected threshold and is reported as descriptive only.

3.3.2 Scope of the Reported Analyses

The estimates reported throughout this study correspond to the fitted-model summaries (coefficients, standard errors, *t/z* statistics, *p*-values, 95% confidence intervals, and R^2 or pseudo- R^2 values) from the analysis, and the statistical validity of the reported tests rests on those standard inferential statistics. Several diagnostic and robustness analyses complement the fitted models reported above. The simple slope of awareness at high transparency is reported with its full standard error in Section 4. Exact variance inflation factors for the moderation model are reported in Section 4.4.1. The between-conditions comparison of perceived transparency with Cohen's *d* is reported in Section 4 and in the Methods (manipulation-check paragraph). Per-group sample sizes are reported in Section 4 alongside the independent-samples *t*-tests. A bivariate (joint) logistic regression on the two binary behaviors is reported in Section 4.4. Poisson and ordinal-logistic specifications for the privacy-protective behavior index are reported in Section 4.4 as robustness checks for alternative specifications. Full tables for these analyses are provided in the Supplementary Material (Tables S1–S10), and the dataset is referenced in the Data Availability statement.

3.3.3 Model Specification: Scope and Limits

Three modeling choices for the privacy-protective behavior index warrant explicit comment. First, the index is the simple sum of two binary indicators that record qualitatively different actions (adjusting Instagram privacy settings, refusing to use a platform, or deleting an account). The sum is not intended to assert that the two actions are exchangeable units of the same underlying behavior, but rather to provide a parsimonious count of how many of the two protective behaviors a participant reported. The same two binary outcomes are also reported separately as logistic

regressions (Equations 2 and 3) so that the relationship between awareness and each specific behavior can be inspected on its own probability scale alongside the OLS slope on the index.

Second, OLS on a bounded 0/1/2 outcome is a linear approximation rather than a generative model for the count. Predicted values from the OLS slope can fall outside the observed support of the index, and the interpretation of the slope is limited to the local change in the expected count per unit increase in awareness. A model that respects the support of the outcome, such as a Poisson regression on the count, an ordinal logistic regression on the ordered categories 0–1–2, or a bivariate (joint) logistic regression on the two correlated binary indicators, is a more appropriate specification. Poisson and ordinal-logistic specifications for the index are reported as alternative-specification robustness checks in Section 4.4 and Supplementary Table S9. The direction and significance of the awareness coefficient under those specifications are consistent with the OLS estimate on the index, which is reported here for interpretive parsimony.

Third, the two univariate logistic regressions on the individual binary behaviors do not model the within-participant correlation between the two outcomes, and the OLS coefficient on the index is on a different metric (expected count) from the logistic coefficients (log-odds of a single behavior). The two univariate specifications are reported as complementary views of the same H1 question rather than as a single coherent estimator. A bivariate (joint) logistic specification on the pair of binary outcomes, fitted as a generalized estimating equation (GEE) with an exchangeable working correlation that places both outcomes on a common log-odds metric and accounts for their within-participant correlation, is reported as a robustness analysis in Section 4.4 and Supplementary Table S4. Its awareness coefficient is consistent in direction and significance with the two univariate logistic regressions.

3.3.4 Timing of Measures

All measured constructs were captured in a single survey wave administered after exposure to the assigned onboarding vignette. The two privacy-protective behavior indicators were retrospective self-reports of prior actions taken by the participant (“Have you ever adjusted your Instagram privacy settings?” and “Have you ever refused to use a platform or deleted an account due to privacy concerns?” in Appendix A), and therefore refer to behaviors that occurred before the survey. The web-scraping awareness, perceived transparency, and user-trust items refer to the participant’s beliefs and evaluations at the time of the survey, that is, after the vignette had been shown. The behavior → trust association reported under H3 is therefore an association between retrospective behavioral self-reports and post-vignette trust beliefs measured in the same wave; the design does not establish temporal precedence between the two within the study itself, and the constraint is repeated in the Limitations.

4 RESULTS

This section presents the study’s statistical outcomes, including descriptive statistics, reliability testing, correlation analyses, and regression results for the hypothesized relationships.

4.1 Descriptive Analysis of the Sample

Descriptive statistics were computed to examine the central tendencies and variability of all key constructs. The final sample consisted of 325 participants, primarily university students and early-career professionals aged 18 to 35. Although demographic variables (e.g., gender, platform usage

frequency) were collected as controls, they were not central to the theoretical model and were therefore excluded from primary analyses.

Table 3 presents the descriptive statistics for the continuous constructs and the 0–2 privacy-protective behavior index. Web Scraping Awareness showed moderate levels across the sample ($M = 3.14$, $SD = 1.10$, 1–5 scale), suggesting general familiarity with automated data collection practices. User Trust showed moderate levels ($M = 2.49$, $SD = 0.97$, 1–5 scale). The privacy-protective behavior index, computed as the sum of two binary indicators, ranged from 0 to 2 and was approximately centered between the two condition means reported in Table 9. The scenario-based moderator, Transparency Condition (coded 0 = low; 1 = high), was used to segment participants in subsequent analyses.

Table 3. Descriptive Statistics for Continuous Constructs and the Behavior Index

Construct	Range	Mean	Standard Deviation
Web Scraping Awareness	1–5	3.14	1.10
User Trust	1–5	2.49	0.97
Privacy-Protective Behavior Index	0–2	0.87	0.76

Note. Means and standard deviations are reported for the full sample ($N = 325$); the behavior-index values are the count of two binary protective behaviors per participant.

Internal consistency reliability was assessed using Cronbach's Alpha for each multi-item scale (Table 4). Web Scraping Awareness ($\alpha = 0.787$), User Trust ($\alpha = 0.847$), and Perceived Transparency ($\alpha = 0.817$) demonstrated strong reliability. The privacy-protective behavior index ($\alpha = 0.305$) had a low coefficient because it consists of only two binary items, which severely limits its item variance. Cronbach's α is reported here for completeness but should not be interpreted as the reliability of a reflective attitudinal scale, because no attitudinal items were administered. The index is treated as a formative count of two distinct behaviors rather than as a homogeneous latent construct. Factor-analytic and discriminant validity tests were not conducted; construct distinction is supported by theory and operational definitions rather than by additional validation of latent variables.

Table 4. Cronbach's Alpha for Multi-Item Constructs and for the Two-Item Behavior Index

Construct	Number of Items	Cronbach's Alpha
Web Scraping Awareness	3 (Likert)	0.787
User Trust	4 (Likert)	0.847
Perceived Transparency	3 (Likert)	0.817
Privacy-Protective Behavior Index	2 (binary)	0.305

Note. The two binary items composing the behavior index are not a reflective Likert scale; the reported α reflects the limited item variance associated with binary indicators rather than the reliability of an attitudinal scale.

Figure 2 presents the tested research model with standardized coefficients and significance levels. The analysis revealed significant relationships among web scraping awareness, the transparency condition, and the privacy-protective behavior index, whereas the link between the behavior index and trust was not significant. Specifically, H1 and H2 were supported, whereas H3 was not.

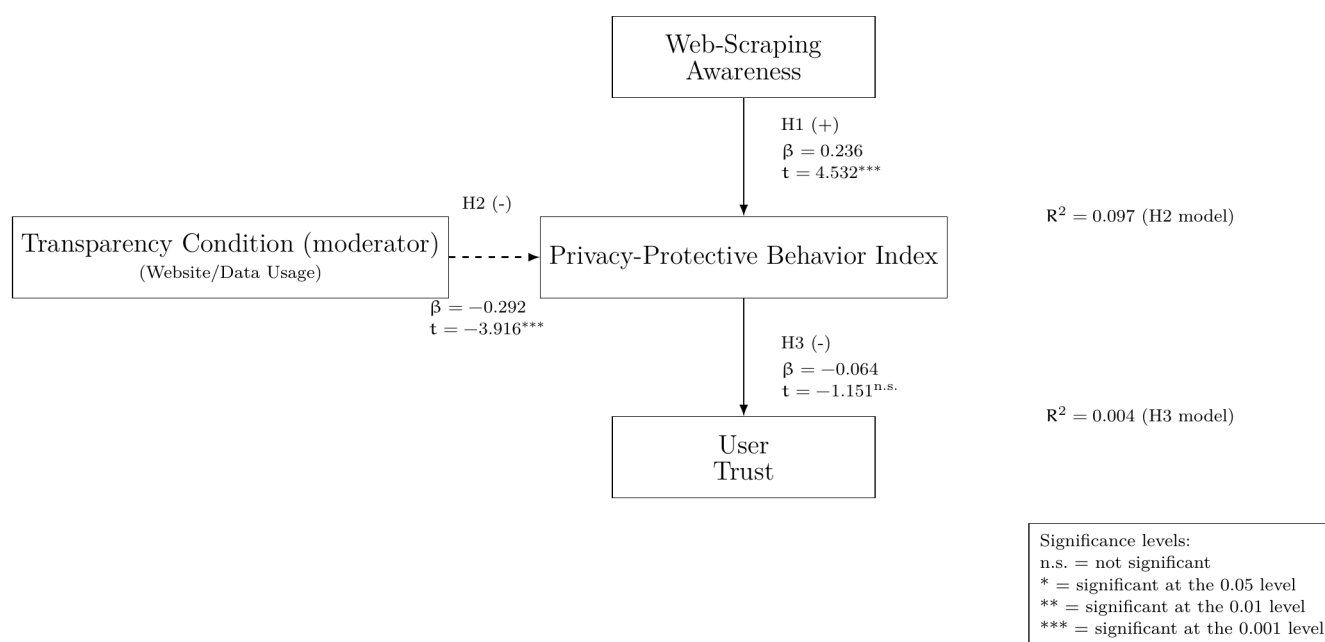


Figure 2. Tested research model with standardized path coefficients and significance levels. H1 and H2 were supported; H3 (privacy-protective behavior $\rightarrow$ user trust) was not supported. Paths are reported as conditional associations because condition assignment was self-selected and baseline awareness differed across conditions. Path coefficients are reported from the moderation model in Table 6; the bivariate H1 model is reported separately in Table 5.

