



## UvA-DARE (Digital Academic Repository)

### Is explanation the cure?

*A human-centered framework for explainable recommender systems*

Wang, C.

### Publication date

2026

[Link to publication](#)

### Citation for published version (APA):

Wang, C. (2026). *Is explanation the cure? A human-centered framework for explainable recommender systems*. [Thesis, fully internal, Universiteit van Amsterdam].

### General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

### Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

# Chapter 1. Introduction

Every day, without asking, algorithms decide what we see. They choose the next series we binge, the headlines we read over breakfast, and the posts that shape our conversations. These choices are rarely explained, yet they quietly guide our attention, influence our opinions, and frame our understanding of the world. As the reach of AI-driven recommender systems grows, so does a simple but urgent question: Why am I seeing this recommendation and not something else?

In recent years, the promise of Explainable AI (XAI) has captured the attention of scholars, policymakers, and designers alike. Explanations are seen as a way to make algorithmic decisions more transparent, fostering understanding, trust, and fairness. Yet they are no simple cure-all. The way explanations are designed, whether users notice them, and how well they align with users' knowledge and skills can mean the difference between empowerment and confusion. This dissertation takes up the challenge of designing explanations that empower rather than confuse users by examining *for whom*, *in what ways*, and *under what conditions* such explanations can make AI-driven recommender systems more comprehensible, effective, and inclusive.

### **Setting the scene: Toward a human-centered conceptual framework**

Artificial intelligence (AI) technologies are increasingly embedded in everyday life. Among them, AI-driven recommender systems (RS) have become key intermediaries shaping how individuals access entertainment and news content. Such systems use algorithms to filter, rank, and present items to users based on their inferred preferences, thereby personalizing information access across domains such as entertainment and news (Jannach et al., 2022; Ricci et al., 2021). Whether helping users discover a new movie or selecting which headlines appear in their news feeds, these systems influence attention, shape choices, and ultimately affect perceptions of the world (e.g., Yu et al., 2024; Joris et al., 2024).

By tailoring content to individual preferences, recommender systems enhance discovery, satisfaction, and ease of use (Zhang et al., 2021). However, alongside these benefits, they raise societal concerns about fairness, bias, and ethical use. Research has shown that algorithmic personalization can lead to

information isolation and ideologically narrow feeds (Cinelli et al., 2021; Whittaker et al., 2021). Personalization mechanisms may also unintentionally reinforce biases or amplify polarizing and extremist content, with documented effects on public discourse and democratic participation (Bakshy et al., 2015; Ribeiro et al., 2020).

These risks are further magnified by the opacity of AI systems. Users are rarely aware of why they are shown particular items, which limits their ability to evaluate relevance, detect potential bias, or exercise informed choice (Burrell, 2016; Nunes & Jannach, 2017). This opacity not only undermines trust but also reduces users to passive recipients of algorithmic decisions (Eslami et al., 2016; Swart, 2021). Without adequate understanding, people cannot meaningfully contest, calibrate, or resist algorithmic influence, particularly in contexts such as news consumption, where exposure diversity and fairness are essential (Thurman, 2019; Napoli, 2014).

These challenges have prompted growing calls among the public, scholars, and policymakers for transparency, accountability, and more user-centered approaches to AI design. Regulatory frameworks such as the EU *Digital Services Act* (Regulation (EU) 2022/2065) and the EU *AI Act* (OJ 2024, June 13) reflect a broader shift toward promoting transparency and accountability in high-impact AI systems including recommender algorithms. These regulations require that users receive meaningful information about how automated decisions are made and how they influence access to information. In the Netherlands, the *Autoriteit Persoonsgegevens* and other national bodies are developing guidance to align domestic AI governance and transparency practices with these EU-level requirements, further reinforcing the expectation of explainable and user-oriented algorithmic systems (Autoriteit Persoonsgegevens, 2024). Within this context, explanations that clarify how recommendations are generated have come to be seen as a practical way to address these ethical, legal, and operational imperatives.

