Transparency Features in Robo-Advice: Evidence, Mechanisms, and EU Policy Implications for Trust, Reliance, and Financial Inclusion

Inga Jonaityte[a]

[a] *Ca' Foscari University of Venice, Venice School of Management, Italy*

Abstract

Digital investment advice increasingly combines automated recommendations with varying degrees of human support. Policy and design debates often assume that "more transparency"—explanations, disclosures, uncertainty communication, performance information, or human oversight cues—will unambiguously raise trust and improve consumer outcomes. This position/review paper maps the mechanisms through which transparency features can help or harm retail users, and synthesizes experimental and quasi-experimental evidence with a focus on outcomes relevant to trust/legitimacy, behavioral reliance, and access-to-human-support implications.

Methodologically, we implement a Scopus-based scoping-and-extraction protocol (1,000 candidate records retrieved; 101 machine-extracted records after user-defined screening), and then restrict substantive claims to a full-text verified subset of studies for which PDFs were obtained (N=7). The verified evidence suggests systematic trade-offs: (i) providing performance information can increase perceived trustworthiness but may backfire when presented in

*Draft v1.2 (2026-02-18). Working paper — do not cite without permission.

cognitively demanding formats that reduce advice-taking; (ii) human-in-the-loop designs can increase uptake, yet may reduce decision accuracy and shift responsibility perceptions; and (iii) trust is highly sensitive to experienced accuracy, with "bad advice" producing discrete drops in trusting intentions relative to near-perfect advice, consistent with asymmetric trust updates.

We integrate these findings with EU sectoral and cross-cutting governance constraints to derive testable propositions and research design implications for future survey and experimental research. The paper's core design recommendation is *calibrated transparency*: layered, low-burden disclosure that improves understanding and accountability while preserving user agency and access to meaningful human support.
Keywords: Robo-advice, transparency, explanations, disclosure, uncertainty communication, human oversight, trust, advice-taking, financial inclusion, MiFID II, AI Act

---

1. Motivation and scope

Robo-advisors and more broadly "AI-mediated advice" now appear throughout retail finance: portfolio recommendations, risk profiling, pension dashboards, and hybrid advisory workflows. In parallel, regulators emphasize transparency, fairness, and accountability in digital finance, while practitioners deploy interface features intended to increase adoption and engagement. A common narrative is linear: add explanations and disclosures $\rightarrow$ users trust more $\rightarrow$ users make better decisions. The evidence and mechanisms are less linear. Transparency can *educate* but also *overwhelm*; it can *signal competence* but also *highlight uncertainty* in ways that trigger aversion; it can

clarify accountability but also *invite overreliance* by creating a false sense of safety.

This paper targets a narrow, policy-relevant question aligned with this scope: among retail consumers/retail investors, what is the effect of adding transparency features to robo-advice on (a) trust-related outcomes and (b) behavioral reliance outcomes? We focus on five feature families commonly proposed in robo-advice interfaces:

1. **Explanation/rationale** (why this portfolio / why this risk score);
2. **Disclosure** of fees, risks, conflicts, and limitations;
3. **Uncertainty communication** (confidence bands, scenario ranges, "model is unsure");
4. **Model performance information** (historical accuracy, backtests, benchmark comparisons);
5. **Human oversight / escalation information** (human-in-the-loop, review, ability to request a human).

2. Conceptual framework: why transparency works, and why it fails

We treat "transparency" as an *intervention on beliefs* under limited attention. A single feature can shift multiple latent beliefs at once: perceived competence, perceived benevolence, perceived integrity, perceived control, and perceived accountability. These beliefs map into two outcome families:

- **Trust outcomes**: trust/credibility, perceived transparency/understandability, perceived fairness/legitimacy.

- **Reliance outcomes**: advice-taking/compliance, delegation/override behavior, adoption/usage, switching to human advice, portfolio/risk-taking choices.

2.1. Mechanism map

We organize mechanisms into six channels (testable in experiments and surveys):

1. **Comprehension channel**: transparency improves understanding of inputs and outputs; effects depend on cognitive load and literacy.
2. **Competence signal**: performance metrics and explanations serve as cues of skill, potentially increasing trust.
3. **Uncertainty salience**: uncertainty communication can calibrate expectations but also increase perceived risk and reduce reliance.
4. **Accountability and responsibility**: human oversight cues can shift perceived accountability and affect moral comfort with delegation.
5. **Procedural fairness and legitimacy**: disclosure of rules and process can increase perceived fairness even without changing outcomes.
6. **Agency and control**: features that preserve user choice (override, escalation) can increase acceptance; excessive automation can trigger reactance.

3. Evidence base and review method

3.1. Scopus-based scoping and extraction (lead generation, not evidence)

We used a structured Scopus query over robo-advice/algorithmic advice with transparency-related keywords and experimental identification terms.