4.2 Hypothesis Testing and Empirical Findings

4.2.1 H1: Web Scraping Awareness and Privacy-Protective Behavior

H1 posited that higher awareness of web scraping would predict more frequent privacy-protective behavior. Because the instrument captured this construct only through two binary indicators (adjusting privacy settings; refusing platform use), H1 was evaluated with one OLS regression on the 0–2 behavior index and one logistic regression for each of the two binary behaviors. The dependent variable in the OLS model was the privacy-protective behavior index (range 0–2). The logistic models examined the likelihood of (a) adjusting privacy settings and (b) refusing or deleting a platform.

The results shown in Table 5 indicate that web scraping awareness was significantly associated with the privacy-protective behavior index ($\beta = 0.16$, $p = 0.006$). Participants with higher awareness were more likely to adjust their privacy settings (logistic regression, $B = 0.33$, $p = 0.003$), although awareness did not significantly predict platform refusal ($B = 0.15$, $p = 0.196$). The findings partially support H1, with an association with low-barrier protective actions but not with higher-effort disengagement behaviors.

In addition, Figure 3 shows the logistic regression of web-scraping awareness on the binary indicator of privacy-settings adjustment, indicating a positive association between greater awareness and a higher likelihood of reporting this protective behavior.

Table 5. Regression Results Predicting Privacy-Protective Behaviors from Web Scraping Awareness

Predictor	B	SE	β	t/z	p	95% CI [LL, UL]
<i>OLS Regression (Behavior Index)</i>						
Constant	0.58	0.14	–	4.17***	<0.001	[0.31, 0.85]
Web Scraping Awareness	0.11	0.04	0.16	2.78**	0.006	[0.03, 0.20]
Model $R^2 = 0.03$; Adjusted $R^2 = 0.02$						
<i>Logistic Regression (Adjusted Privacy Settings)</i>						
Constant	-0.83	0.38	–	-2.20*	0.028	[-1.57, -0.09]
Web Scraping Awareness	0.33	0.11	–	2.93**	0.003	[0.11, 0.55]
Pseudo $R^2 = 0.02$						
<i>Logistic Regression (Refused Platform Use)</i>						
Constant	-0.92	0.38	–	-2.40*	0.016	[-1.67, -0.17]
Web Scraping Awareness	0.15	0.11	–	1.29	0.196	[-0.08, 0.37]
Pseudo $R^2 < 0.01$						

Note. Dependent variables: privacy-protective behavior index (OLS); adjusted privacy settings and refused platform use (logistic).
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

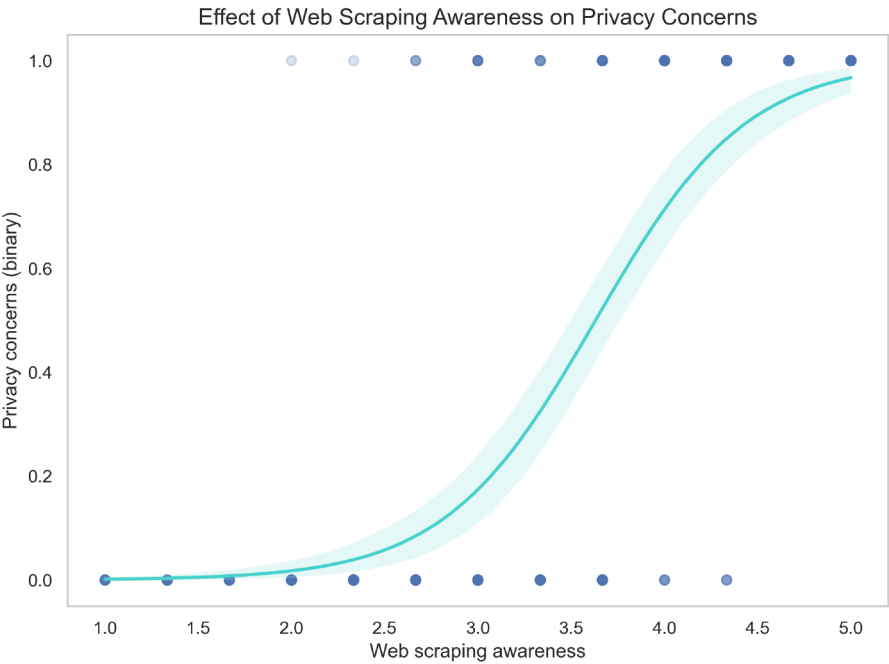


Figure 3. Logistic regression of web-scraping awareness on the binary indicator of having adjusted Instagram privacy settings.

520 4.2.2 H2: Moderating Effect of the Transparency Condition

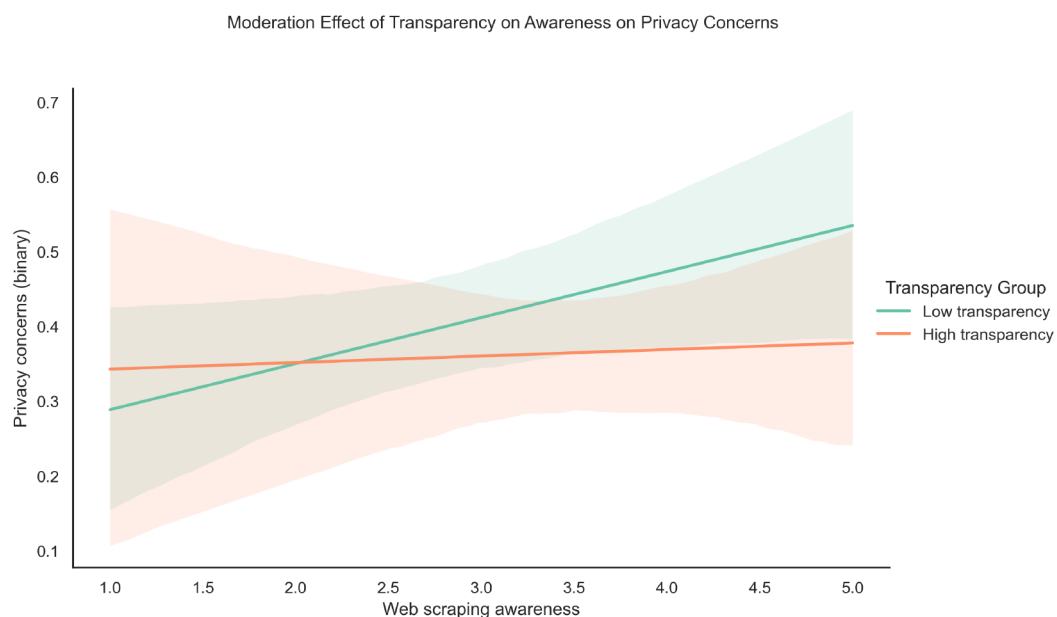
521 H2 proposed that the transparency condition moderates the relationship between awareness and
522 the privacy-protective behavior index. The OLS model, incorporating both main effects and their
523 interaction, was significant ($p < 0.001$), with an adjusted $R^2 = 0.09$. The interaction term was
524 negative and statistically significant ($\beta = -0.24$, $p < 0.001$), supporting the proposed moderation
525 pattern.

Table 6. Moderation Analysis: Privacy-Protective Behavior Index from Web-Scraping Awareness and Transparency Condition

Predictor	B	SE	β	t	p	95% CI [LL, UL]
Constant	0.04	0.16	–	0.23	.818	[-0.28, 0.36]
Web Scraping Awareness	0.24	0.05	0.26	4.53***	<0.001	[0.13, 0.34]
Transparency Group	1.17	0.25	0.27	4.70***	<0.001	[0.68, 1.66]
Awareness $\times$ Transparency	-0.29	0.08	-0.24	-3.92***	<0.001	[-0.44, -0.15]

Note. Dependent variable: privacy-protective behavior index (0–2). Transparency is coded as 0 (low) and 1 (high). * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Figure 4 illustrates the interaction pattern. Simple slopes can be read directly from Table 6. Because the transparency condition is coded 0 = low and 1 = high, the awareness coefficient in the model corresponds to the simple slope of awareness on the privacy-protective behavior index at low transparency. This slope was significantly positive ($B = 0.24$, $SE = 0.05$, $t = 4.53$, $p < 0.001$, 95% CI [0.13, 0.34]). The simple slope at high transparency is the sum of the awareness coefficient and the interaction coefficient: $0.24 + (-0.29) = -0.05$. Using the full coefficient covariance matrix of the moderation model, the standard error of the implied high-transparency simple slope is $SE = 0.054$, yielding $t(321) = -0.90$, $p = .371$, 95% CI [-0.15, 0.06]; the slope at high transparency is therefore not significantly different from zero. The contrast between the two slopes is identical to the interaction coefficient in Table 6 ($B = -0.29$, $SE = 0.08$, $t = -3.92$, $p < 0.001$), which confirms that the two slopes differ significantly. The pattern indicates that the association between awareness and privacy-protective behavior is substantially weaker, and not significantly different from zero in direction, under the high-transparency condition.

**Figure 4.** Predicted privacy-protective behavior index as a function of web-scraping awareness, by transparency condition.

4.2.3 H3: Privacy-Protective Behavior and User Trust

H3 hypothesized a negative association between the privacy-protective behavior index and user trust. As shown in Table 7, the regression coefficient was negative but not significant ($\beta = -0.06$, $p = .250$), and the model explained negligible variance in trust ($R^2 = 0.004$). Thus, H3 was not supported. As shown in Figure 5, the slope of the regression line between the behavior index and user trust is nearly flat, visually reinforcing the absence of a meaningful association.