Consequently, Explainable AI (XAI) has emerged as a pivotal response to the opacity of AI systems, shaped by both regulatory pressures and technical efforts to make algorithms more interpretable. XAI refers to methods and techniques that make model outputs or internal processes understandable to human users without unduly sacrificing performance (Arrieta et al., 2020; Naiseh, 2024; Patil & Patil,

2023; Vilone & Longo, 2021). In the context of recommender systems, explainability seeks to make the system's reasoning transparent by clarifying how recommendations are generated and why particular items are suggested, using XAI techniques such as feature-based justifications, similarity explanations, or model-agnostic approaches such as SHAP or LIME (Ribeiro et al., 2020; Lundberg & Lee, 2017).

However, making algorithms interpretable is only one part of the challenge. Technical transparency alone does not guarantee that users can comprehend or act upon explanations. This limitation has prompted a shift from system-centered transparency toward a human-centered perspective in XAI, which emphasizes the importance of designing explanations that align with users' cognitive needs, expectations, and goals (Liao & Varshney, 2021; Wang et al., 2019). Within recommender systems, this approach highlights that explanations should not only reveal how recommendations are produced but also support meaningful user understanding and engagement.

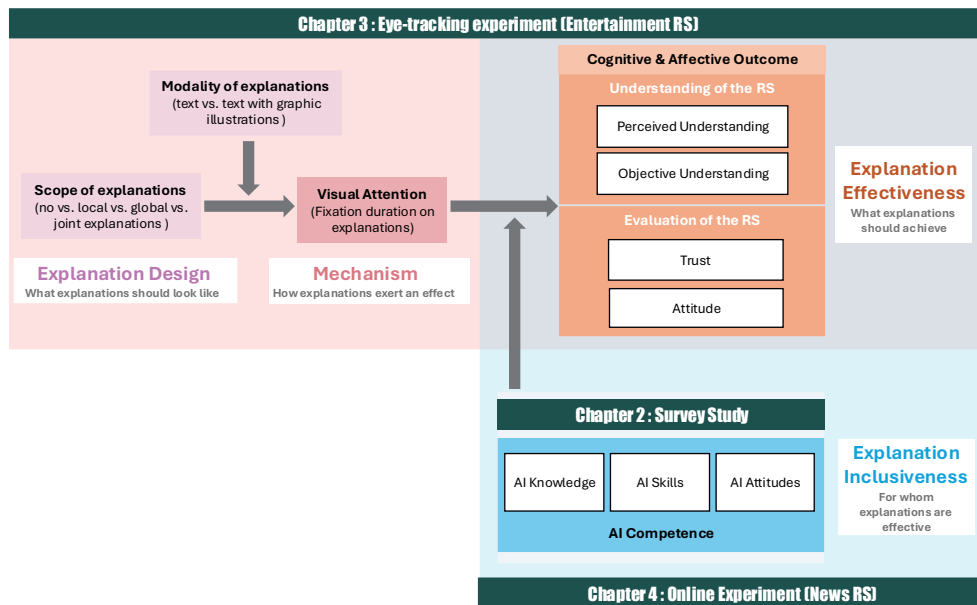
Human-centered explanations are therefore seen as a key strategy to help mitigate the societal risks associated with opaque recommender systems such as filter bubbles and echo chambers (Pariser, 2011; Zuiderveen Borgesius et al., 2016; Helberger et al., 2018; Stray, 2024), biased or manipulative recommendations (Binns et al., 2018; Eslami et al., 2016), and declining trust and autonomy in algorithmic environments (Shin, 2021; Mittelstadt et al., 2019). Viewing explainability through this lens situates it as both a technical and socio-technical design challenge, requiring a careful balance between algorithmic performance, communicative clarity, and user empowerment.

However, despite this growing recognition, research on explainable recommender systems remains conceptually and empirically fragmented. Many studies treat explanations as a binary feature, focusing only on their presence or absence rather than examining them as interrelated design choices (Abusitta et al., 2024; Liao & Varshney, 2021). Research has also tended to examine cognitive and affective outcomes separately, despite evidence that understanding and trust are closely related (Liao & Vaughan, 2023; Miller, 2019). Finally, there has been limited attention to user heterogeneity, as most studies assume uniform responses

and overlook differences in users' competence (Ehsan et al., 2021; Ahn et al., 2025).

