Educational Requirements for Trustworthy Recommender Systems

Alexandru Mateescu
Université Paris 1 Panthéon-Sorbonne
École Doctorale 280, Department of Philosophy
Paris, France
alexandru.mateescu@etu.univ-paris1.fr

***Abstract*—Recommender systems are widely deployed AI technologies that influence users' beliefs, preferences, and behaviors — often through adaptive mechanisms that remain opaque and difficult to contest. While technical and ethical safeguards have been proposed to improve system trustworthiness, they typically focus on internal properties such as fairness or robustness, and treat the user as a passive recipient of algorithmic output. In this position paper, we argue for a shift: trust must be co-constructed through mechanisms that actively support user understanding, evaluation, and control. We introduce the concept of *educational requirements* — design elements that embed epistemic support into the system itself — and outline how they can be specified, implemented, and assessed throughout the development lifecycle. We illustrate their relevance through the case of echo chambers in personalized content feeds, and show how educational requirements transform users from engagement targets into epistemic participants. Trust, we contend, is not merely engineered — it must be learned, enacted, and sustained.**

I. Introduction

Among the many AI systems that interact with users, recommender systems occupy a distinctive and influential role: they do not simply respond to user queries or assist with tasks, but actively shape what users see, think about, and ultimately choose — often without their awareness. This is not incidental: recommender systems based on personalization are, by definition, adaptive. Their outputs evolve in response to user behavior, optimizing for metrics such as engagement or retention rather than epistemic value. While these systems are nominally designed to deliver relevant content, in practice, they are typically tuned to maximize behavioral signals like click-through rate or watch time — not understanding or utility. To date, most approaches to trustworthy AI focus on internal system properties: explainability, fairness, non-discrimination, or robustness. These are crucial, but incomplete. Trustworthiness is not solely a matter of system architecture — it is also a matter of how users understand, evaluate, and respond to the system's behavior over time. In adaptive environments, where system outputs evolve based on user input, trust cannot be engineered into the system alone. We argue that recommender systems require not just ethical and functional safeguards, but also educational requirements: design features that support users in understanding how recommendations are generated, how their actions shape outcomes, and how to retain agency in the face of persuasive optimization. This proposal draws on a long intellectual tradition, beginning with Aristotle's theory of rhetoric, which emphasized not only the art of persuasion but also the capacity to recognize and resist it. In classical contexts, rhetorical influence was governed by shared norms and open deliberation. In contrast, modern recommender systems operate through structured appeals — personalization, reinforcement, emotional salience — but without shared rules of engagement or accessible means of contestation. In the digital environment, such awareness cannot be assumed. Users are often unaware not only of how they are being influenced, but that influence is occurring at all. In what follows, we examine the limitations of current approaches to trust in recommender systems, which tend to frame users as passive recipients of algorithmic output rather than as active participants in shaping their informational environments. This asymmetry is particularly problematic in adaptive systems, where user behavior not only triggers responses but gradually trains the system itself. We argue that users must be given the means to understand (intelligibility), question (deliberation), and influence (control) the systems that increasingly mediate their choices — and the risks those systems entail. Trust, in this view, is not a static property but a co-constructed relationship, one that requires sustained epistemic engagement on both sides.

II. Recommender Systems and the Limits of Internal Trustworthiness

Recommender systems are typically powered by machine learning algorithms trained on large volumes of user data to optimize predefined metrics such as click-through rate, watch time, or engagement. These systems do not rely on explicit rules or symbolic reasoning, but instead generate predictions through statistical correlations learned from past behavior. As a result, their outputs are adaptive (they evolve in response to user behavior, e.g., showing more extreme content after repeated engagement), opaque (the logic driving recommendations is distributed across high-dimensional representations, making it difficult to trace why a particular item was shown), and often unpredictable (subtle changes in input or model parameters can yield divergent outputs) — even for their developers [1]. This raises a fundamental question: how can we meaningfully speak of trustworthiness when users are unable to observe, question, or influence how the system behaves? Much of the recent work on trustworthy AI focuses on internal features such as fairness constraints, robustness under distributional shift, or post hoc explainability. While these are essential for regulatory compliance and technical soundness, they are insufficient for fostering user-level trust. They frame trust as a property that can be engineered into the system, rather than as a relationship co-constructed through ongoing interaction [2]. This internalist framing also underestimates the persuasive nature of recommendations. Unlike traditional tools, recommender systems do not merely reflect users' preferences — they actively shape them over time. This becomes especially problematic when users are unaware that their choices train the system or interpret recommendations as neutral suggestions rather than goal-oriented nudges. In such contexts, opacity is not a minor usability issue; it is an epistemic vulnerability [3]. Users who cannot anticipate how a system will respond are more likely to defer to it — even when its outputs are skewed, manipulative, or harmful. Internal trustworthiness, while necessary, is therefore insufficient. It must be complemented by mechanisms that foster user understanding, critical evaluation, and the capacity to intervene. In the next section, we examine how current research in explainability, persuasion, and trust addresses these challenges — and where it falls short.

