Does transparency about AI limitations or non-human identity improve or worsen user trust and decision-making, and what evidence exists for perverse effects such as moral licensing, compliance pressure, or reduced engagement following AI self-disclosure?

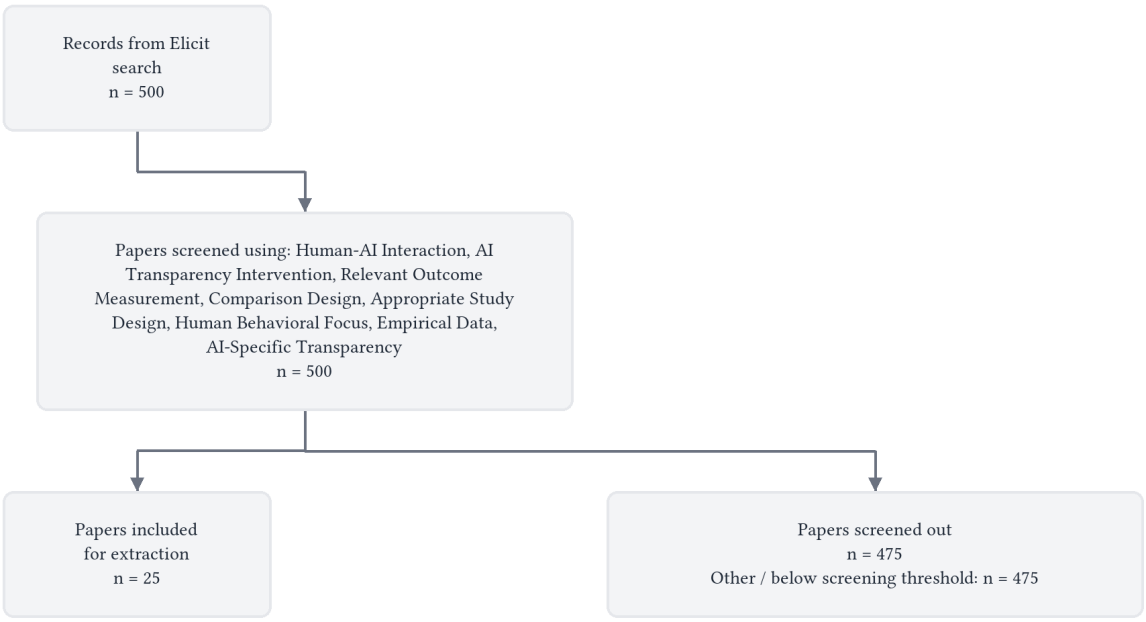
Transparency about AI limitations improves trust calibration and decision quality in high-stakes decision support contexts but identity disclosure severely damages trust and outcomes in consumer settings, with perverse effects including over-trust, compliance pressure, cognitive overload, and discrimination against AI appearing in over half of studies.

Abstract

Transparency about AI limitations and non-human identity produces markedly heterogeneous effects on trust and decision-making, ranging from substantial benefits to severe harms depending on transparency content, domain context, user characteristics, and implementation details. Disclosure of AI limitations and capability boundaries consistently improves trust calibration and performance in high-stakes decision support contexts, with effect sizes reaching d=1.78 for trust resilience [1] and appropriate reliance rates of 76.4% versus 34.2% in non-disclosure conditions [1]. Such transparency reduces cognitive biases by 37-41% [1] and improves risk-adjusted investment returns by 7.2% [1]. Conversely, disclosure of AI identity without accompanying capability information severely damages trust and behavioral outcomes in consumer contexts, reducing purchases by 79.7% [2] and lowering trust scores significantly (M=4.81 vs. 5.26) [3] due to perceptions of reduced knowledge and empathy [2]. Feature-based explanations show mixed-to-negative effects, decreasing decision accuracy by 8.9-9.8% [4] through confirmation bias and mental model rigidity [4].

Perverse effects emerged in 13 of 25 studies (52%), with over-trust and compliance pressure representing the most common consequences. Transparency paradoxically induces over-trust that crowds out unique human knowledge [5], with agreement rates reaching 90.7% even when AI lacks critical information [5]. Users continue accepting AI advice despite transparency revealing warnings against the algorithm [6]. Cognitive overload from excessive transparency reduces AI adoption through an inverted U-shaped relationship [7, 8], while detailed rationales reduce perceived autonomy (d=-0.19) [9]. Discrimination against disclosed AI intensifies in creative and social domains, with both human and algorithmic evaluators penalizing AI assistance [10, 11]. Trust calibration support produces the counterintuitive effect of increased AI disuse despite reducing misuse [12]. However, these heterogeneous outcomes reflect systematic variation rather than contradictory effects: limitation-focused transparency benefits high-stakes decision support with sophisticated users, while identity disclosure requires social presence cues and capability information in low-stakes consumer contexts to avoid triggering anthropomorphic expectation violations and speciesism-driven trust decline [3].