Bringing these strands together and addressing these limitations, this dissertation proposes an *integrated conceptual framework* that connects explanation design, attentional mechanisms, user competencies, and outcomes (see Figure 1). It addresses four interrelated questions: (1) what explanations should look like, that is, how design dimensions such as scope and modality constitute effective explanation; (2) what explanations should achieve, namely, how they influence users' cognitive and affective outcomes such as understanding, trust, and attitudes toward the system; (3) how explanations exert effects, specifically through attentional and cognitive mechanisms, with visual attention functioning as a key underlying process; and (4) for whom explanations are effective, considering variations in users' AI competence, including differences in AI-related knowledge, skills, and attitudes. The framework serves as a bridge between communication science, cognitive psychology, and human-computer interaction, offering a cumulative perspective for studying effective and inclusive explainable recommender systems. The next section elaborates each component of the framework in turn; together, these components form the basis for the empirical studies presented in later chapters.

Figure 1. Integrated conceptual model of human-centered explanations in recommender systems



Note. The model connects four dimensions: explanation design, mechanism, explanation effectiveness, and explanation inclusiveness. It illustrates how design choices influence user outcomes across entertainment and news recommender systems. See the following section on core concepts and theories for further explanation.

## Unpacking the framework: Components and theoretical grounding

### What explanations should look like: Explanation design

A central starting point for understanding human-centered explanations lies in their design. Prior research has identified various dimensions of explanation design, including the type of information provided, the level of detail, timing, interactivity, and personalization (e.g., Miller, 2019; Ribeiro et al., 2016; Ehsan et al., 2021). Within this broad landscape, this dissertation focuses on two key dimensions: *scope* and *modality*. These dimensions capture the essential aspects of explanation design: what information they communicate to users and in what form (e.g., Herlocker et al., 2000; Tintarev & Masthoff, 2011; Doshi-Velez & Kim, 2017; Nunes & Jannach, 2017). Together, scope and modality define the core structure of explanation design, forming the foundation for the conceptual framework developed in this dissertation.

Explanation scope refers to the level at which a system’s decision-making process is communicated. *Local* explanations focus on the rationale for specific recommendations, enabling users to verify or contest individual outputs. For instance, “*This film is recommended because it stars the same actor as another movie you watched*” (Doshi-Velez & Kim, 2017; Tintarev & Masthoff, 2011). *Global* explanations describe the overall logic, principles, or algorithmic strategies of the system, supporting the development of accurate mental models. For example, “*This system recommends content based on genre popularity*” (Herlocker et al., 2000; Kulesza et al., 2013). *Joint* explanations combine local and global perspectives, offering a more complete understanding by integrating instance-level justifications with broader system-level context (Radensky et al., 2022; Kunkel et al., 2019). Empirical studies suggest that such combined approaches can enhance both perceived and objective understanding, fostering greater trust and satisfaction than single-scope explanations (Zhang & Chen, 2020; Guesmi et al., 2023). This taxonomy builds on mental model theory, which proposes that richer conceptual representations of system functioning can improve comprehension and enable more informed decision making (Norman, 2023; Kulesza et al., 2013). By clarifying what is being explained and at what level, scope serves as a key structural dimension of explanation design that links algorithmic fidelity with user comprehensibility (Bhatnagar & Agrawal, 2024; Mersha et al., 2024).

While explanation scope determines *what* information is communicated, modality shapes *how* it is conveyed. Explanation modality concerns the format in which explanations are delivered, typically through text, visuals (e.g., charts, icons), or multimodal combinations. The Cognitive Theory of Multimedia Learning (Mayer, 2005) and Cognitive Load Theory (Sweller, 1988) suggest that combining textual and visual information can enhance understanding by engaging dual processing channels, provided that the materials are well-aligned and avoid extraneous cognitive load (Mayer & Moreno, 2003). In recommender systems, visual modalities such as graphs, icons, or highlighted features can make algorithmic reasoning more tangible and engaging, helping users grasp the rationale behind recommendations (Kouki et al., 2019; Guesmi et al., 2023). For example, visual cues that link recommended items to user preferences or similarity clusters have been shown to increase perceived transparency and trust (Pu & Chen, 2007; Hernandez-Bocanegra & Ziegler, 2021). Yet, visual explanations are not

universally beneficial. Poorly aligned visuals can introduce cognitive noise or mislead interpretation, obscuring rather than clarifying the recommendation logic (Chromik & Butz, 2021; Meza Martínez et al., 2023). The key challenge, therefore, lies in achieving modality alignment, ensuring that form and content reinforce one another to support cognitive processing rather than overwhelm it. In this sense, modality serves as the presentational counterpart to scope, determining whether explanations are not only accurate but also accessible and meaningful to diverse users.