III. Related Work: Explainability, Persuasion, and Trust

Research on trustworthy recommender systems spans several overlapping domains. In this section, we focus on three strands particularly relevant to our argument: explainability, persuasion, and trust. These correspond to the core capacities that we argue users must develop in adaptive algorithmic environments: the ability to understand, evaluate, and influence system behavior. While each area offers valuable contributions, we show that they often fall short of supporting sustained user agency. Educational requirements, as we propose them, help fill that gap.

A. Explainability and the Limits of Interpretation

Explainability research aims to make model outputs interpretable to users, typically by providing textual or visual rationales for why a given item was recommended. These approaches fall into two categories: model-based explanations (integrated into the recommendation logic) and post hoc explanations (generated after the fact). While these techniques can improve transparency at the surface level, they rarely support deeper understanding of how user behavior shapes recommendations or how model parameters evolve over time. As Zhang and Chen note, most explainable recommendation methods are evaluated in terms of perceived trust or satisfaction — not actual comprehension [4]. Educational requirements shift the goal from perception to epistemic engagement: understanding is treated not as a byproduct, but as a design objective in itself.
.

B. Persuasion and the Need for Critical Awareness

Recommender systems do not merely reflect user preferences — they also shape them. As persuasive technologies, they operate through attention-capture strategies, behavioral nudges, and reinforcement loops. This dynamic has been criticized for bypassing rational deliberation and undermining user autonomy. Susser, Roessler, and Nissenbaum argue that personalization systems can operate as mechanisms of manipulation, especially when their influence is opaque or unaccountable [3]. This concern recalls Aristotle's warning that rhetoric is not only a tool for persuasion, but also a practice that must be scrutinized and resisted [5]. Yet most recommender systems are optimized for engagement while offering users no means to interrogate the system or contest its effects. Educational requirements address this imbalance by embedding prompts for reflection and opportunities for resistance into the interaction flow.

C. Trust and the Incomplete Account of the User

Most frameworks for trustworthy AI emphasize internal system attributes — fairness, robustness, transparency — but offer only a thin account of the user. The user is typically seen as a recipient of outputs, not as a co-constitutor of the trust relation. However, as Binns has shown, even principles like fairness are normatively contested and must be interpreted through political and moral reasoning [6]. Meanwhile, regulatory efforts such as the EU's Digital Services Act invoke user empowerment, but tend to operationalize it through auditability and interface-level choice architectures. These efforts rarely equip users with the means to evaluate or shape the logic of personalization itself. Educational requirements help close this gap by enabling users to engage critically with the dynamics of trust — not just once, but across time.

Across these three domains — explainability, persuasion, and trust — we find that the user is often under-empowered: given little support in understanding how the system works, in resisting undue influence, or in negotiating the normative terms of interaction. Educational requirements help fill this gap by equipping users with the means to understand, evaluate, and influence how trust is formed, maintained, or withdrawn in relation to adaptive AI systems. We argue that these requirements must be treated as essential components of trustworthy system design — not as optional enhancements, but as foundational to any meaningful account of user-centered AI.