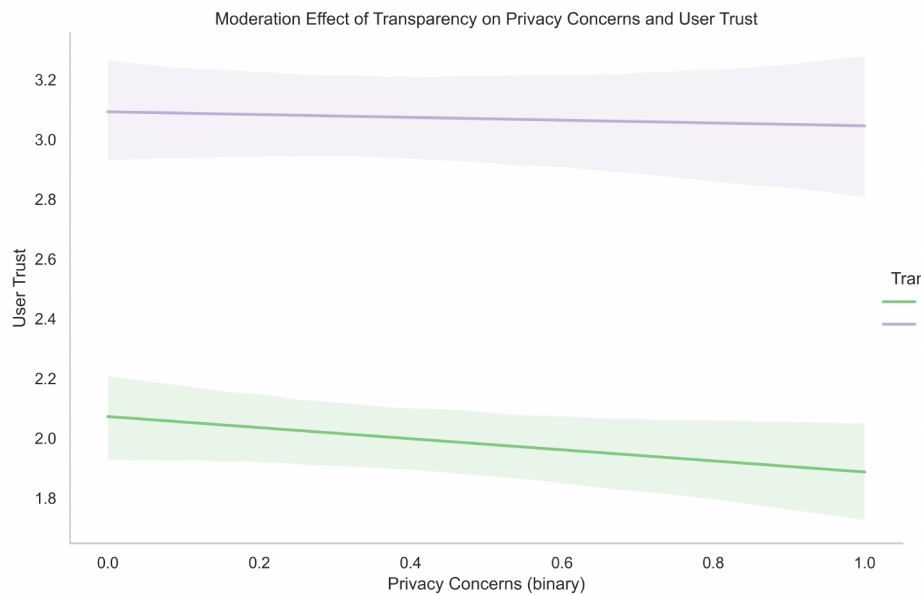


Figure 5. Relationship between the privacy-protective behavior index and user trust.

Table 7. Regression Predicting User Trust from the Privacy-Protective Behavior Index

Predictor	B	SE	β	t	p	95% CI [LL, UL]
Constant	2.56	0.08	—	31.22***	<0.001	[2.40, 2.73]
Privacy-Protective Behavior Index	-0.08	0.07	-0.06	-1.15	.250	[-0.22, 0.06]

Note. Dependent variable: user trust score. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Lastly, Table 8 summarizes the outcomes of all hypotheses. H1 and H2 were supported in the tested models, indicating that awareness was associated with the privacy-protective behavior index and that the transparency condition moderated this relationship. H3 was not supported, indicating that the behavior index did not significantly predict user trust.

4.3 Baseline Comparisons and Group Differences

Because participants self-selected into transparency conditions, the between-condition comparisons reported in this subsection are interpreted as group differences rather than as evidence of the manipulation’s causal effect. Table 9 presents independent-samples t-tests for the three focal variables (awareness, trust, and the privacy-protective behavior index) across the low-transparency (Group A) and high-transparency (Group B) conditions, together with Cohen’s d effect sizes.

Table 8. Summary of Hypotheses Testing Results

Hypothesis	Statement	Result	Interpretation
H1	Increased awareness of web scraping is positively associated with the privacy-protective behavior index.	Supported	Greater awareness of automated data collection was associated with higher scores on the 0–2 privacy-protective behavior index, driven primarily by adjusting privacy settings.
H2	The transparency condition moderates the relationship between web-scraping awareness and the privacy-protective behavior index.	Supported	The high-transparency condition weakened the awareness–behavior link in the moderation model; the two simple slopes differed significantly.
H3	Higher scores on the privacy-protective behavior index negatively predict user trust.	Not supported	The behavior index did not significantly predict user trust in the tested model.

Web-scraping awareness differed significantly between conditions ($M_A = 2.93$, $M_B = 3.35$, $t(323) = -3.44$, $p = .001$, Cohen's $d = -0.39$). Because awareness is the focal independent variable in H1 and H2, this difference indicates baseline non-equivalence on the predictor of primary interest and is consistent with selection effects associated with voluntary condition assignment. The interpretation of H2 should therefore be read as a conditional association rather than as evidence that the high-transparency vignette alone produced the observed difference in the awareness–behavior slope. User trust ($t(323) = -2.80$, $p = .006$, $d = -0.32$) and the privacy-protective behavior index ($t(323) = -3.53$, $p < .001$, $d = -0.39$) were also higher in the high-transparency condition, but, given the baseline non-equivalence on awareness, these differences cannot be unambiguously attributed to the vignette condition.

Baseline equivalence on age, gender, frequency of Instagram use, familiarity with privacy policies, and perceived transparency is reported in Table 9 alongside the focal variables. Conditions did not differ on age ($t(323) = -0.62$, $p = .54$, $d = -0.07$), gender ($\chi^2(3) = 0.88$, $p = .83$), or frequency of Instagram use ($t(323) = 1.50$, $p = .14$, $d = 0.17$), supporting baseline equivalence on these characteristics. Conditions differed in familiarity with privacy policies ($t(323) = -3.17$, $p = .002$, $d = -0.35$), with participants in the high-transparency condition reporting greater familiarity. Perceived transparency was higher in the high-transparency condition (Welch's $t(323) = -4.94$, $p < .001$, $d = -0.55$), as reported in the manipulation-check paragraph of the Methods. The baseline non-equivalence on awareness and on policy familiarity, together with the manipulation-consistent difference in perceived transparency, is incorporated into the interpretation of H2 in the Discussion.

4.4 Diagnostic and Robustness Analyses

This subsection reports the diagnostic and robustness analyses that complement the primary regression models. Full tables are provided in Supplementary Material (Tables S4, S8, and S9).

Table 9. Independent Samples t-Tests Between Transparency Conditions, with Cohen's d

Variable	Mean (A)	Mean (B)	t	p	Cohen's d
Awareness	2.93	3.35	-3.44	0.0007	-0.39
User Trust	2.35	2.66	-2.80	0.0055	-0.32
Privacy-Protective Behavior Index	0.73	1.03	-3.53	0.0005	-0.39
Perceived Transparency	2.40	2.93	-4.94	<0.001	-0.55
Age (years)	29.41	29.93	-0.62	0.539	-0.07
Instagram-use frequency (0–5)	3.78	3.66	1.50	0.136	0.17
Policy familiarity (1–5)	2.42	2.82	-3.17	0.002	-0.35

Note. Group A = low-transparency condition; Group B = high-transparency condition. Welch's t tests on the per-group sample sizes ($n_A = 173$, $n_B = 152$); Cohen's d is reported with the sign of the mean difference (Group A minus Group B), so negative values indicate higher means in the high-transparency condition. Gender by condition did not differ ($\chi^2(3) = 0.88$, $p = .83$). Sensitivity analyses using a broader seven-item composite that combines the trust items with the three perceived-transparency items are reported in Supplementary Table S3.

4.4.1 Variance Inflation Factors for the Moderation Model

Variance inflation factors (VIFs) for the predictors in the moderation model (Equation 4) are well below conventional thresholds for problematic collinearity: $VIF(Awareness_c) = 2.02$, $VIF(Condition) = 1.04$, $VIF(Awareness_c \times Condition) = 1.99$. The intercept-only condition number is 2.95. Multicollinearity is not a concern for the reported moderation estimates.

4.4.2 Bivariate Logistic Specification for the Two Binary Behaviors

A bivariate (joint) logistic specification was estimated by fitting a generalized estimating equation (GEE) with a binomial family, a logit link, and an exchangeable working correlation over the two binary behavioral outcomes per participant. The GEE models adjust privacy settings and refuse the platform, using a common log-odds metric and directly estimating the within-participant correlation between the two behaviors. The awareness coefficient remained positive and significant ($B = 0.26$, $SE = 0.08$, $z = 3.07$, $p = .002$, 95% CI [0.09, 0.42]), consistent in direction and significance with the two univariate logistic regressions reported under H1. A behavior-type indicator (settings vs refusal) was negative and significant ($B = -0.41$, $SE = 0.15$, $z = -2.80$, $p = .005$), reflecting the lower base rate of platform refusal. The estimated within-participant correlation was small ($\hat{\rho} = 0.17$).

4.4.3 Alternative Specifications for the 0/1/2 Behavior Index

Two alternative specifications were estimated for the privacy-protective behavior index: a Poisson regression on the count and an ordinal-logistic regression on the ordered categories 0–1–2. The awareness coefficient was positive and significant in both: Poisson $B = 0.14$, $SE = 0.06$, $z = 2.59$, $p = .010$, 95% CI [0.04, 0.25]; ordinal logistic $B = 0.31$, $SE = 0.10$, $z = 3.20$, $p = .001$, 95% CI [0.12, 0.50]. The Poisson moderation model ($Awareness_c$, $Condition$, $Awareness_c \times Condition$) also reproduces the interaction pattern reported under H2: awareness $B = 0.32$, $p < .001$; condition $B = 0.34$, $p = .005$; interaction $B = -0.37$, $p < .001$. The direction and significance of the awareness effect, and of the moderation pattern, are robust to the choice of specification.

4.5 Additional Correlational Analyses

A Pearson correlation matrix was computed to explore the relationships among the constructs further, and is presented in Figure 6. Awareness and the privacy-protective behavior index were

weakly correlated ($r = 0.18$), whereas awareness and trust were moderately positively correlated ($r = 0.36$). The behavior index and trust were weakly and negatively correlated ($r = -0.06$). These correlations are reported descriptively without multiplicity correction.