Despite these advances, important gaps remain in how explanation design has been studied. Much prior research treats explanations as a binary feature, namely present or absent, rather than examining how multiple design dimensions jointly affect user experience (Tintarev & Masthoff, 2011; Nunes & Jannach, 2017; Abusitta et al., 2024). In particular, few studies have examined scope and modality in combination, leaving unanswered questions about how different levels and formats interact to balance explanatory completeness with cognitive effort (Herm & Janiesch, 2023; Chromik & Butz, 2021). To address these gaps, this dissertation foregrounds scope and modality as the core dimensions of explanation design. Together, they outline both the content and the presentation of explanations, offering a structured foundation for understanding how explanations shape user responses.

### **What explanations should achieve: Cognitive vs. affective outcomes**

Having outlined how explanations are designed through scope and modality, the next question is what those explanations are meant to achieve for users. The central aim of explainable AI is to enhance transparency while supporting users' informed engagement with algorithmic systems. Policy frameworks such as the EU AI Act and the EU Digital Services Act link explainability to principles of accountability, user empowerment, and trustworthiness (European Commission, 2023; European Union, 2022). These goals translate into two complementary dimensions of user response: cognitive outcomes, which concern what users understand about the system, and affective outcomes, which concern how users evaluate the system and whether they are willing to rely on it. Explanations are therefore expected to influence both cognitive and affective outcomes, each serving as a pathway through which transparency achieves its societal and regulatory aims.

Cognitive outcomes capture the extent to which users understand an AI system. They include *perceived understanding*, the subjective sense of comprehension, and *objective understanding*, the accuracy of users' mental representations of how the system functions (Hoffman et al., 2023). Drawing on sensemaking theory, which views understanding as a process of constructing coherence in the face of uncertainty (Miller, 2019), and mental model frameworks, which describe how users form internal representations that guide prediction and interaction (Kulesza et al., 2013), explanations help users build structured understandings of algorithmic processes. These cognitive processes, in turn, help reduce system opacity and enable users to evaluate more accurately the relevance and fairness of algorithmic outcomes. Empirical work supports these theoretical claims: Kulesza et al. (2013) showed that explanations engineered for soundness and completeness improved mental model accuracy and users' ability to predict system behavior. In applied recommender system settings, layered and transparent explanations have been found to raise users' perceived clarity about why items are suggested (Guesmi et al., 2023; Zhang & Chen, 2020). However, most prior studies have examined either perceived or objective understanding in isolation, with few comparing them directly or exploring how they jointly reflect user comprehension. Addressing this gap, the present dissertation measures both dimensions simultaneously to provide a more complete account of how explanations influence user understanding of AI-driven recommendations.

Beyond informing users, explanations also shape how they evaluate and choose to engage with algorithmic systems (Liao & Varshney, 2021; Shin, 2021). Affective outcomes capture these evaluative and emotional reactions, reflecting how users appraise, trust, and choose to engage with AI systems. Among the various affective responses studied, *trust* and *attitude* consistently emerge as central constructs in explainable AI and recommender-system research (e.g., Tintarev & Masthoff, 2015; Zhang & Chen, 2020; Shin, 2021; Liao et al., 2022). Trust refers to users' willingness to rely on a system's outputs even when its internal workings cannot be fully verified, reflecting confidence in its competence and integrity (Mayer et al., 1995; Lee & See, 2004). Attitude, in contrast, captures users' broader evaluative orientation toward the system. While related concepts such as satisfaction and acceptance are often used interchangeably, attitude provides a more stable and generalized evaluative lens (Abadi et al., 2025).

Together, trust and attitude serve as complementary indicators of affective engagement: trust denotes calibrated reliance and confidence, while attitude reflects overall favorability and acceptance.