IV. Educational Requirements: Definition and Rationale

If recommender systems are to be genuinely trustworthy, they must do more than optimize outputs or comply with internal safeguards. They must support users in becoming informed, reflective, and capable participants in the algorithmic environments that shape their choices. This requires a new class of requirements — educational requirements — that embed learning, critical awareness, and agency directly into system design. We define educational requirements as design and evaluation criteria aimed at enabling users to understand how a system works, evaluate its influence, and deliberately modulate or resist its behavior. These are not peripheral features or usability enhancements; they are central to ensuring that trust is not passive but actively constructed through ongoing interaction. In adaptive systems, where personalization is continuous and opaque, the absence of such requirements leaves users exposed to automated influence without epistemic footing. The rationale for educational requirements is grounded in the dual role recommender systems increasingly play: they are not only content delivery mechanisms but also persuasive agents that shape user attention, reinforce preferences, and influence belief formation over time. In such contexts, opacity is not a minor design flaw — it is a structural asymmetry. Traditional explainability mechanisms may clarify what was recommended, but rarely illuminate why recommendations evolve or how user actions shape them. Educational requirements respond to this asymmetry by embedding support for iterative sense-making: mechanisms that enable users to form, test, and revise mental models of system behavior. Rather than offering isolated justifications or static transparency tools, these mechanisms aim to cultivate a user who can track system dynamics across changing contexts and interact with them reflectively. The interface becomes not just a site of exposure, but a site of epistemic engagement. At the core of educational requirements is a threefold commitment to the user as an epistemic agent: • Intelligibility: the ability to understand how the system behaves and adapts • Deliberation: the capacity to evaluate its influence and relevance • Control: the means to intervene in or redirect recommendation trajectories These capacities form the foundation of a relational model of trust — one in which users are not merely protected from harm, but equipped to engage with algorithmic systems on their own terms.

V. Operationalizing Educational Requirements

To be effective, educational requirements must be embedded explicitly at each stage of the system development lifecycle: design, implementation, and evaluation. This section outlines how educational objectives can be translated into concrete mechanisms that support user intelligibility, deliberation, and control.

A. Design Phase: Articulating Learning Objectives

In the design phase, educational requirements should be specified alongside functional and ethical goals. Designers must consider what users should learn about the system, how this learning will be supported, and how users' epistemic capacities might evolve over time. Educational intent should be formalized into design questions such as: What mental models will the system help users develop? and What interpretive tools will they need to retain agency in adaptive settings?

B. Implementation Phase: Interface-Level Mechanisms

Educational requirements can be realized through specific interaction mechanisms that transform the user interface into a learning environment. Examples include: • Explanatory overlays: contextual indicators that surface how and why recommendations are generated • What-if controls: tools that allow users to modify input signals or simulate alternative outcomes • Diversity nudges: prompts that raise awareness of content concentration and suggest epistemic expansion • Goal-setting prompts: options that allow users to align recommendations with their stated intentions These mechanisms do not aim to reduce complexity, but to help users navigate it meaningfully. They recast the interface from a site of passive content delivery into a space for critical interaction.

C. Evaluation Phase: Measuring Epistemic Agency

Evaluation must go beyond standard usability metrics. Educational requirements demand assessment criteria that reflect user understanding and agency. These may include:

- **The user's ability to predict system behavior**
- **Their capacity to articulate the system's logic**
- **Their willingness or ability to adjust their interaction strategies over time**

Behavioral indicators — such as increased diversity-seeking or strategic modulation of preferences — can serve as proxies for epistemic resilience. By integrating educational requirements across the development lifecycle, designers can close the gap between system behavior and user understanding — a condition necessary not only for trust, but for meaningful user participation in algorithmic environments.

VI. Use Case: Echo Chambers and Behavioral Drift

The dynamics of behavioral drift and echo chamber formation are well documented on platforms such as YouTube and TikTok. These platforms use adaptive recommender systems that respond to patterns of engagement — clicks, views, watch time — to continuously recalibrate the content presented to users [7], [8]. Consider a user who watches a few videos expressing skepticism about climate policy. The system detects increased engagement with this genre and begins prioritizing similar content. Over time, the user's informational environment begins to drift — not randomly, but directionally, toward content that reinforces those initial signals. What results is not a deliberate search for confirmation, but an emergent narrowing of epistemic exposure. From the user's perspective, this drift is often imperceptible. Without clear indicators of why content is shown or how prior behavior shaped the feed, the platform appears to merely "understand" the user — rather than actively guiding their informational trajectory. The echo chamber, in such cases, is not a product of malicious design, but of missing affordances: the system offers no cues for reflection, no tools for questioning, and no means for course correction. Educational requirements could intervene in this process through a combination of interface mechanisms:

- **Explanatory aids**: visual maps or summaries that clarify how user activity influences recommendations.
- **Interactive controls**: adjustable parameters that allow users to shape or diversify their content flow.
- **Reflective prompts**: nudges that surface homogeneity or repetition in the feed and invite reassessment.