Flow Diagram



Paper search

We performed a semantic search across over 138 million academic papers from the Elicit search engine, which includes all of Semantic Scholar and OpenAlex.

We ran this query: "Does transparency about AI limitations or non-human identity improve or worsen user trust and decision-making, and what evidence exists for perverse effects such as moral licensing, compliance pressure, or reduced engagement following AI self-disclosure?"

The search returned 500 total results from Elicit.

We retrieved 500 papers most relevant to the query for screening.

Screening

We screened in sources based on their abstracts that met these criteria:

- **Human-AI Interaction**: Does this study involve human participants interacting with AI systems?
- **AI Transparency Intervention**: Does this study include interventions involving AI transparency about limitations, capabilities, or non-human identity disclosure?
- **Relevant Outcome Measurement**: Does this study measure trust, decision-making outcomes, or behavioral changes as primary or secondary outcomes?

- **Comparison Design**: Does this study include comparison groups or baseline measurements to assess the impact of transparency interventions?
- **Appropriate Study Design**: Is this study an experimental, quasi-experimental, observational study, systematic review, or meta-analysis (rather than a theoretical paper, opinion piece, or commentary)?
- **Human Behavioral Focus**: Does this study examine human behavioral outcomes rather than focusing solely on technical AI performance without human interaction?
- **Empirical Data**: Does this study present empirical data rather than being purely theoretical, opinion-based, or commentary without data?
- **AI-Specific Transparency**: Does this study examine specific AI transparency elements rather than only general technology acceptance without AI-specific components?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

At abstract screening, the number of papers excluded for each primary reason was:

- **Other / below screening threshold**: n = 475

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **AI Transparency Type**:

  Extract details about what type of AI transparency or disclosure was tested, including:

- Disclosure of AI limitations (confidence levels, uncertainty, error rates, capability boundaries)
- Disclosure of AI identity (revealing non-human nature vs. concealing it)
- Explanation of AI reasoning or decision process
- Interactive transparency (user feedback mechanisms, control options)
- Any combination of transparency types
- Specific content and format of transparency information provided to users

- **Trust Outcomes**:

  Extract all trust-related outcomes specifically related to AI transparency effects, including:

- How trust was measured (behavioral, self-reported, calibrated trust)
- Direction of trust effects (increased, decreased, no change)
- Magnitude of effects with confidence intervals if available
- Trust in AI vs. trust in AI recommendations vs. trust in human-AI collaboration
- Changes in appropriate reliance vs. over-reliance vs. under-reliance
- Trust calibration (alignment between trust and AI actual performance)

- **Decision-Making Effects**:

  Extract all decision-making and performance outcomes following AI transparency disclosure, including:

- Decision accuracy, quality, or effectiveness measures
- Speed or efficiency of decision-making
- User compliance with AI recommendations
- Human-AI collaboration performance

- Task completion rates or success metrics
- Changes in decision strategy or process
- Any measures of decision confidence or certainty

- **Perverse Effects**:

  Extract evidence of any negative or unintended consequences of AI transparency, specifically:

- Moral licensing (users becoming less careful after disclosure)
- Compliance pressure (feeling obligated to follow AI despite concerns)
- Reduced engagement or motivation
- Increased anxiety or cognitive burden
- Overconfidence or complacency
- Discrimination or bias against AI
- Decreased willingness to use AI systems
- Any paradoxical effects where transparency backfired

- **AI System Context**:

  Extract characteristics of the AI system and task domain relevant to transparency effects, including:

- Type of AI system (decision support, content moderation, predictive, conversational, etc.)
- Domain of application (healthcare, hiring, content review, etc.)
- AI performance level (high-performing, low-performing, miscalibrated)
- Task complexity and stakes (high-risk vs. low-risk decisions)
- Whether AI was presented as human, AI, or ambiguous identity

- **User Characteristics**:

  Extract participant demographics and characteristics that may moderate transparency effects, including:

- Sample size and demographics (age, education, tech experience)
- Prior AI experience or familiarity
- Domain expertise relevant to the task
- Cognitive abilities or individual differences mentioned
- Any subgroup analyses showing differential effects of transparency
- Recruitment source and setting (lab, online, workplace, etc.)