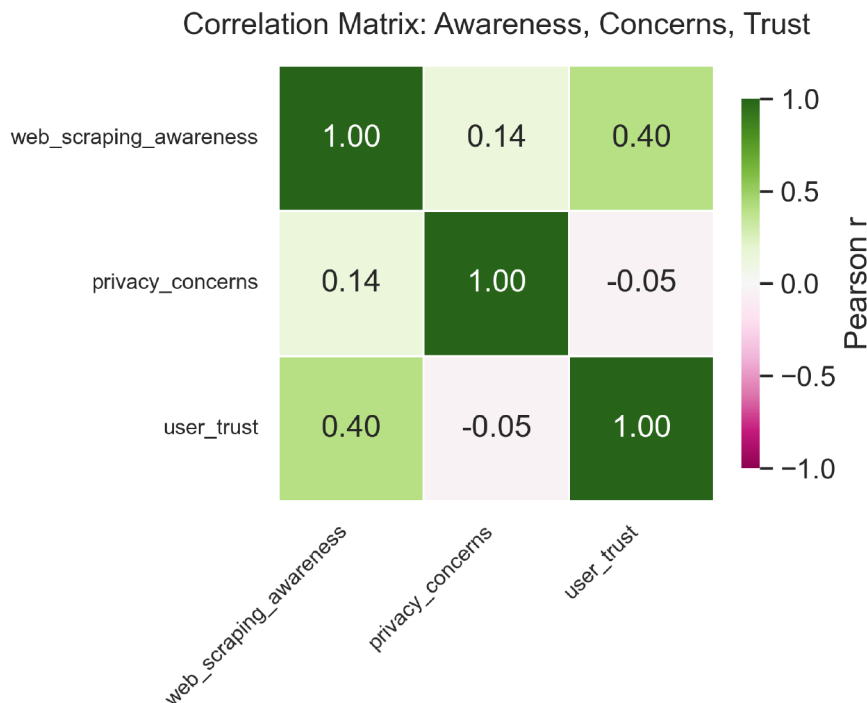


Figure 6. Pearson correlation matrix among primary constructs.

Demographic analyses (Figures 7, 10) indicate that most participants were between 20 and 35 years old, used Instagram daily, and were only moderately familiar with its privacy policies. These distributions contextualize the findings: awareness is common, but a deep understanding of policy remains limited.

5 DISCUSSION

This study examined how awareness of web scraping and the transparency vignette condition in onboarding disclosures were associated with privacy-protective behavior and user trust. The discussion interprets these associations in light of established privacy and trust frameworks.

5.1 Contributions to Theory

The findings refine the way awareness, transparency, and behavioral privacy outcomes are connected. Awareness of automated data collection was associated with privacy-protective behavior; the transparency condition moderated that association; and the behavior index did not predict user trust in this sample. None of these associations can be interpreted causally, because condition assignment was self-selected and the predictor of primary interest (awareness) was not equivalent across conditions at baseline.

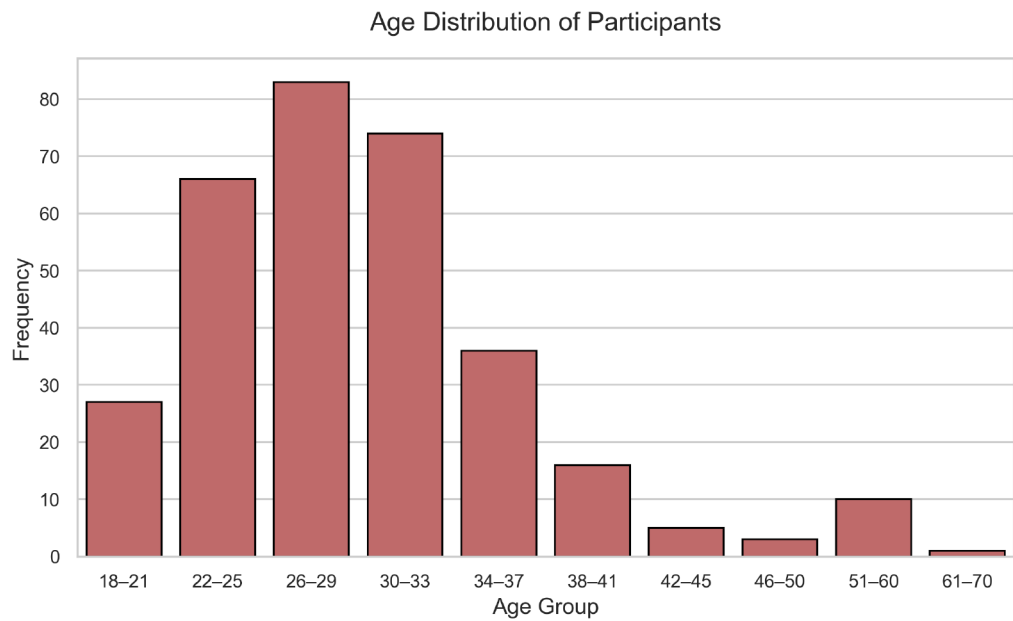


Figure 7. Age distribution of participants.

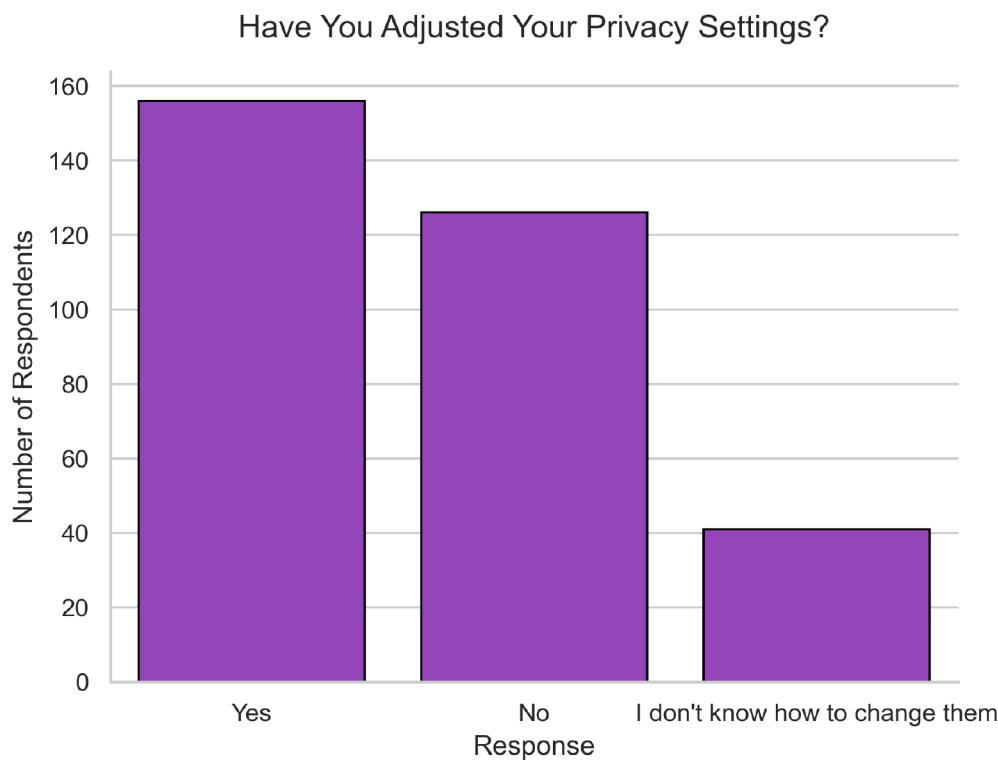


Figure 8. Distribution of privacy settings adjustments.

624 First, the IUIPC model (Malhotra et al., 2004) frames awareness as a cognitive antecedent of
625 concern. The present results extend this account by showing that, at the behavioral level, awareness

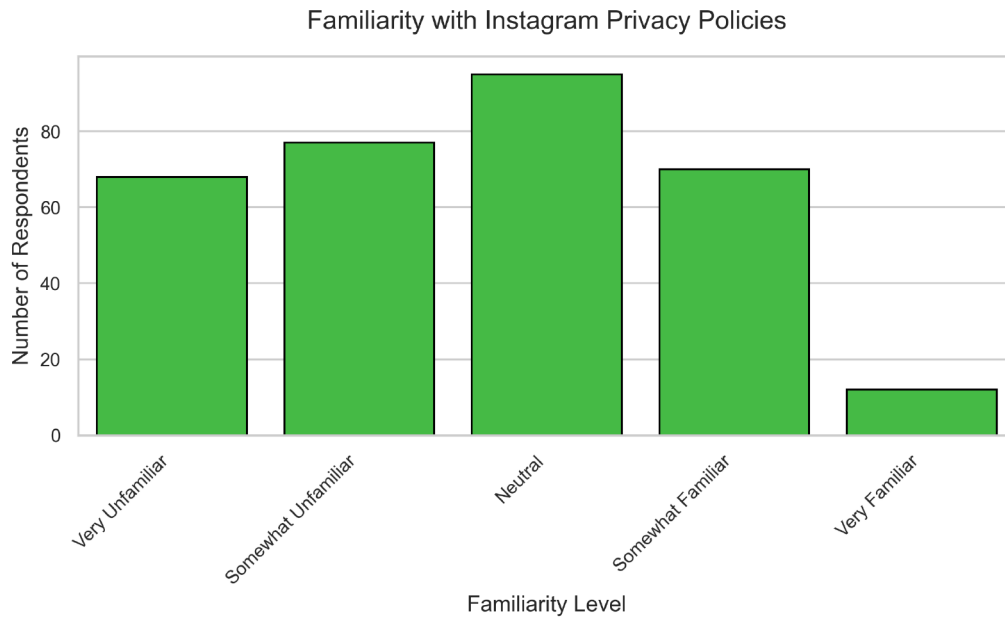


Figure 9. Participants' familiarity with Instagram privacy policies.