In recommender system contexts, explanations that link recommended items to user preferences or item features have been shown to increase satisfaction, acceptance, and perceived transparency, which function as proximal indicators of trust and positive attitude (Tintarev & Masthoff, 2015; Zhang & Chen, 2020; de Campos et al., 2024). Theories of trust further emphasize that appropriate reliance should stem from informed understanding (Mayer et al., 1995; Lee & See, 2004), and explanations therefore play a central role in trust calibration, enabling users to assess whether system behavior aligns with their expectations and values (Liao & Sundar, 2022; Shin, 2021).

However, despite the conceptual link between understanding and trust, cognitive and affective outcomes have often been examined in isolation. Although theory suggests that greater understanding should foster greater trust and more favorable attitudes, empirical research rarely examines how these processes interact (Kaplan et al., 2023; Liao & Sundar, 2022). This separation limits insight into how these two outcome domains relate and whether cognitive gains reliably translate into affective benefits. Addressing this gap, the present dissertation integrates cognitive and affective outcomes within a unified framework. It argues that effective explanations should enhance understanding while fostering appropriate trust and evaluation, offering a more comprehensive account of when explanations succeed and when they fall short. In doing so, it advances explainable AI from a technical ideal to a communicative practice that fosters informed and trustworthy human-AI interaction.

### **How explanations exert effects: Visual attention as an underlying mechanism**

Even the most carefully designed explanations cannot achieve their goals if users do not notice or process them. Research on transparency in domains such as advertising and news shows that disclosure messages often fail when audiences overlook them or fail to grasp their meaning (Boerman et al., 2014, 2017; Wojdyski & Evans, 2016). Similar patterns are observed in explainable

recommender systems, where users frequently ignore explanations or even misremember having seen them (Andrienko et al., 2022; Chromik & Butz, 2021; Poursabzi-Sangdeh et al., 2021). Consequently, even well-designed explanations cannot influence user outcomes if they are not visually attended to and cognitively processed (Kulesza et al., 2015; Poursabzi-Sangdeh et al., 2021; Meza Martínez et al., 2023). Hence, visual attention is central to understanding how explanation design shapes user responses in human-AI interaction.

Visual attention refers to the selective allocation of perceptual and cognitive resources to elements within a visual display (Wolfe, 2000). In human-computer interaction, it determines which parts of an interface are perceived, in what sequence, and for how long. Although attention can sometimes occur without direct eye movement (Posner, 1980), in visual interfaces, eye fixations are generally regarded as reliable indicators of cognitive engagement (Rayner, 2009; Holmqvist et al., 2011). This assumption is also formalized in the Eye-Mind Hypothesis, which posits a close temporal link between where people look and what they process cognitively (Just & Carpenter, 1980; Rayner, 2009; Holmqvist et al., 2011).

Design choices play a crucial role in directing users' visual attention. Research in human-computer interaction shows that layout, prominence, spacing, and visual grouping influence how people scan and focus on specific elements (Goldberg & Kotval, 1999; Jacob & Karn, 2003). These insights are consistent with limited-capacity models of information processing, which propose that users possess finite cognitive resources that must be allocated across competing stimuli (Lang, 2000). Within explainable AI, design features such as the scope and modality of explanations thus determine not only how information is presented but also how users allocate attention to and engage with explanatory content.

Visual attention, in turn, functions as the key pathway through which explanations influence user outcomes. Fixations and viewing time on explanatory elements create opportunities for users to encode causal relations and construct mental models, supporting cognitive outcomes such as users' perceived and objective understanding (Kulesza et al., 2013; Hoffman et al., 2023). The Limited Capacity Model of Motivated Mediated Message Processing (Lang, 2000) further suggests that people allocate more cognitive resources to information they find

personally meaningful or engaging. Sustained attention to clear and relevant explanations may therefore foster calibrated trust, namely reliance that matches perceived system competence and integrity (Mayer et al., 1995; Lee & See, 2004; Liao et al., 2022). However, the relationship between attention and benefit is not linear: too little attention implies non-exposure, whereas excessive attention may signal processing difficulty rather than deeper comprehension (Sweller, 1988; Holmqvist et al., 2011). Effective explanation design must therefore capture users' attention efficiently while avoiding cognitive overload.