The query yielded 1,000 candidate records for screening in this review. Screening columns were defined to enforce (i) human participants, (ii) retail consumer/investor relevance, (iii) a comparison group (baseline robo and/or human/hybrid), and (iv) experimental or quasi-experimental designs. We then produced structured, machine-extracted fields for 101 records (based on the workflow's eligibility classification and available text). This dataset is useful for triage, but *machine-extracted output is not treated as evidence.*

3.2. Full-text verification (evidence used in this draft)

To avoid unverified claims, we restrict substantive empirical statements to studies for which we obtained and read full-text PDFs. In the current draft, this yields 7 verified papers spanning lab/online experiments and a field experiment. Table 1 summarizes these studies and maps them to transparency-feature families.

4. Synthesis by transparency feature family

4.1. Model performance information and accuracy cues

A recurring theme is that users respond strongly to cues about algorithmic *competence*—but the presentation format and experiential context matter.

Presentation format: information can help trust but hurt reliance.. In an experimental comparison of algorithmic versus human advice, You et al. [14] show that presenting algorithm prediction performance in an aggregated format can increase trust without reducing advice-taking, whereas an elaborated format can increase cognitive load and reduce "algorithm appreciation" and advice-taking. The result supports a comprehension-cost mechanism: holding

Table 1: Full-text verified evidence base used for empirical claims in this draft (N=7).

Study	**Year**	**Design**	**Transparency feature(s) and main outcome(s)**
Germann & Merkle, "Algorithm aversion in delegated investing" [4]	2019/2022	Experiment	Information about algorithmic management in delegated investing; adoption/take-up and delegation choices.
Niszczota & Kaszás, "Robo-investment aversion" [6]	2020	Online experiment	Algorithmic vs human investment management framing; trust and choice between robo vs human.
Daschner & Obermaier, "Algorithm aversion?" [1]	2022	Online experiments	Advice accuracy manipulation (nearly-perfect vs bad) as a performance/quality signal; trusting beliefs and intentions.
You, Yang & Li, "Does presenting prediction performance matter?" [14]	2022	Online experiments	Performance information format (aggregated vs elaborated) for algorithmic vs human advice; trust and advice-taking.
Sele & Chugunova, "Putting a human in the loop" [12]	2023/2024	Laboratory/online experiments	Human-in-the-loop cue increases uptake but decreases accuracy of automated decisions.
Yang et al., "My advisor, her AI, and me" [13]	2025	Field experiment	Disclosure/framing of AI assistant within advisory workflow; investor outcomes and satisfaction.
Kramer et al., "On-	2024	Randomized	Explanatory dashboard

informational content fixed, presentation complexity can shift reliance in opposite directions.

Experienced accuracy: asymmetric trust updates.. Daschner and Obermaier [1] experimentally vary advice accuracy and observe that trusting intentions are highly sensitive to "bad advice." A key empirical regularity is an asymmetry: near-perfect advice does not necessarily increase trust relative to a moderate baseline, but low accuracy reduces trust markedly. This pattern is consistent with loss-aversion-like updating in trust formation (negative surprises dominate).

Delegated investing: adoption and control.. In delegated investing settings, performance and process cues interact with delegation comfort. Germann and Merkle [4] examine algorithm aversion in delegated investing, focusing on adoption/take-up decisions and the willingness to delegate to algorithmic management. While the design and outcomes differ from standard robo-advisor UI tests, the paper highlights that consumers treat algorithmic delegation as a qualitatively distinct act, with separate concerns about control and accountability.

4.2. Human oversight and human-in-the-loop cues

Human oversight is often positioned as "reassurance" (someone is monitoring the machine). Evidence indicates that oversight cues can increase uptake, but can also degrade performance by altering incentives and responsibility perceptions.

Uptake–accuracy trade-off.. Sele and Chugunova [12] provide direct evidence: adding a human-in-the-loop increases uptake of automated recommendations

but decreases decision accuracy. For policy and product design, this highlights a trade-off: "hybrid" is not automatically superior; it can induce overreliance or reduce the discipline of algorithmic decision rules.

Hybrid framing in the field.. In a retail-advice field experiment, Yang et al. [13] study the consequences of integrating an AI component into advisor interactions. The field setting strengthens external validity for take-up and behavioral outcomes. The results are consistent with the view that *how* the AI component is positioned (assistant vs advisor; disclosure and social cues) can meaningfully affect outcomes, implying that transparency is closely intertwined with *framing*.

4.3. Explanation/rationale and informational dashboards

Explanation features are rarely "just" explanation: they can alter the salience of risks, future expectations, and confidence in one's knowledge.