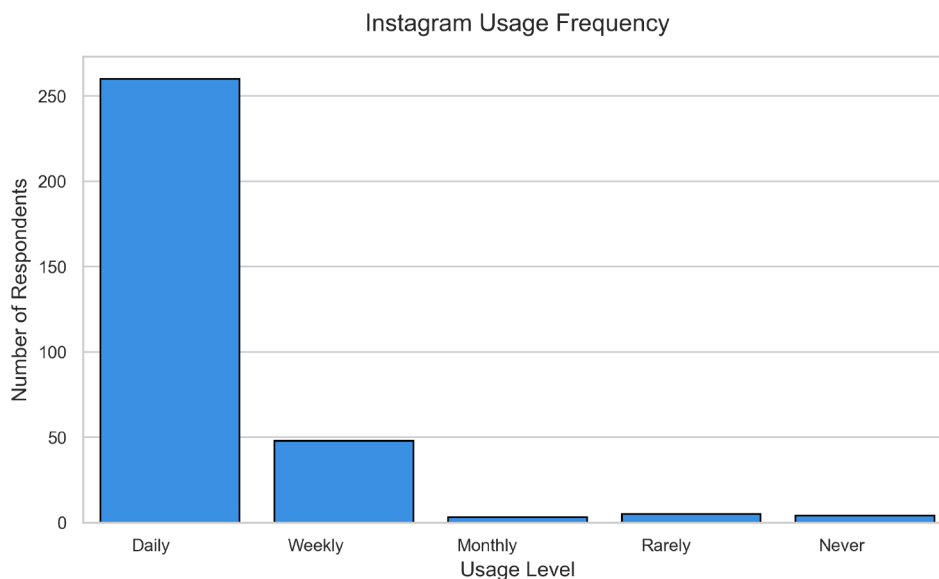


Figure 10. Frequency of Instagram use among participants.

was associated with low-barrier protective actions (adjusting privacy settings) but not with higher-effort disengagement (refusing or deleting the platform). Greater technical understanding of web scraping was therefore associated with concrete privacy management but not with platform exit. Recent evidence on AI-based inferences (Asthana et al., 2024) provides a complementary explanation: when users become aware that platforms can infer information that was never explicitly disclosed, they tend to adjust granular controls rather than abandon the service.

Second, the results are consistent with transparency operating as a contextual buffer of the awareness–behavior link rather than as a direct determinant of behavior (Schnackenberg and Tomlinson, 2016; Li, 2012; Dehling and Sunyaev, 2024). Under the high-transparency vignette, the slope of awareness on the behavior index was not significantly different from zero, while under the low-transparency vignette, the slope was positive and significant. This pattern is consistent with transparency functioning as a situational moderator, although baseline non-equivalence on awareness limits the strength of this inference.

Third, the absence of a significant association between the behavior index and user trust is consistent with prior accounts that distinguish trust from reliance (Banovic et al., 2023; Mcknight et al., 2011; Bélanger and James, 2020; Lukyanenko et al., 2022). Trust did not appear to decline in proportion to reported protective behavior in this sample. One reading is that the index captured reliance-related actions (such as platform refusal) that are conceptually distinct from attitudinal trust. The present design cannot identify the mechanisms underlying this pattern.

5.2 Scope of Interpretation

The association between web-scraping awareness and the privacy-protective behavior index underscores the need to distinguish general privacy familiarity from awareness of automated data extraction specifically. Awareness was measured rather than experimentally manipulated and may reflect prior attitudes, privacy sensitivity, or digital experience as well as knowledge of scraping itself. Baseline awareness also differed between conditions ($t(323) = -3.44$, $p = .001$), which constrains causal interpretation of the moderation finding.

The moderating role of the transparency condition should be interpreted within the boundaries of the present design. Because participants self-selected into conditions and no item was explicitly designated as a manipulation check, the findings support an association between the transparency vignette and the awareness–behavior relationship, but they do not establish that the vignette alone caused the observed difference in slopes. The between-conditions comparison of perceived transparency reported in the Methods is a post hoc consistency check on the vignette contrast and is not a substitute for a pre-specified manipulation-check item. Interpretation of the moderation pattern, therefore, remains constrained by the quasi-experimental design and by the absence of a pre-registered manipulation check, and the comparison cannot be read as independent evidence that the vignette was perceived exactly as intended.

Finally, the weak association between the privacy-protective behavior index and trust indicates that the behavioral indicator used here did not explain trust in this sample. The result should not be read as evidence that privacy concerns never affect trust, but rather as evidence that the expected negative association was not observed with the behavior index under the conditions and measures studied here (Vereschak et al., 2021). The Trust analyses reported in the main text rest on the four-item User Trust composite specified in the Measures section (Section 3), and the H3 conclusions should be read against that four-item operationalization. A parallel seven-item composite that combines the trust items with the three perceived-transparency items is reported in the Supplementary Material as a sensitivity analysis, and the seven-item results are consistent in direction and significance with the four-item results.

5.3 Practical Implications

The findings offer tentative guidance for the design of onboarding disclosures, framed as implications rather than as demonstrated causal effects, given the study's quasi-experimental nature.

First, interactive transparency mechanisms that make data flows visible and controllable, such as privacy dashboards, step-by-step onboarding explanations of collection practices, and granular privacy toggles, are consistent with the observed pattern that the awareness-behavior link was weaker under the high-transparency vignette. The present study does not directly test empowerment or perceived control, so the underlying mechanism remains open.

Second, communication design matters. Simplified, visually guided disclosures, such as short explanatory videos or tooltips, can make transparency cognitively accessible. Because the privacy-protective behavior index did not predict trust in this study, any claim that transparency maintains trust by reducing privacy concerns should be treated as a design hypothesis rather than a demonstrated effect in the present data.

Third, continuous transparency, in the form of ongoing privacy-change alerts or open reporting channels, may be a useful complement to one-time onboarding disclosures (Schnackenberg and Tomlinson, 2016; Dehling and Sunyaev, 2024). Whether such mechanisms shift long-term trust or engagement is beyond what the present data can establish.

For practitioners (designers, data officers, and product managers), the broader implication is that privacy communication that pairs awareness-raising content with usable controls is more consistent with the pattern observed here than disclosure alone. Awareness without usable controls was associated with reactive protective behavior; the high-transparency condition was associated with a flatter awareness-behavior slope. Trust may depend on factors that the behavior index used here does not capture; multi-item attitudinal trust and reliance measures would strengthen future work (Vereschak et al., 2021; Banovic et al., 2023).

The privacy-protective behavior index combines qualitatively different actions (adjusting platform privacy settings versus refusing to use a platform or deleting an account), and the joint 0/1/2 index should be read as an aggregate count of low-barrier protective actions. Design recommendations should also be evaluated against the two separate per-behavior logistic models reported under H1, which retain the distinction between in-platform settings adjustment and outright platform exit. The joint reading supports design choices that target the overall protective response, whereas the separate readings inform design choices that target one behavior over the other, such as lowering the cost of granular settings adjustment or addressing the conditions under which users disengage from the platform entirely.

5.4 Limitations

Several limitations qualify the interpretation of the results.

1. Condition assignment was self-selected rather than randomized, and baseline awareness differed significantly between conditions ($t(323) = -3.44$, $p = .001$, Cohen's $d = -0.39$); the moderation finding is therefore interpreted as a conditional association rather than as evidence that the vignette caused the difference in slopes.

2. No item was explicitly designated as a manipulation check; a post hoc between-conditions comparison of perceived transparency is reported in the Methods (manipulation-check paragraph) and in Section 4 as a consistency check. A dedicated, pre-registered manipulation-check item remains a recommendation for future research building on this design.
3. Awareness of web scraping was measured rather than manipulated and may reflect prior privacy attitudes, platform experience, or digital literacy.
4. Privacy-related outcomes were captured exclusively through two binary self-report indicators (adjusting privacy settings; refusing platform use) summed into a behavior index with values 0, 1, or 2; no attitudinal Likert privacy-concern scale was administered, so the index should not be read as a measure of attitudinal concern.
5. Baseline comparisons on age, gender, frequency of Instagram use, and familiarity with privacy policies are reported in Section 4 (Table 9). Conditions did not differ in age, gender, or frequency of Instagram use, but did differ in policy familiarity, which is incorporated into the interpretation of H2.
6. The simple slope of awareness at high transparency is reported with its full standard error in Section 4 ($SE = 0.054$, $t = -0.90$, $p = .371$); the implied slope is not significantly different from zero, consistent with the interaction-coefficient test that contrasts the two simple slopes.
7. The sample was concentrated among university students and early-career professionals, and the scenario was modeled after Instagram, limiting generalization to older populations, non-student populations, and platforms with different data practices.
8. Factor-analytic and discriminant validity evidence was not produced in the present study and is identified as a step for future work.
9. The trust items were analyzed as a single composite for parsimony; future studies should test the dimensional structure of trust, transparency, and related constructs using exploratory or confirmatory factor analysis.
10. Alternative-specification robustness checks for the privacy-protective behavior index (Poisson regression on the count, ordinal-logistic regression on the ordered categories, and a bivariate joint logistic GEE on the two binary indicators) are reported in Section 4.4. The direction and significance of the awareness effect, as well as the moderation pattern, are consistent across specifications. Outlier-exclusion sensitivity analyses remain a focus for future studies building on the present design.
11. The privacy-protective behavior index is a 0/1/2 count and was modeled with OLS for interpretive parsimony. This is a linear approximation rather than a generative model for the outcome, and predictions from the slope can fall outside the observed support of the index. Specifications that respect the support of the outcome (Poisson regression on the count and ordinal-logistic regression on the ordered categories) are reported as robustness analyses in Section 4.4 and yield consistent inferences for the awareness coefficient.
12. The two binary behavioral outcomes (settings adjustment and platform refusal) were modeled in separate logistic regressions and then summed into an index. This decomposition does not estimate the within-participant correlation between the two behaviors, and the OLS coefficient on the index (expected count) is measured on a different scale than the logistic coefficients (log-odds of a single behavior). A bivariate (joint) logistic GEE specification on the two correlated binary outcomes is reported in Section 4.4. The within-participant correlation between the two

behaviors is small ($\hat{\rho} = 0.17$), and the awareness coefficient under the bivariate specification is consistent in direction and significance with the two univariate logistic regressions.