- **Study Design**:

  Extract methodological details necessary for assessing study quality and generalizability, including:

- Experimental design (between-subjects, within-subjects, factorial)
- Control/comparison conditions (no transparency, human comparison, etc.)
- Randomization and blinding procedures
- Study setting (laboratory, field, online)
- Duration of interaction or follow-up period
- Key confounds controlled for or limitations acknowledged

- **Causal Mechanisms**:

  Extract any evidence or theoretical explanations for why transparency improved or worsened outcomes, including:

- Cognitive mechanisms (cognitive load, mental models, understanding)
- Affective mechanisms (anxiety, confidence, perceived competence)
- Behavioral mechanisms (attention, effort, strategy changes)
- Social mechanisms (anthropomorphism, perceived agency, social pressure)
- Authors' explanations for observed effects
- Mediator or moderator analyses showing causal pathways

Results

Characteristics of Included Studies

The review identified 25 studies examining the effects of AI transparency on user trust and decision-making across diverse domains and transparency types. Table 1 presents the key characteristics of included studies.

Study	Full Text Retrieved?	Design	Sample Size	Domain	Transparency Type	Primary Outcomes
Tobias Rieger et al., 2023	No	Between-subjects online experiment [13]	128 [13]	Healthcare (simulated medical task) [13]	Disclosure of AI limitations (error-prone color) [13]	Trust behavior, decision deviation [13]
Zenan Chen et al., 2025	Yes	3×2 factorial between-within subjects [5]	752 [5]	Hiring decisions [5]	Explanation of AI reasoning (brief vs. extensive) [5]	Agreement with AI, unique human knowledge utilization [5]
Hou Tsung-Yu et al., 2024	No	3×3 between-subjects [14]	207, 223 [14]	Gender-biased hiring [14]	Explanation richness (none, lean, rich) [14]	Bias perception, AI recommendation adoption [14]
Ancuta Margondai et al., 2026	No	Between-subjects [9]	557 [9]	Organizational decision-making [9]	Detailed rationales vs. basic recommendations [9]	Perceived autonomy (M=3.73 vs. 3.84, d=-0.19) [9]
Toan Khang Trinh et al., 2025	Yes	Longitudinal within-subjects [1]	428 [1]	Financial advisory [1]	System limitation disclosure, explainability features [1]	Trust resilience (M=4.76 vs. 3.28, d=1.78) [1], portfolio performance [1]

Study	Full Text Retrieved?	Design	Sample Size	Domain	Transparency Type	Primary Outcomes
Xueming Luo et al., 2019	No	Field experiment [2]	6,200+ [2]	Conversational commerce [2]	AI identity disclosure [2]	Purchase rates (79.7% reduction) [2], call length [2]
Divya K. Srivastava et al., 2024	Yes	Two-phase between-subjects [15]	180 (90 per phase) [15]	Space mission trajectory planning [15]	Information sharing for shared situation awareness [15]	Situation awareness, task performance, overreliance [15]
Pooja Prajod et al., 2026	Yes	3×2×2 mixed factorial [16]	40 [16]	News journalism [16]	Disclosure levels (none, one-line, detailed) [16]	Trust questionnaire scores, subscription rates, source-checking [16]
Philipp Schmidt et al., 2020	Yes	Full-factorial between-subjects [17]	200 [17]	Text classification (movie reviews) [17]	Word highlighting, confidence scores [17]	Behavioral trust (1-3 percentage point decrease) [17], decision accuracy [17]
V. M. Ngo et al., 2025	No	Within-subjects web experiment [7]	491 [7]	Fake news detection, friend recommendations [7]	Real, placebic, or absent transparency [7]	Trust, certainty, AI use intention [7]
V. M. Ngo et al., 2025a	No	Between-subjects web experiment [8]	491 [8]	Fake news detection, friend recommendations [8]	None, placebo, real transparency [8]	Trust, certainty, AI use intention [8]
Jaymari Chua et al., 2025	No	Within-subjects using synthetic personas [18]	32 [18]	Competitive gaming (StarCraft II) [18]	Superhuman capability disclosure [18]	Trust, fairness perception, strategic defeatism [18]
Sebastian Krügel et al., 2021	No	Online experiment [6]	Not reported	Ethical dilemmas [6]	Transparent vs. opaque algorithm [6]	Advice acceptance, overtrust [6]