Although visual attention is central to understanding explanation effects, most research has examined it only indirectly through self-reported engagement or recall (Chromik & Butz, 2021; Liao & Varshney, 2021). Only a limited number of studies in explainable AI and human–computer interaction have applied process-tracing methods such as eye-tracking, cursor trajectories, and interaction logs to observe how users allocate attention to explanations and how these patterns relate to comprehension and evaluation (Eiband et al., 2019; Rader et al., 2018; Kunkel et al., 2019; Chromik & Butz, 2021). Yet such approaches remain rare, leaving open questions about when and how explanation designs succeed in capturing users' attention and translating it into meaningful user outcomes.

This dissertation addresses this gap by empirically examining visual attention as a mediating mechanism linking explanation design, specifically its scope and modality, to user outcomes including understanding, trust, and attitudes toward explainable recommender systems. Eye-tracking is employed to capture users' visual attention to explanatory elements, providing direct behavioral evidence of how design features shape perceptual engagement (Holmqvist et al., 2011; Eiband et al., 2018; Fan et al., 2022). This approach contributes to the literature by integrating process-level evidence into human-centered research on explainable AI, clarifying how attention mediates the link between explanatory features and user responses.

### **For whom explanations are effective: Explanation inclusiveness**

A human-centered account of explainability must also ask for whom explanations are effective. Even carefully designed, theory-grounded explanations only succeed

if users can notice, interpret, and act on them, which makes user competence a central concern for inclusive explainable AI.

The notion of AI competence refers to individuals' ability to understand, interact with, and critically assess AI systems. Building on digital-competence frameworks (Ferrari, 2012; van Deursen & van Dijk, 2014) and AI-literacy scholarship (e.g., Carolus et al., 2023; Long & Magerko, 2020), it comprises three interrelated dimensions: AI-related knowledge, skills, and attitudes. *Knowledge* is the conceptual understanding of how algorithmic systems function; *skills* are the practical abilities to recognize, control, and influence algorithmic processes; and *attitudes* are evaluative orientations that shape motivation to engage. Together, these dimensions determine how effectively users can make sense of and benefit from explanations.

Research consistently shows that such competences are unevenly distributed. Communication and information-science studies document persistent disparities in digital and algorithmic skills, reflecting broader inequalities by education, age, and socioeconomic status (Hargittai, 2001; van Dijk, 2020; Gran et al., 2021). Scholars further warn that as AI systems become ubiquitous, these gaps may widen differences in users' ability to understand and meaningfully engage with algorithmic decisions (Cotter & Reisdorf, 2020; Zarouali et al., 2021). Individuals with lower AI competence are more likely to misinterpret or overtrust algorithmic outputs, overlook bias or manipulation, and lack the capacity to contest or adjust recommendations (e.g., Voorveld et al., 2024; Park & Young, 2025). This, in turn, can heighten susceptibility to misinformation and algorithmic persuasion, reduce autonomy in digital environments, and exacerbate inequities in algorithmically mediated systems (Eynon et al., 2023; Wang et al., 2022). Accordingly, explainable systems that assume a uniform user base may overlook these differences.

Several theoretical perspectives help explain why AI competence is expected to condition explanation effectiveness. Constructivist learning theory (Bruner, 1961) posits that new information is more easily assimilated when it connects to existing knowledge. Users with stronger AI knowledge may therefore integrate explanations into more coherent mental models of how the system works (Kulesza et al., 2013). The Elaboration Likelihood Model (Petty & Cacioppo,

1986) further suggests that comprehension depends on both ability and motivation. Users with higher skills and more positive attitudes toward AI are more likely to process explanations systematically, leading to deeper understanding and more stable evaluations. Finally, cognitive-fit theory (Vessey, 1991) proposes that explanation effectiveness depends on the match between information format and user expertise. Explanations that are aligned with users' cognitive resources tend to facilitate understanding and improve performance.

From these perspectives, inclusiveness should be a design principle rather than an afterthought. Ensuring that users with different levels of AI competence can derive meaningful understanding and appropriately calibrated trust also aligns with responsible-AI governance, including the EU AI Act and the Digital Services Act (European Union, 2022; European Commission, 2023). However, much existing work on explainable AI and recommender systems still assumes a uniform user base, overlooking competence differences that shape how people interpret and benefit from explanations (Ehsan et al., 2021; Liao et al., 2022). Empirical reviews show that many studies report average effects but seldom examine how outcomes vary by user background, competence, or motivation (Wardatzky et al., 2024; Rong et al., 2023; Naveed et al., 2024). Without sufficient attention to user heterogeneity, evaluations may provide an incomplete picture of explanation effectiveness, masking variations in who benefits and how.