13. All constructs were measured in a single survey wave administered after vignette exposure. The privacy-protective behavior indicators are retrospective self-reports of prior actions, whereas user trust reflects post-vignette attitudes; the cross-sectional, single-wave design therefore does not establish temporal precedence between behavior and trust within the study itself, thereby constraining the causal interpretation of the behavior-to-trust association.

6 CONCLUSION

This study examined how awareness of web scraping and a transparency vignette during onboarding were associated with a privacy-protective behavior index and with user trust. Awareness was associated with the behavior index; the transparency condition moderated this association; and the behavior index did not predict trust in this sample.

Theoretically, the results refine accounts of awareness, transparency, and behavioral privacy outcomes by showing that awareness was associated with low-barrier protective behavior but not with platform exit, that the awareness-behavior slope was substantially flatter under the high-transparency vignette, and that the behavior index did not explain trust. In practice, the results are consistent with onboarding disclosures that pair awareness-raising content with usable controls, while leaving causal claims about trust formation to designs with random assignment, a pre-registered manipulation check, and attitudinal and behavioral measures of privacy concern.

Findings should be interpreted within the boundaries of a self-selected, scenario-based survey with baseline non-equivalence on awareness. Within those boundaries, the study supports the view that awareness, transparency, privacy-protective behavior, and trust do not operate as a simple linear sequence.

ETHICS STATEMENT

This study involved human participants through an anonymous online survey. Under the institutional policy at Universidad Panamericana, anonymous, minimal-risk survey research that does not collect identifiable or sensitive personal data is exempt from formal Institutional Review Board (IRB) review. The present study fell within that exempt category. Before participation, all individuals were informed about the study's nature and purpose, the voluntary nature of their involvement, the confidentiality of their responses, and their right to withdraw at any time without consequences. Electronic informed consent was obtained from all participants before the questionnaire began.

DATA AVAILABILITY

The per-participant, de-identified dataset that supports the analyses reported in this article is provided in the Supplementary Material ([dataset.xlsx](#)). The Supplementary Material also includes the full output tables for reliability (Table S1), descriptives (Table S2), baseline group comparisons (Table S3), H1 model specifications including the bivariate logistic GEE (Table S4), the H2 moderation model with simple slopes (Table S5), the H3 regression for the four-item and seven-item Trust composites (Table S6), Pearson correlations (Table S7), variance inflation factors

(Table S8), alternative-specification robustness analyses (Table S9), and the perceived-transparency consistency check (Table S10).

ACKNOWLEDGEMENT(S)

The authors thank the Vicerrectoría General de Investigación, the Engineering Faculty, the NexEd Hub Research Group, and the Computing Department of Universidad Panamericana, Mexico City Campus.

FUNDING

The authors declare that no financial support was received for this article's research, authorship, and/or publication.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest. The funders had no role in the study's design, data collection, analysis, interpretation, manuscript writing, or decision to publish the results.

REFERENCES

- Agozie, D. Q. and Kaya, T. (2021). Discerning the effect of privacy information transparency on privacy fatigue in e-government. *Government Information Quarterly* 38, 101601. doi:10.1016/j.giq.2021.101601
- Al-Natour, S., Cavusoglu, H., Benbasat, I., and Aleem, U. (2020). An empirical investigation of the antecedents and consequences of privacy uncertainty in the context of mobile apps. *Information Systems Research* 31, 1037–1063. doi:10.1287/isre.2020.0931
- Asthana, S., Im, J., Chen, Z., and Banovic, N. (2024). “I know even if you don’t tell me”: Understanding users’ privacy preferences regarding AI-based inferences of sensitive information for personalization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA: Association for Computing Machinery), CHI '24, 1–21. doi:10.1145/3613904.3642180
- Awad, N. F. and Krishnan, M. S. (2006). The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS Quarterly* 30, 13–28. doi:10.2307/25148715
- Banovic, N., Yang, Z., Ramesh, A., and Liu, A. (2023). Being trustworthy is not enough: How untrustworthy artificial intelligence (AI) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction* 7, 1–17. doi:10.1145/3579460
- Bansal, G., Zahedi, F. M., and Gefen, D. (2010). The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems* 49, 138–150. doi:10.1016/j.dss.2010.01.010
- Barth, S. and de Jong, M. D. (2017). The privacy paradox – investigating discrepancies between expressed privacy concerns and actual online behavior – a systematic literature review. *Telematics and Informatics* 34, 1038–1058. doi:10.1016/j.tele.2017.04.013
- Bélanger, F. and James, T. L. (2020). A theory of multilevel information privacy management for the digital era. *Information Systems Research* 31, 510–536. doi:10.1287/isre.2019.0900

- Bélanger, F. and Crossler, R. E. (2011). Privacy in the digital age: A review of information privacy research in information systems. *MIS Quarterly* 35, 1017–1041. doi:10.2307/41409971
- Crossler, R. E. and Bélanger, F. (2019). Why would i use location-protective settings on my smartphone? motivating protective behaviors and the existence of the privacy knowledge–belief gap. *Information Systems Research* 30, 995–1006. doi:10.1287/isre.2019.0846
- Culnan, M. J. and Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science* 10, 104–115
- Dehling, T. and Sunyaev, A. (2024). A design theory for transparency of information privacy practices. *Information Systems Research* 35, 956–977. doi:10.1287/isre.2019.0239
- Elliott, D. and Soifer, E. (2022). Ai technologies, privacy, and security. *Frontiers in Artificial Intelligence* 5, 826737. doi:10.3389/frai.2022.826737
- Esmailzadeh, P. (2019). The impacts of the perceived transparency of privacy policies and trust in providers for building trust in health information exchange: Empirical study. *JMIR Medical Informatics* 7, e14050. doi:10.2196/14050
- Gallese, C. (2023). Web scraping ed addestramento dei modelli generativi nella direttiva 790/19. *i-lex* 16, 1–16. doi:10.6092/issn.1825-1927/18871
- Gefen, D., Karahanna, E., and Straub, D. W. (2003). Trust and tam in online shopping: An integrated model. *MIS Quarterly* 27, 51–90
- [Dataset] General Data Protection Regulation (GDPR) (2018). General data protection regulation gdpr. Retrieved from <https://gdpr-info.eu/>
- Gierlich-Joas, M., Teebken, M., and Hess, T. (2022). A synthesized perspective on privacy and transparency in the digital workplace. In *Proceedings of the 55th Hawaii International Conference on System Sciences* (Hawaii, USA: Hawaii International Conference on System Sciences), 10. doi:10.24251/HICSS.2022.633
- [Dataset] Gomstyn, A. and Jonker, A. (2025). Exploring privacy issues in the age of ai. Retrieved from <https://www.ibm.com/think/insights/ai-privacy>
- Groß, T. (2021). Validity and reliability of the scale internet users' information privacy concerns (iuipc). *Proceedings on Privacy Enhancing Technologies* 2021, 235–258. doi:10.2478/popets-2021-0026
- Hirsh, H., Basu, C., and Davison, B. D. (2000). Learning to personalize. *Commun. ACM* 43, 102–106. doi:10.1145/345124.345159
- Hu, X., Wu, G., Wu, Y., and Zhang, H. (2010). The effects of web assurance seals on consumers' initial trust in an online vendor: A functional perspective. *Decision Support Systems* 48, 407–418. doi:10.1016/j.dss.2009.10.004
- Jiang, Z., Heng, C. S., and Choi, B. C. F. (2013). Research note—privacy concerns and privacy-protective behavior in synchronous online social interactions. *Information Systems Research* 24, 579–595. doi:10.1287/isre.1120.0441
- Kim, Y., Kim, S. H., Peterson, R. A., and Choi, J. (2023). Privacy concern and its consequences: A meta-analysis. *Technological Forecasting and Social Change* 196, 122789. doi:10.1016/j.techfore.2023.122789
- Kyi, L., Mhaidli, A., Santos, C. T., Roesner, F., and Biega, A. J. (2024). “it doesn't tell me anything about how my data is used”: User perceptions of data collection purposes. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA: Association for Computing Machinery), CHI '24, 1–12. doi:10.1145/3613904.3642260