Study	Full Text Retrieved?	Design	Sample Size	Domain	Transparency Type	Primary Outcomes
Jingshu Li et al., 2024	Yes	Between-subjects [12]	252 (126 per experiment) [12]	City photo recognition [12]	Confidence calibration states, trust calibration support [12]	Trust in AI accuracy, misuse, disuse [12]
Nika Mozafari et al., 2021	Yes	Between-subjects online [19]	257 [19]	Liability insurance chatbot [19]	Identity disclosure, expertise/weakness communication [19]	Trust in competence, benevolence, integrity [19]
Maria D. Molina et al., 2022	Yes	3×3×2 between-subjects online [20]	676 [20]	Content moderation (hate speech, suicidal ideation) [20]	Source disclosure (AI/human/both), interactive transparency [20]	Trust (attitudinal and behavioral) [20], user agency [20]
Guanglu Zhang et al., 2022	No	Factorial between-subjects [21]	Not reported	Human-AI cooperation [21]	Identity deception (”human” vs. AI) [21]	Decision acceptance, joint performance [21]
Monika Westphal et al., 2023	No	Three experimental studies [22]	Not reported	Workplace AI collaboration [22]	Decision control, explanations [22]	Trust, understanding, compliance [22]
Inyoung Cheong et al., 2025	Yes	2×3×3 between-subjects online [10]	1,970 human + 2,520 LLM ratings [10]	News writing evaluation [10]	AI assistance disclosure statements [10]	Quality perception (-0.112 to -0.133 points) [10]
Aaron Springer et al., 2019	Yes	Within-subjects counterbalanced [23]	74 [23]	Emotional personal informatics [23]	Progressive transparency (word highlighting) [23]	Trust expectations vs. experience [23]
Roberta De Cicco et al., 2025	Yes	Three between-subjects studies [3]	160, 214 [3]	Consumer chatbot sales [3]	Non-human identity disclosure timing and social presence [3]	Trust (M=4.81 vs. 5.26) [3], purchase intention [3]

Study	Full Text Retrieved?	Design	Sample Size	Domain	Transparency Type	Primary Outcomes
Kevin Bauer et al., 2024	No	Two-stage mixed-method [11]	Graduate design students [11]	Creative image generation [11]	GenAI involvement disclosure [11]	Creativity and quality perception [11], effort beliefs [11]
Holly L. Wilson et al., 2019	No	Between-subjects VR simulation [24]	Not reported	Autonomous vehicle moral decisions [24]	Transparent vs. opaque AI decision-making [24]	Distress from socio-demographic decision criteria [24]
Kevin Bauer et al., 2023	Yes	Two studies (within and between) [4]	607, 153 [4]	Investment and real estate decisions [4]	Feature-based XAI (LIME, SHAP) [4]	Decision accuracy (8.9-9.8% decrease) [4], mental model adjustments [4]
Ángel Alexander Cabrera et al., 2023	Yes	Between-subjects online [25]	225 [25]	Fake review, satellite, bird classification [25]	AI behavior descriptions on subgroups [25]	Human-AI accuracy, reliance calibration [25]

Eighteen of 25 studies (72%) had full texts available for detailed analysis. Studies spanned diverse domains including healthcare [13], hiring [5, 14], finance [1], commerce [2], journalism [16], content moderation [20], and creative work [11]. Sample sizes ranged from 32 [18] to over 6,200 [2], with most studies (68%) employing between-subjects experimental designs [3, 8–10, 12–15, 17, 19–21, 24, 25]. Field experiments were rare, with only one study conducting real-world sales interactions [2].

Transparency interventions tested fell into four primary categories: AI identity disclosure (revealing non-human nature) [2, 3, 10, 11, 18, 19, 21], limitation disclosure (error rates, confidence levels, capability boundaries) [1, 12, 13, 15, 17, 25], reasoning explanations (rationales, feature importance) [4, 5, 9, 14], and interactive transparency (user feedback mechanisms, control options) [15, 20, 22]. Several studies tested multiple transparency dimensions simultaneously [17, 20, 22].

Effects of AI Transparency on Trust

Trust outcomes exhibited substantial heterogeneity across studies, with effects varying by transparency type, measurement approach, and context. Table 2 presents quantitative trust findings where available.