- **Feedback simulations**: tools that visualize how small changes in user behavior shift future outputs.

Such tools would not block the formation of echo chambers, but would give users the capacity to detect, evaluate, and resist drift. Educational requirements, in this context, serve not as constraints on personalization, but as instruments of epistemic resilience and reorientation.

VII. Discussion and Implications

Educational requirements reframe trust not as a static property of the system, but as a dynamic relationship constructed between user and system. This shift challenges internalist approaches to trustworthy AI, which often treat user understanding as peripheral or optional. In adaptive environments, where system behavior evolves in response to user input, this omission becomes a structural risk. From a requirements engineering perspective, educational requirements redefine what it means to specify and validate a system. Fixed properties — fairness constraints, robustness checks — remain necessary, but must be complemented by criteria that assess how users track, interpret, and influence system behavior over time. This requires new forms of evaluation that foreground epistemic outcomes: understanding, reflection, and the capacity to act. Although regulatory frameworks increasingly reference user empowerment, they rarely operationalize it meaningfully. Educational requirements provide a concrete pathway to do so — by making empowerment measurable through indicators of user intelligibility, deliberation, and control. They also raise new design challenges: What does it mean to support user sense-making? How can feedback loops be made visible without sacrificing usability or overwhelming attention? Finally, educational requirements imply a conceptual shift: from paternalism to co-learning. Trust becomes not an assumption or guarantee, but a process — one that unfolds through interaction, adjustment, and reflection. This move carries philosophical weight. Drawing a parallel from Kantian epistemology,[1] educational requirements restore the user not as a variable in optimization, but as a condition of intelligibility — without whom no trustworthy interaction is possible.

VIII. Conclusion

Trustworthy AI cannot be achieved solely through internal technical safeguards. While fairness, robustness, and explainability are essential, they are insufficient if users remain unable to understand, question, or shape the systems influencing them. In adaptive environments like recommender systems, where user behavior continuously feeds back into algorithmic outputs, internal trustworthiness must be complemented by mechanisms that support epistemic agency. We introduced the concept of educational requirements as a framework for designing such mechanisms. These requirements aim to foster user understanding, deliberation, and control — not as peripheral enhancements, but as core components of trustworthy interaction. We outlined how educational requirements can be specified, implemented, and evaluated across the system lifecycle, and illustrated their value through the case of echo chambers in personalized content feeds. This approach reframes users not as passive recipients, but as co-learners — capable of interpreting, challenging, and shaping the systems they engage with. By embedding pedagogical strategies into design, we move toward systems that are not only safer and more transparent, but more intelligible and participatory. We invite the requirements engineering community to treat user understanding not as a secondary objective, but as a foundational condition for systems that deserve trust — and for users who deserve to understand them.

[1]We briefly draw a Kantian parallel to clarify this point. Internal trustworthiness treats the system as analogous to *understanding*: a mechanism that imposes order according to internal rules. The user, by contrast, plays the role of *sensibility*: the site of experiential input and responsiveness. In Kant's terms, "understanding without sensibility is empty; sensibility without understanding is blind." Similarly, trust cannot arise from system integrity alone. It requires user-facing intelligibility as a condition of co-constructed reliability.

References

[1] S. Milano, M. Taddeo, and L. Floridi, "Recommender systems and their ethical challenges," *AI & Society*, vol. 36, no. 1, pp. 59–70, 2021.

[2] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[3] D. Susser, B. Roessler, and H. Nissenbaum, "Online manipulation: Hidden influences in a digital world," *Georgetown Law Technology Review*, vol. 4, 2019.

[4] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," *Foundations and Trends in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2018.

[5] Aristotle, *Rhetoric*

[6] R. Binns, "Fairness in machine learning: Lessons from political philosophy," *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, pp. 149–159, 2018.

[7] M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr, "Auditing radicalization pathways on youtube," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 131–141.

[8] M. N. Ibrahim, K. Fairley, A. Molina, A. Azad, and S. Thakur, "Tiktok's recommendations skewed towards republican content during the 2024 u.s. presidential race," *Media Manipulation Casebook*, 2025. [Online]. Available: https://mediamanipulation.org/research/tiktoks-recommendations-skewed-towards-republican-content-\during-2024-us-presidential