- 869 Li, Q. and Kollnig, K. (2024). Data scraping for the training of generative ai — lessons from
870 chinese case law and regulation. *Computer Law Review International* 25, 33–41. doi:10.9785/
871 cri-2024-250201
- 872 Li, Y. (2012). Theories in online information privacy research: A critical review and an integrated
873 framework. *Decision Support Systems* 54, 471–481. doi:10.1016/j.dss.2012.06.010
- 874 Lins, S., Greulich, M., Löbbers, J., Benlian, A., and Sunyaev, A. (2024). Why so skeptical?
875 investigating the emergence and consequences of consumer skepticism toward web seals.
876 *Information & Management* 61, 103920. doi:10.1016/j.im.2024.103920
- 877 Lotfi, C., Srinivasan, S., Ertz, M., and Latrous, I. (2022). Web scraping techniques and applications:
878 A literature review. In *SCRS Conference Proceedings on Intelligent Systems*, eds. R. Pal and
879 P. K. Shukla (India: Society for Computer Research and Science (SCRS)), 381–394. doi:10.
880 52458/978-93-91842-08-6-38
- 881 Lukyanenko, R., Maass, W., and Storey, V. C. (2022). Trust in artificial intelligence: From a
882 foundational trust framework to emerging research opportunities. *Electronic Markets* 32, 1993–
883 2020. doi:10.1007/s12525-022-00605-4
- 884 Malhotra, N. K., Kim, S. S., and Agarwal, J. (2004). Internet users' information privacy concerns
885 (iuipe): The construct, the scale, and a causal model. *Information Systems Research* 15, 336–355
- 886 Marengo, A. (2024). Navigating the nexus of ai and iot: A comprehensive review of data analytics
887 and privacy paradigms. *Internet of Things* 27, 101318. doi:10.1016/j.iot.2024.101318
- 888 Martin, K. (2018). The penalty for privacy violations: How privacy violations impact trust online.
889 *Journal of Business Research* 82, 103–116. doi:10.1016/j.jbusres.2017.08.034
- 890 Mcknight, D. H., Carter, M., Thatcher, J. B., and Clay, P. F. (2011). Trust in a specific technology:
891 An investigation of its components and measures. *ACM Trans. Manage. Inf. Syst.* 2. doi:10.
892 1145/1985347.1985353
- 893 Obar, J. A. (2022). Defining and assessing data privacy transparency: A third study of canadian
894 internet carriers. *International Journal of Communication* 16, 25
- 895 Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science
896 research methodology for information systems research. *Journal of Management Information*
897 *Systems* 24, 45–77. doi:10.2753/MIS0742-1222240302
- 898 Phelps, J., Nowak, G., and Ferrell, E. (2000). Privacy concerns and consumer willingness to provide
899 personal information. *Journal of Public Policy & Marketing* 19, 27–41. doi:10.1509/jppm.19.1.
900 27.16941
- 901 Pichiyan, V., Muthulingam, S., G, S., Nalajala, S., Ch, A., and Das, M. N. (2023). Web scraping
902 using natural language processing: Exploiting unstructured text for data extraction and analysis.
903 *Procedia Computer Science* 230, 193–202. doi:10.1016/j.procs.2023.12.074. 3rd International
904 Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023)
- 905 Schnackenberg, A. K. and Tomlinson, E. C. (2016). Organizational transparency: A new perspective
906 on managing trust in organization-stakeholder relationships. *Journal of Management* 42, 1784–
907 1810. doi:10.1177/0149206314525202
- 908 Sengar, S. S., Hasan, A. B., Kumar, S., and Carroll, F. (2025). Generative artificial intelligence: a
909 systematic review and applications. *Multimedia Tools and Applications* 84, 23661–23700. doi:10.
910 1007/s11042-024-20016-1
- 911 Smith, H. J., Milberg, S. J., and Burke, S. J. (1996). Information privacy: Measuring individuals'
912 concerns about organizational practices. *MIS Quarterly* 20, 167–196

- 913 Sparkes, M. (2023). The war against ai web scraping. *New Scientist* 258, 12. doi:10.1016/
914 S0262-4079(23)00836-9
- 915 Sutanto, J., Palme, E., Tan, C.-H., and Phang, C. W. (2013). Addressing the personalization-
916 privacy paradox: An empirical assessment from a field experiment on smartphone users. *MIS*
917 *Quarterly* 37, 1141–1164
- 918 Tsai, J. Y., Egelman, S., Cranor, L., and Acquisti, A. (2011). The effect of online privacy
919 information on purchasing behavior: An experimental study. *Information Systems Research*
920 22, 254–268. doi:10.1287/isre.1090.0260
- 921 Vereschak, O., Bailly, G., and Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision
922 making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer*
923 *Interaction* 5, 1–39. doi:10.1145/3476068
- 924 Weerasinghe, M., Maduranga, M. W. P., and Kawya, M. V. T. (2023). Enhancing web scraping with
925 artificial intelligence: A review. In *4th Student Symposium of Faculty of Computing* (Ratmalana,
926 Sri Lanka: Faculty of Computing, General Sir John Kotelawala Defence University), 22
- 927 Xu, H., Dinev, T., Smith, H. J., and Hart, P. (2008). Examining the formation of individuals’
928 privacy concerns: Toward an integrative view. In *Proceedings of the International Conference on*
929 *Information Systems (ICIS 2008)* (Paris, France: Association for Information Systems (AIS)),
930 Paper 6. Available at the AIS Electronic Library (AISeL)
- 931 Xu, H., Teo, H.-H., Tan, B. C. Y., and Agarwal, R. (2012). Research note—effects of individual
932 self-protection, industry self-regulation, and government regulation on privacy concerns: A study
933 of location-based services. *Information Systems Research* 23, 1342–1363. doi:10.1287/isre.1120.
934 0416

APPENDIX A. QUESTIONNAIRE ITEMS BY CONSTRUCT

935 This appendix lists the questionnaire items used in the study, organized by construct. Each set of
936 items was adapted from previously validated instruments, as indicated in parentheses.

937 Web Scraping Awareness

938 (Adapted from Bélanger and Crossler (2011); Phelps et al. (2000).)

- 939 1. I am aware that websites can collect my personal data through automated methods such as
940 web scraping.
- 941 2. I understand how companies use web scraping to gather information from users’ online activities.
- 942 3. I am knowledgeable about the implications of web scraping for my personal data privacy.

943 User Trust

944 (Adapted from Mcknight et al. (2011); Gefen et al. (2003).)

- 945 1. I believe that Instagram can handle personal data responsibly.
- 946 2. I believe that Instagram acts in its users’ best interests.
- 947 3. I believe that Instagram is honest about its data collection policies.
- 948 4. I am willing to rely on Instagram to handle my personal data appropriately.

Perceived Transparency

(Adapted from Schnackenberg and Tomlinson (2016); Li (2012); items reflect the disclosure, clarity, and accuracy dimensions of the framework.)

1. The platform provides substantial information about its data collection practices.
2. The platform's privacy policies are easy to understand.
3. I trust the accuracy of the data collection information provided by the platform.

Privacy-Protective Behavior

(Adapted from Smith et al. (1996); Malhotra et al. (2004).)

1. Have you ever adjusted your Instagram privacy settings? (Binary: Yes/No)
2. Have you ever refused to use a platform or deleted an account due to privacy concerns? (Binary: Yes/No)

All items, except the two binary privacy-protective behavior indicators, were measured using a five-point Likert scale ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*). No multi-item attitudinal Likert scale of privacy concern was administered; "privacy concern" in this study refers to the behavioral indicators above.

Capítulo 3. Discusión

Esta tesis examinó cómo la conciencia sobre prácticas de *web scraping*, la transparencia en el uso de datos y las preocupaciones de privacidad se articulan e influyen en la confianza del usuario en plataformas digitales. En conjunto, los resultados confirman parcialmente el modelo teórico propuesto: por un lado, respaldan la idea de que una mayor conciencia técnica sobre la recolección automatizada de datos incrementa las preocupaciones de privacidad (H1) y muestran que la transparencia actúa como moderador que atenúa esa relación (H2); por otro lado, los hallazgos no apoyan la hipótesis de que las preocupaciones de privacidad reduzcan de forma directa la confianza en la plataforma (H3), lo que sugiere que la dinámica de la confianza es más compleja que la planteada por algunos modelos tradicionales.

3.1 Interpretación de los resultados

El apoyo a H1 indica que la conciencia sobre *web scraping* funciona como un arma de doble filo. Tal como señala la literatura sobre privacidad y alfabetización digital, una mayor comprensión de las prácticas de recolección automatizada puede empoderar a los usuarios, pero también aumenta su percepción de riesgo y vulnerabilidad cuando no existen garantías claras de control. En este estudio, los participantes más conscientes de estas técnicas reportaron niveles más altos de preocupación y, en algunos casos, manifestaron conductas más protectoras, como el ajuste de configuraciones de privacidad. Esos cambios no se tradujeron necesariamente en el abandono de la plataforma. Esto refuerza la idea de que la conciencia, por sí sola, no basta para proteger la privacidad: debe ir acompañada de mecanismos de control efectivos y de una comunicación adecuada para no derivar en mayores niveles de ansiedad o desconfianza.

Los resultados relativos a H2 confirman que la transparencia opera como amortiguador contextual entre la conciencia y la preocupación. Cuando las prácticas de *web scraping* se comunican de manera más clara, específica y justificada durante el *onboarding*, la relación positiva entre conciencia y preocupación se debilita de forma significativa. En otras palabras, el mismo nivel de conciencia genera menos inquietud cuando las personas perciben que la plataforma explica qué datos se recopilan, para qué se utilizan y qué opciones de control están disponibles. Este hallazgo matiza visiones simplistas de la transparencia entendida como factor que siempre reduce las preocupaciones y la sitúa más bien como recurso de diseño cuya efectividad depende de su forma de implementación: el momento en que se presenta la información, el nivel de detalle y la carga cognitiva asociada resultan determinantes para que la transparencia contribuya a mitigar, y no a intensificar, la ansiedad relacionada con la privacidad.

En contraste, la falta de apoyo para H3 muestra que las preocupaciones de privacidad no

erosionan automáticamente la confianza del usuario. Aunque el coeficiente entre preocupación y confianza tuvo el signo negativo esperado, su magnitud fue pequeña y estadísticamente no significativa, y el modelo explicó solo una proporción mínima de la varianza en la confianza. Esos resultados sugieren que la confianza en plataformas de uso intensivo, como las redes sociales, puede mantenerse a pesar de la preocupación, posiblemente debido a factores como la habituación, la percepción de indispensabilidad del servicio o la creencia de que “todas las plataformas funcionan igual”. Desde la teoría de la confianza, ese comportamiento apunta a un escenario en el que los usuarios racionalizan el riesgo y priorizan la funcionalidad, de modo que la confianza se muestra relativamente resistente incluso frente a niveles moderados de preocupación.