Study	Transparency Condition	Trust Measure	Direction	Magnitude
Toan Khang Trinh et al., 2025	High vs. low explainability	Trust resilience score [1]	Increased	M=4.76 vs. 3.28, d=1.78 [1]
Toan Khang Trinh et al., 2025	Limitation disclosure vs. non-disclosure	Appropriate reliance patterns [1]	Increased	76.4% vs. 34.2% [1]
Ancuta Margondai et al., 2026	Detailed rationales vs. basic	Perceived autonomy [9]	Decreased	M=3.73 vs. 3.84, d=-0.19 [9]
Xueming Luo et al., 2019	Identity disclosure vs. non-disclosure	Purchase rates (behavioral trust) [2]	Decreased	79.7% reduction [2]
Nika Mozafari et al., 2021	Mere disclosure vs. non-disclosure	Self-reported trust [19]	Decreased	ΔTrust=-0.365, $p<0.1$ [19]
Nika Mozafari et al., 2021	Disclosure + expertise vs. mere disclosure	Self-reported trust [19]	Increased	ΔTrust=0.378, $p<0.1$ [19]
Roberta De Cicco et al., 2025	Disclosure vs. non-disclosure	Trust towards firm [3]	Decreased	M=4.81 vs. 5.26 [3]
Philipp Schmidt et al., 2020	Word highlighting	Behavioral trust (following AI) [17]	Decreased	1-1.5 percentage points [17]
Philipp Schmidt et al., 2020	Confidence scores	Behavioral trust [17]	Decreased	2.5-3 percentage points [17]
Inyoung Cheong et al., 2025	AI disclosure (LLM ratings)	Trustworthiness perception [10]	Decreased	-0.112 to -0.133 points [10]
V. M. Ngo et al., 2025	Moderate transparency	Trust and certainty [7]	Increased	Not quantified [7]
V. M. Ngo et al., 2025	Excessive transparency	Trust [7]	Decreased	Not quantified [7]
Guanglu Zhang et al., 2022	AI vs. "human" identity	Decision acceptance (behavioral trust) [21]	Increased for AI	Not quantified [21]
Sebastian Krügel et al., 2021	Transparent vs. opaque	Advice acceptance [6]	No change	Overtrust in both conditions [6]
Monika Westphal et al., 2023	Decision control	Trust [22]	Increased	Not quantified [22]
Ángel Alexander Cabrera et al., 2023	Behavior descriptions	Self-reported trust [25]	No change	Not significant [25]

The financial advisory study by Toan Khang Trinh et al. demonstrated the strongest positive effects, with high-explainability conditions yielding trust resilience scores nearly 1.5 standard deviations higher than low-explainability conditions (M=4.76 vs. 3.28, d=1.78) [1]. Transparency about system limitations produced dramatically improved trust calibration, with 76.4% of participants in transparent conditions exhibiting appropriate reliance patterns versus only 34.2% in non-disclosure conditions [1]. Institutional affiliation (β=0.68, $p<0.001$) and perceived competence (β=0.57, $p<0.001$) significantly influenced initial trust formation [1], while longitudinal analysis revealed three distinct trust trajectory classes: increasing (58.2%), stable (23.7%), and degrading (18.1%) over nine months [1].

Conversely, AI identity disclosure consistently produced negative trust effects across consumer contexts. Luo et al.'s field experiment with over 6,200 customers found that disclosing chatbot identity before conversation reduced purchase rates by 79.7% [2], driven by perceptions of AI as less knowledgeable and less empathetic despite objective performance equivalence [2]. Similarly, De Cicco et al. found that disclosing artificial agent identity significantly reduced trust towards the firm (M=4.81 in disclosure vs. M=5.26 in non-disclosure) [3]. Transparency penalties extended to creative domains, with both human and LLM evaluators consistently penalizing disclosed AI assistance in news writing (reductions of 0.112-0.133 points) [10].

Several studies revealed non-linear relationships between transparency and trust. Ngo et al. identified an inverted U-shaped effect where moderate transparency fostered trust and certainty, but excessive transparency induced cognitive overload and heightened scrutiny, reducing AI adoption [7, 8]. Trust and certainty served as significant mediators of transparency's effects on AI use intention [7, 8].

Trust calibration—the alignment between trust and AI actual performance—emerged as a critical but often neglected outcome. Schmidt et al. found that transparency features led participants to mistrust correct AI predictions while maintaining or increasing reliance on incorrect ones [17], suggesting transparency can paradoxically worsen trust calibration. Li et al. demonstrated that uncalibrated AI confidence (either overconfident or underconfident) led to both misuse and disuse [12], and while trust calibration support helped participants recognize uncalibration and reduced misuse, it simultaneously fostered distrust and caused greater disuse of AI [12].

Effects on Decision-Making and Performance

Transparency's impact on decision quality and human-AI collaboration outcomes showed marked context dependency. Table 3 summarizes performance effects across domains.

Study	Task Domain	Transparency Intervention	Performance Metric	Effect
Toan Khang Trinh et al., 2025	Financial investment	High explainability	Risk-adjusted returns (Sharpe ratio) [1]	AI-advised: 0.83 vs. self-directed: 0.65, $p<0.01$ [1]
Toan Khang Trinh et al., 2025	Financial investment	Transparency	Overconfidence bias reduction [1]	41.3% [1]
Toan Khang Trinh et al., 2025	Financial investment	Transparency	Disposition effect reduction [1]	37.8% [1]
Toan Khang Trinh et al., 2025	Financial investment	Transparency	Decision error rate [1]	Reduced from 0.38 to 0.17 [1]
Ángel Alexander Cabrera et al., 2023	Classification tasks	Behavior descriptions	Overall accuracy [25]	Improved in 2 of 3 tasks [25]
Divya K. Srivastava et al., 2024	Trajectory planning	Information sharing	Shared situation awareness [15]	Boosted [15]
Divya K. Srivastava et al., 2024	Trajectory planning	Information sharing	Task performance [15]	Improved [15]
Divya K. Srivastava et al., 2024	Trajectory planning	Information sharing	Overreliance on AI [15]	Reduced [15]
Kevin Bauer et al., 2023	Investment decisions	XAI explanations	Decision accuracy [4]	Decreased 8.9% (observing) to 9.8% (not observing) [4]