Kramer et al. [5] analyze online pension dashboards in a randomized survey experiment. While the intervention is not robo-advice per se, the study provides a relevant analogue: explanatory dashboards and default exposure can shift pension knowledge and expectations, which are upstream determinants of trust and reliance in retirement decisions.

4.4. Baseline preferences: robo-investment aversion

Transparency features operate against baseline preferences. If there is strong default aversion to robo-management, even highly transparent interfaces may not close the gap.

Niszczota and Kaszás [6] document "robo-investment aversion"—a preference for human over algorithmic investment management in experimental

choice tasks. This baseline aversion matters for interpretation: transparency interventions may be insufficient if the binding constraint is perceived legitimacy, moral comfort, or identity-based distrust of algorithms.

5. Moderators and heterogeneity: who benefits, who is left out?

Across the verified studies, moderators that plausibly govern transparency effectiveness include:

- **Financial and digital literacy**: complex performance metrics or uncertainty displays may widen gaps between high- and low-literacy users, raising inclusion concerns.
- **Prior experience**: experienced investors may treat transparency as diagnostic information; novices may treat it as noise.
- **Baseline trust and attitudes**: "algorithm appreciation" vs "algorithm aversion" predispositions shape both trust and compliance responses [14, 1].
- **Availability of human support**: the option to escalate can increase comfort, but may also change responsibility attributions and accuracy [12].

These moderators motivate an emphasis on inclusion: transparency features should be evaluated not only in terms of average effects, but also in terms of whether they reduce or exacerbate inequality in access to effective advice.

6. EU regulatory and policy relevance

Retail investment advice in the EU is governed by suitability and disclosure obligations (MiFID II) and standardized product information (PRIIPs KIDs) [7, 8]. Supervisors have explicitly examined automation in financial advice and the consumer risks associated with digital channels [3, 11]. Beyond sectoral rules, data protection and AI governance introduce cross-cutting constraints:

- **GDPR Article 22** concerns automated individual decision-making and underscores safeguards such as human intervention and contestability [9, 2].
- **EU AI Act** introduces transparency and human oversight requirements for certain AI systems, shaping what forms of disclosure and oversight cues may be mandated or expected [10].

From a design perspective, the regulatory landscape encourages more informative interfaces. From a behavioral perspective, the verified evidence cautions that adding information without attention to cognitive burden and responsibility allocation can backfire.

7. Testable propositions (mechanism-first)

We state propositions as falsifiable claims suitable for survey hypotheses and experimental tests.

P1 (format-sensitive transparency).. Holding informational content fixed, *lower cognitive-burden* performance disclosures increase trust and do not reduce advice-taking; *higher burden* disclosures reduce advice-taking via cognitive load [14].

P2 (asymmetric trust updating).. Negative performance surprises produce larger absolute changes in trust and reliance than positive surprises of equal magnitude [1].

P3 (oversight reassurance vs overreliance).. Human-in-the-loop cues increase uptake of automated recommendations, but can decrease objective decision quality by shifting perceived responsibility and increasing overreliance [12].

P4 (inclusion constraint).. The net effect of transparency features on welfare is more positive for high literacy/experience users than for low literacy/experience users, unless transparency is explicitly "inclusive-by-design" (layering, plain language, actionable escalation). [CHECK: This proposition is stated without a supporting verified citation in the current draft.]

8. Implications for future research

8.1. Survey measurement modules

The mechanism map implies that survey-based work should treat "trust" as a multi-dimensional construct rather than a single attitude. In particular, measurement should separately capture perceived competence, benevolence, and integrity, alongside perceived transparency/understandability and perceived fairness/legitimacy. These facets are conceptually distinct and should be measured separately.

In parallel, surveys should elicit reliance-related proxies that correspond to how advice is used in practice. Beyond stated advice-following, this includes willingness to delegate versus retain control, willingness to override recommendations, switching propensity (including a preference for human advice),

and willingness-to-pay for human support. To support interpretation, it is also useful to measure exposure and interpretation: whether respondents remember seeing performance metrics, uncertainty displays, fee/risk disclosures, and information about human oversight or escalation.

Finally, to keep an inclusion lens, surveys should pre-specify moderators that shape both comprehension and uptake. At minimum, this includes objective and subjective financial literacy, digital literacy, and prior investing experience. Where feasible, it should also capture access constraints that plausibly condition exposure and use, such as language barriers, disability-related constraints, and the practical availability of support.

8.2. Experimental design suggestions

Future designs are likely to be more diagnostic when they vary *format* and *actionability*, not only the informational content of transparency features. One natural contrast is aggregated versus elaborated performance disclosure formats while holding the stated level of performance fixed. A second contrast is uncertainty communication that includes explicit guidance on how to act under uncertainty (e.g., what to do when uncertainty is high) versus non-actionable ranges that merely display variability.