3.2 Relación con la literatura

En conjunto, estos hallazgos aportan matices importantes a los modelos previos de privacidad y confianza. En lugar de asumir una cadena lineal en la que la conciencia genera preocupación y esta, a su vez, reduce la confianza, los resultados de esta tesis apoyan una visión más condicionada: la conciencia incrementa la preocupación, pero ese efecto se ve moderado por la transparencia percibida, y la preocupación no basta por sí sola para explicar la confianza. Para la investigación en interacción humano-computadora, esto implica que la transparencia no debería tratarse únicamente como un antecedente directo de la confianza o de la preocupación, sino como un mecanismo relacional que reconfigura cómo los usuarios interpretan y gestionan el riesgo asociado a la recolección automatizada de datos.

3.3 Implicaciones prácticas

Los resultados tienen implicaciones directas para el diseño de interfaces. El experimento muestra que es posible incrementar la conciencia sobre *web scraping* sin disparar las preocupaciones de privacidad hasta un punto crítico, siempre que la transparencia se implemente como un proceso interactivo, comprensible y continuo, y no como un simple texto legal. Para los diseñadores de experiencia de usuario y responsables de producto, esto implica pasar de una transparencia centrada en el cumplimiento normativo a una transparencia centrada en el usuario, que combine explicaciones claras, controles accionables y retroalimentación constante sobre el uso de los datos. En esa dirección, esta tesis aporta evidencia empírica que respalda la incorporación de mecanismos de transparencia bien diseñados como parte esencial de las estrategias de construcción y mantenimiento de confianza en plataformas de datos intensivos.

Capítulo 4. Conclusiones y trabajo futuro

4.1 Conclusiones

El objetivo de esta tesis fue analizar cómo la transparencia en el uso de datos modera la relación entre la conciencia sobre prácticas de *web scraping* y las preocupaciones de privacidad, así como explorar el vínculo entre dichas preocupaciones y la confianza del usuario en una plataforma digital. Para ello se diseñó y llevó a cabo un experimento de *onboarding* en una red social simulada, en el que se manipularon dos condiciones de transparencia sobre la recolección automatizada de datos y se midieron tres constructos centrales: conciencia de *web scraping*, preocupaciones de privacidad y confianza en la plataforma.

La tesis mostró que un mayor nivel de conciencia sobre *web scraping* se asocia con un incremento significativo de las preocupaciones de privacidad, lo que respalda la primera hipótesis planteada. La transparencia percibida actúa como moderador relevante de esa relación: bajo condiciones de mayor transparencia, el efecto de la conciencia sobre las preocupaciones de privacidad se debilita, apoyando la segunda hipótesis. Las preocupaciones de privacidad no resultaron un predictor significativo de la confianza, por lo que la tercera hipótesis no fue respaldada empíricamente.

4.2 Contribuciones

En el plano teórico, la tesis refuerza la idea de que la conciencia sobre prácticas de recolección automatizada no es neutral: conocer más sobre *web scraping* tiende a aumentar la percepción de riesgo y, con ello, las preocupaciones de privacidad. Al mismo tiempo, muestra que la transparencia no debe entenderse únicamente como antecedente directo de la confianza, sino como mecanismo modulador que reconfigura la relación entre conciencia y preocupación. La evidencia obtenida sugiere que la transparencia puede amortiguar, aunque no eliminar, el impacto de la conciencia sobre las inquietudes de los usuarios, lo que matiza los modelos lineales que asumen efectos uniformemente positivos o negativos de la transparencia y de la preocupación sobre la confianza.

En el plano práctico, la tesis ofrece orientaciones concretas para el diseño de interfaces en plataformas de datos intensivos. Los resultados indican que es posible incrementar la conciencia de los usuarios sin disparar sus preocupaciones de privacidad hasta un nivel que erosione su experiencia, siempre que la transparencia se implemente de forma clara, contextual y comprensible. Esto implica superar los avisos legales extensos y apostar por mecanismos de transparencia integrados en el flujo de uso, que expliquen qué datos se recolectan, con qué fines y qué opciones de control están disponibles. Para equipos de diseño y producto, estos hallazgos refuerzan la importancia de tratar la transparencia como componente central de la experiencia de usuario, y no solo como

requisito normativo.

4.3 Limitaciones

El estudio presenta limitaciones que deben considerarse. Primero, se trabajó con un escenario simulado de *onboarding* y una muestra específica, lo que restringe la generalización de los resultados a otros contextos, tipos de plataformas o poblaciones. Segundo, las medidas utilizadas se basan en autoinforme, por lo que capturan principalmente percepciones y no necesariamente comportamientos reales a largo plazo. Por último, el modelo analizado se centró en tres constructos principales y dejó fuera otros factores que podrían influir en la confianza, como la reputación percibida de la plataforma, la utilidad del servicio o la experiencia previa de los usuarios con incidentes de privacidad.

4.4 Trabajo futuro

A partir de estas limitaciones se abren varias líneas de trabajo futuro. Sería relevante replicar y extender el experimento en plataformas reales, con diferentes niveles de dependencia del servicio, e incorporar medidas conductuales (por ejemplo, la configuración efectiva de opciones de privacidad o el abandono de la plataforma) que complementen los datos de autoinforme. También sería valioso explorar otras formas de transparencia, como explicaciones personalizadas, visualizaciones de flujos de datos o notificaciones continuas, y analizar cómo interactúan con la confianza y las preocupaciones de privacidad a lo largo del tiempo. En conjunto, se espera que este trabajo contribuya a una comprensión más matizada de la relación entre transparencia, privacidad y confianza, y sirva como base para el diseño de sistemas más claros, responsables y centrados en las personas usuarias.

Bibliografia

- Al-Natour, S., Cavusoglu, H., Benbasat, I., and Aleem, U. (2020). An empirical investigation of the antecedents and consequences of privacy uncertainty in the context of mobile apps. *Information Systems Research*, 31(4):1037–1063.
- Awad, N. F. and Krishnan, M. S. (2006). The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS Quarterly*, 30(1):13–28.
- Bélanger, F. and James, T. L. (2020). A theory of multilevel information privacy management for the digital era. *Information Systems Research*, 31(2):510–536.
- Bélanger, F. and Crossler, R. E. (2011). Privacy in the digital age: A review of information privacy research in information systems. *MIS Quarterly*, 35(4):1017–1041.
- Crossler, R. E. and Bélanger, F. (2019). Why would i use location-protective settings on my smartphone? motivating protective behaviors and the existence of the privacy knowledge–belief gap. *Information Systems Research*, 30(3):995–1006.
- Culnan, M. J. and Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science*, 10(1):104–115.
- Dehling, T. and Sunyaev, A. (2024). A design theory for transparency of information privacy practices. *Information Systems Research*, 35(3):956–977.
- Esmailzadeh, P. (2019). The impacts of the perceived transparency of privacy policies and trust in providers for building trust in health information exchange: Empirical study. *JMIR Medical Informatics*, 7(4):e14050.
- Gallese, C. (2023). Web scraping ed addestramento dei modelli generativi nella direttiva 790/19. *i-lex*, 16(2):1–16.
- Gefen, D., Karahanna, E., and Straub, D. W. (2003). Trust and tam in online shopping: An integrated model. *MIS Quarterly*, 27(1):51–90.
- General Data Protection Regulation (GDPR) (2018). General data protection regulation gdpr. Retrieved from <https://gdpr-info.eu/>.

- Gierlich-Joas, M., Teebken, M., and Hess, T. (2022). A synthesized perspective on privacy and transparency in the digital workplace. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, page 10, Hawaii, USA. Hawaii International Conference on System Sciences.
- Hirsh, H., Basu, C., and Davison, B. D. (2000). Learning to personalize. *Commun. ACM*, 43(8):102–106.
- Kyi, L., Mhaidli, A., Santos, C. T., Roesner, F., and Biega, A. J. (2024). “it doesn’t tell me anything about how my data is used”: User perceptions of data collection purposes. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, pages 1–12, New York, NY, USA. Association for Computing Machinery.
- Li, Q. and Kollnig, K. (2024). Data scraping for the training of generative ai — lessons from chinese case law and regulation. *Computer Law Review International*, 25(2):33–41.
- Lotfi, C., Srinivasan, S., Ertz, M., and Latrous, I. (2022). Web scraping techniques and applications: A literature review. In Pal, R. and Shukla, P. K., editors, *SCRS Conference Proceedings on Intelligent Systems*, pages 381–394, India. Society for Computer Research and Science (SCRS).
- Lukyanenko, R., Maass, W., and Storey, V. C. (2022). Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities. *Electronic Markets*, 32(4):1993–2020.
- Marengo, A. (2024). Navigating the nexus of ai and iot: A comprehensive review of data analytics and privacy paradigms. *Internet of Things*, 27:101318.
- Martin, K. (2018). The penalty for privacy violations: How privacy violations impact trust online. *Journal of Business Research*, 82:103–116.
- Mcknight, D. H., Carter, M., Thatcher, J. B., and Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manage. Inf. Syst.*, 2(2).
- Obar, J. A. (2022). Defining and assessing data privacy transparency: A third study of canadian internet carriers. *International Journal of Communication*, 16:25.
- Pichiyan, V., Muthulingam, S., G, S., Nalajala, S., Ch, A., and Das, M. N. (2023). Web scraping using natural language processing: Exploiting unstructured text for data extraction and analysis. *Procedia Computer Science*, 230:193–202. 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023).

- Sutanto, J., Palme, E., Tan, C.-H., and Phang, C. W. (2013). Addressing the personalization-privacy paradox: An empirical assessment from a field experiment on smartphone users. *MIS Quarterly*, 37(4):1141–1164.
- Xu, H., Dinev, T., Smith, H. J., and Hart, P. (2008). Examining the formation of individuals' privacy concerns: Toward an integrative view. In *Proceedings of the International Conference on Information Systems (ICIS 2008)*, page Paper 6, Paris, France. Association for Information Systems (AIS). Available at the AIS Electronic Library (AISeL).