Study	Task Domain	Transparency Intervention	Performance Metric	Effect
Philipp Schmidt et al., 2020	Text classification	Transparency features	Decision accuracy [17]	Decreased [17]
Zenan Chen et al., 2025	Hiring decisions	AI reasoning display	Agreement with AI [5]	Increased to 90.7% (extensive) from 88.7% (baseline) [5]
Zenan Chen et al., 2025	Hiring decisions	AI reasoning display	UHK utilization [5]	Decreased (crowded out) [5]
Pooja Prajod et al., 2026	News consumption	Detailed AI disclosure	Subscription rates [16]	Decreased [16]
Pooja Prajod et al., 2026	News journalism	AI disclosure	Source-checking behavior [16]	Increased [16]
Hou Tsung-Yu et al., 2024	Biased hiring	Rich explanations	Bias recognition [14]	Improved, especially for females [14]

Financial advisory contexts demonstrated the most consistently positive performance effects. AI-advised portfolios using transparent explainability features outperformed self-directed investments by 7.2% on risk-adjusted returns (Sharpe ratio: 0.83 vs. 0.65, $p<0.01$) [1]. Transparency reduced overconfidence bias by 41.3% and disposition effect by 37.8% [1], while decision error rates decreased from 0.38 to 0.17 [1]. However, financial literacy moderated these benefits, with high-literacy investors showing smaller performance differentials between AI-advised and self-directed conditions compared to low-literacy participants [1].

Transparency about AI behavior patterns similarly improved decision quality in classification tasks. Cabrera et al. found that showing behavior descriptions significantly improved overall accuracy in two of three tasks (fake reviews and bird classification) [25]. Participants learned to override AI errors and increased reliance on the AI when it was accurate [25]. Srivastava et al.'s trajectory planning study found that providing information about the decision environment—thereby establishing shared situation awareness—boosted both situation awareness and task performance while reducing overreliance on AI recommendations [15].

Conversely, feature-based explainability produced concerning performance decrements in high-stakes domains. Bauer et al.'s investment and real estate studies revealed that XAI explanations using LIME and SHAP decreased decision accuracy by 8.9% while participants observed explanations and 9.8% when not observing them [4]. The mechanism appeared to involve mental model adjustments subject to confirmation bias, allowing misconceptions to persist and accumulate [4]. Similarly, Schmidt et al. found that word highlighting and confidence scores decreased classification accuracy, particularly for correct AI predictions [17].

Chen et al. documented a particularly troubling pattern where revealing AI reasoning—whether brief or extensive—significantly increased agreement with AI recommendations (from 88.7% to 90.0-90.7% without unique human knowledge) [5]. However, this increased compliance crowded out utilization of unique human knowledge (UHK) that participants possessed but the AI lacked [5]. When participants had critical private information, brief reasoning increased agreement from 77.3% to 81.3% ($p<0.001$) and extensive reasoning to 81.7% ($p<0.001$) [5], demonstrating that transparency acted as a powerful persuasive heuristic inducing over-trust rather than appropriate calibration [5].

Explanation quality and content determined effectiveness in bias mitigation contexts. Hou et al. found that comprehensive explanations helped users recognize AI bias and mitigate its influence, with effects particularly pronounced

among females evaluating a gender-biased hiring AI [14]. This suggests transparency can support critical evaluation when explanations provide sufficient detail about problematic decision patterns.

Perverse Effects and Unintended Consequences

Thirteen studies (52%) documented perverse effects where transparency backfired or produced unintended negative consequences. Table 4 categorizes these effects.