A third design lever is human escalation. Experiments can vary whether escalation availability is credibly implemented and frictionless (i.e., easily discoverable and usable) versus nominal or opaque. Finally, a fairness/legitimacy module can be incorporated to test whether transparency features increase perceived procedural legitimacy independently of realized outcomes, thereby separating legitimacy effects from performance beliefs.

9. Limitations and next steps

This draft is deliberately conservative: it uses automated discovery and structured extraction for triage, but limits claims to full-text verified studies available in PDF at the time of writing. The next step for a PRISMA-ready systematic review is to (i) expand full-text retrieval and verification beyond the current subset, (ii) harmonize outcome measures and compute comparable effect sizes where possible, and (iii) pre-specify moderator and risk-of-bias analyses.

10. Conclusion

Transparency in robo-advice is not a monotone "more is better" design choice. The most consistent pattern in the verified evidence is a set of *trade-offs*: information format can flip the sign of reliance effects; oversight cues can increase uptake at the cost of accuracy; and trust responds asymmetrically to negative performance. For EU policy and responsible FinTech design, this implies that transparency obligations should be operationalized as *calibrated transparency*: layered, comprehensible, and paired with meaningful human support and contestability.

References

[1] Daschner, S., Obermaier, R., 2022. Algorithm aversion? On the influence of advice accuracy on trust in algorithmic advice. Journal of Decision Systems 31, 77–97. URL: http://dx.doi.org/10.1080/12460125.2022.2070951, doi:10.1080/12460125.2022.2070951.

[2] (endorsed by the European Data Protection Board), A..W.P., 2018. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (rev.01). Technical Report. European Commission. URL: https://ec.europa.eu/newsroom/article29/redirection/item/612053.

[3] Joint Committee of the European Supervisory Authorities (EBA, ESMA, E., 2018. Joint Committee Report on the results of the monitoring exercise on automation in financial advice. Technical Report. European Supervisory Authorities. URL: https://www.esma.europa.eu/sites/default/files/library/jc_2018_29_-_jc_report_on_automation_in_financial_advice.pdf.

[4] Germann, M., Merkle, C., 2022. Algorithm aversion in delegated investing. Journal of Business Economics 93, 1691–1727. URL: http://dx.doi.org/10.1007/s11573-022-01121-9, doi:10.1007/s11573-022-01121-9.

[5] Kramer, M., Lensink, R., Plantinga, A., 2024. Do online pension dashboards affect pension knowledge and expectations? Evidence from a randomized survey experiment. Journal of Pension Economics and Finance 23, 504–527. URL: http://dx.doi.org/10.1017/s1474747224000040, doi:10.1017/s1474747224000040.

[6] Niszczota, P., Kaszás, D., 2020. Robo-investment aversion. PLOS ONE 15, e0239277. URL: http://dx.doi.org/10.1371/journal.pone.0239277, doi:10.1371/journal.pone.0239277.

[7] Parliament, E., of the European Union, C., 2014a. Directive 2014/65/eu on markets in financial instruments (mifid ii). Official Journal of the European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014L0065.

[8] Parliament, E., of the European Union, C., 2014b. Regulation (eu) no 1286/2014 on key information documents for packaged retail and insurance-based investment products (priips). Official Journal of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2014/1286/oj/eng.

[9] Parliament, E., of the European Union, C., 2016. Regulation (eu) 2016/679 (general data protection regulation). Official Journal of the European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[10] Parliament, E., of the European Union, C., 2024. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[11] Securities, E., Authority, M., 2023. Guidelines on certain aspects of the MiFID II suitability requirements. Technical Report ESMA35-43-3172. ESMA. URL: https://www.esma.europa.eu/sites/default/files/2023-04/ESMA35-43-3172_Guidelines_on_certain_aspects_of_the_MiFID_II_suitability_requirements.pdf.

[12] Sele, D., Chugunova, M., 2023. Putting a human in the loop: Increasing

uptake, but decreasing accuracy of automated decision-making. PLoS ONE URL: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0298037&type=printable, doi:10.1371/journal.pone.0298037.

[13] Yang, C.L., Bauer, K., Li, X., Hinz, O., 2026. My Advisor, Her AI, and Me: Evidence from a Field Experiment on Human–AI Collaboration and Investment Decisions. Management Science 72, 242–264. URL: http://dx.doi.org/10.1287/mnsc.2022.03918, doi:10.1287/mnsc.2022.03918.

[14] You, S., Yang, C.L., Li, X., 2022. Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? Journal of Management Information Systems 39, 336–365. URL: http://dx.doi.org/10.1080/07421222.2022.2063553, doi:10.1080/07421222.2022.2063553.