This dissertation addresses these gaps by embedding user competence at the core of a human-centered framework for explainable recommender systems. It conceptualizes AI competence as a multidimensional moderator of explanation effects on user outcomes, thereby linking inclusiveness to measurable psychological and communicative mechanisms. Framed this way, explanations are evaluated not only by aggregate effectiveness but also by equity: that is, whether different groups can meaningfully benefit.

Taken together, the preceding sections defined the key dimensions that constitute this conceptual framework. The following chapters each present an independent empirical study that examines one or more of these dimensions in distinct contexts. The next subsection outlines the structure of the dissertation and the focus of each chapter.

# Dissertation Outline

This dissertation consists of five chapters. The following three chapters each address different components of the framework, providing complementary perspectives on how explanation design, underlying mechanisms, user characteristics, and contexts shape explanation effectiveness. The final chapter synthesizes the findings and reflects on their broader significance.

**Chapter 2** (*Study 1: The artificial intelligence divide: Who is the most vulnerable?*) establishes the broader context in which human-centered explainable AI for recommender systems operates. Using a large-scale, population-representative survey in the Netherlands, this study investigates individuals' AI-related competencies, including knowledge, skills, and attitudes, and identifies distinct user groups through latent class analysis. The most vulnerable groups, characterized by limited AI knowledge and skills, were typically older, less educated, and less active in privacy protection. These findings reveal demographic and cognitive inequalities that define the contours of the emerging AI divide, framing inclusiveness as a central concern for explanation design and for equitable participation in AI-driven environments. By confirming substantial variation in AI competence across the population, this chapter also provides the empirical foundation for Chapter 4, where AI competencies are examined as moderators of explanation effects in news recommender systems.

**Chapter 3** (*Study 2: When Recommendations Are Explainable: An Eye-Tracking Study Comparing How and What to Explain*) examines the design and mechanism dimensions of the conceptual framework, focusing on how explanation scope and modality influence user outcomes through visual attention. In a controlled between-subjects lab experiment using a movie recommender system, explanation scope (i.e., no, local, global, or joint) and modality (i.e., text-only vs. text with illustrations) were manipulated alongside a no-explanation control group. Eye-tracking was employed to capture visual attention, operationalized as fixation duration on predefined Areas of Interest, as the process through which design features influence user responses. The study assessed both cognitive outcomes (i.e., perceived and objective understanding of the RS) and affective outcomes (i.e., trust and attitudes toward the RS), using structural equation modeling to test mediation and moderation effects. By clarifying how explanation design and attentional

processes determine explanation effectiveness, this chapter lays the groundwork for examining explanation inclusiveness in Chapter 4.

**Chapter 4** (*Study 3: The Role of AI Competencies: How do AI knowledge, skills, and attitudes shape users' understanding and evaluation of explainable news recommendations?*) applies the conceptual framework to the context of news recommender systems and focuses on user inclusiveness. Building on Chapter 2's evidence of AI competence disparities and Chapter 3's insights into explanation design, this study examines whether explanations in news recommenders benefit all users equally or whether their effectiveness depends on individual differences in AI competence. In a between-subjects online experiment with a quota-sampled Dutch population, participants were exposed either to a joint text-only explanation of recommendations or to no explanation. The study measured both cognitive outcomes (i.e., perceived and objective understanding of the RS) and affective outcomes (i.e., trust and attitudes toward the RS), while incorporating validated scales of AI competence (i.e., AI knowledge, skills, and attitudes) as moderators. This design allows for testing whether explanation effects differ across competence levels, thus addressing the inclusiveness aspect of the conceptual framework.

**Chapter 5** synthesizes insights from the three empirical studies, reflecting on their contributions to understanding effective and inclusive explanations in recommender systems. It discusses how explanation design, attentional mechanisms, and user competencies jointly shape explanation outcomes, and it also outlines implications for theory, design practice, and policy. Limitations and directions for future research are also considered.

Taken together, the outline shows how each chapter contributes a distinct yet complementary perspective to the integrated conceptual framework. The next chapter begins the empirical exploration by examining users' AI competence and the AI divide.