Study	Perverse Effect Type	Mechanism	Specific Finding
Ancuta Margondai et al., 2026	Reduced autonomy	Detailed explanations reduce perceived control [9]	M=3.73 vs. 3.84, d=-0.19 [9]
Zenan Chen et al., 2025	Over-trust, UHK crowding	AI reasoning as persuasive heuristic [5]	90.7% agreement despite AI lacking critical information [5]
V. M. Ngo et al., 2025	Cognitive overload	Excessive transparency induces scrutiny [7]	Reduced AI adoption [7]
V. M. Ngo et al., 2025a	Cognitive overload, skepticism	Excessive transparency backfires [8]	Inverted U-shaped effect [8]
Xueming Luo et al., 2019	Discrimination against AI	Perceived lack of knowledge/empathy [2]	79.7% purchase reduction [2]
Pooja Prajod et al., 2026	Transparency-trust trade-off	Information overload from detailed disclosure [16]	Lower subscription rates despite preferring detail [16]
Philipp Schmidt et al., 2020	Mistrust of correct predictions	Unintuitive explanations [17]	Higher error rates following correct AI [17]
Sebastian Krügel et al., 2021	Overtrust, compliance pressure	Continued acceptance despite warnings [6]	No effect of transparency on advice acceptance [6]
Jingshu Li et al., 2024	Disuse, distrust	Calibration support triggers wariness [12]	Increased disuse despite reduced misuse [12]
Jaymari Chua et al., 2025	Strategic defeatism, overreliance	Disclosure of superhuman capability [18]	Novices frustrated; experienced players shift goals [18]
Inyoung Cheong et al., 2025	Discrimination, vanishing alignment	Genre expectations around objectivity [10]	LLM fairness preferences disappear with disclosure [10]
Aaron Springer et al., 2019	Distraction, heuristic violation	Transparency demands attention [23]	Expected benefits not realized in practice [23]
Roberta De Cicco et al., 2025	Speciesism-driven trust decline	Prejudice against non-human entities [3]	High speciesism amplifies negative disclosure effects [3]
Kevin Bauer et al., 2024	Algorithm aversion in creativity	Perceived reduced human effort [11]	Quality penalties for disclosed AI assistance [11]

Kevin Bauer et al., 2023	Confirmation bias, spillover	Mental model rigidity [4]	Biased decisions in unrelated domains [4]

The most common perverse effect was over-trust and reduced critical engagement following transparency. Chen et al. found that revealing AI reasoning increased agreement with recommendations to 90.7% [5], but this reflected over-trust that crowded out utilization of unique human knowledge participants possessed [5]. Similarly, Krügel et al. documented that users trustfully accepted AI ethical advice from both transparent and opaque algorithms, continuing to follow recommendations even when transparency revealed warnings against the algorithm [6]. This suggests transparency failed to promote critical evaluation and instead reinforced overtrust [6].

Cognitive overload emerged as a distinct mechanism limiting transparency benefits. Ngo et al. identified an inverted U-shaped relationship where excessive transparency induced cognitive burden and heightened scrutiny, reducing AI adoption [7, 8]. Prajod et al. documented a "transparency dilemma" in news journalism where detailed AI disclosures led to lower trust and subscription rates despite two-thirds of participants expressing preference for detailed disclosures [16]. This trade-off between readers' desire for transparency and their trust in AI-assisted content suggests information overload from detailed disclosures negatively impacts trust [16].

Transparency paradoxically reduced perceived autonomy in organizational decision-making. Margondai et al. found that detailed AI rationales significantly reduced perceived autonomy (M=3.73) compared to basic recommendations (M=3.84, d=-0.19) [9], with effects moderated by personality traits—Openness to Experience reversed the paradox (interaction=-0.227) while Extraversion amplified autonomy reduction (interaction=0.173) [9]. This represents a fundamental transparency paradox where providing more information reduced users' sense of decision control [9].

Trust calibration support produced the counterintuitive effect of increased AI disuse. Li et al. found that while transparency about uncalibrated confidence helped participants recognize errors and reduced misuse of overconfident AI, it simultaneously fostered distrust and caused greater disuse of the system overall [12]. Participants who received trust calibration support showed increased distrust in AI's predictions despite the support helping them better detect uncalibration [12].

Discrimination against AI and algorithm aversion intensified with transparency in several contexts. Luo et al.'s field experiment demonstrated that chatbot identity disclosure before conversation reduced purchases by 79.7% because customers perceived disclosed bots as less knowledgeable and less empathetic, despite objective competence equivalence [2]. Bauer et al. documented similar algorithm aversion in creative work, where GenAI disclosure led recipients to believe less human effort was invested in production [11], triggering quality and creativity penalties [11]. De Cicco et al. identified speciesism—prejudices against non-human entities—as a moderator, with high-speciesism individuals showing substantial trust decline when non-human identity was disclosed during interaction [3].

A particularly concerning finding involved "vanishing alignment" in algorithmic content evaluation. Cheong et al. found that LLM raters favored articles attributed to women or Black authors when no AI disclosure was present, but these fairness-oriented preferences completely disappeared when AI assistance was revealed [10]. This suggests AI transparency can paradoxically undermine fairness considerations in algorithmic decision systems [10].

Confirmation bias and mental model rigidity emerged as mechanisms limiting explainability benefits. Bauer et al. demonstrated that feature-based XAI methods evoked mental model adjustments subject to confirmation bias, allowing misconceptions to persist and accumulate [4]. Moreover, these adjustments created spillover effects altering user behavior in related but distinct domains where they lacked AI access [4], risking manipulation and promoting discriminatory biases [4].

Context-specific perverse effects arose in competitive and high-stakes domains. Chua et al. found that disclosing superhuman AI capabilities in competitive gaming provoked frustration and strategic defeatism among novices while triggering overreliance in competitive contexts [18]. Experienced players interpreted disclosure as confirmation of an unbeatable opponent and shifted to suboptimal goals [18]. Springer et al. documented that transparency can be distracting and undermine simple heuristics users form about system operation [23], with anticipated transparency benefits failing to materialize in actual use [23].

Synthesis: Reconciling Heterogeneous Findings

The substantial heterogeneity in transparency effects—ranging from 79.7% purchase reductions [2] to 1.78 standard deviation trust improvements [1] —cannot be dismissed as methodological noise. Instead, systematic analysis reveals four critical factors that predict when transparency improves versus worsens outcomes.

Transparency Content and Granularity

Studies finding transparency benefits predominantly tested limitation disclosure and behavior descriptions rather than identity disclosure or feature-based explanations. Toan Khang Trinh et al.'s financial advisory study, which achieved the largest positive effects (d=1.78) [1], provided transparency about system limitations, confidence indicators, and capability boundaries rather than merely revealing AI identity [1]. The 76.4% appropriate reliance rate in transparent conditions versus 34.2% in non-disclosure [1] suggests limitation-focused transparency supports trust calibration. Similarly, Cabrera et al.'s behavior descriptions—detailing how AI performs on subgroups with specific metrics and common patterns [25] —improved accuracy in two of three tasks [25] by helping users identify when to rely on versus override AI [25].

Conversely, studies documenting large negative effects primarily tested identity disclosure without accompanying capability information. Luo et al.'s 79.7% purchase reduction [2] followed simple chatbot identity disclosure [2] without capability communication. Mozafari et al. directly demonstrated this distinction: mere identity disclosure reduced trust (ΔTrust=-0.365, $p<0.1$) [19], but pairing disclosure with selective presentation of capabilities (either expertise or weaknesses) eliminated the negative effect (ΔTrust=0.378, $p<0.1$ for expertise) [19]. The critical difference lies not in whether transparency is provided, but in whether it communicates actionable information about when to rely on the AI.

Feature-based XAI explanations (LIME, SHAP) occupied an intermediate position, generally producing negative or null effects on trust and performance. Bauer et al. found 8.9-9.8% accuracy decreases [4] from LIME/SHAP explanations [4], while Schmidt et al.'s word highlighting decreased trust by 1-1.5 percentage points [17]. These granular feature attributions appear to trigger confirmation bias [4] and cognitive overload [17] without providing the high-level behavioral patterns needed for appropriate reliance decisions.

The inverted U-shaped relationship documented by Ngo et al. provides a unifying framework: moderate transparency (likely limitation disclosure) enhanced trust and certainty, but excessive transparency (potentially detailed feature explanations) induced cognitive overload and reduced adoption [7, 8]. This suggests an optimal transparency level exists between uninformative identity disclosure and overwhelming technical detail.

Domain Stakes and Decision Support Role

Transparency effects exhibited strong domain dependency aligned with decision stakes and AI's role. Financial advisory contexts, where AI served as a decision support tool for high-stakes portfolio choices, showed the most consistent positive effects. The 7.2% performance advantage [1] and dramatic bias reductions (41.3% overconfidence, 37.8% disposition effect) [1] occurred because transparency about system limitations helped investors calibrate when AI advice should influence high-consequence financial decisions. Srivastava et al.'s trajectory planning study—another

high-stakes domain where errors could doom space missions—similarly found transparency improved performance by reducing overreliance [15].

In contrast, low-stakes consumer interactions produced the most negative transparency effects. Luo et al.'s conversational commerce field experiment found 79.7% purchase reductions [2] because customers applied different evaluation criteria: they perceived disclosed chatbots as less knowledgeable and empathetic [2] despite objective performance equivalence with humans. De Cicco et al. documented similar patterns in fast-food delivery chatbots where disclosure reduced trust (M=4.81 vs. 5.26) [3] unless high social presence through design cues compensated [3]. The mechanism appears to involve anthropomorphic expectations: in low-stakes social contexts, users prioritize warmth and social connection over technical competence, making identity disclosure harmful without design accommodations.

This stakes-based pattern explains why content moderation contexts showed mixed effects. Molina et al. found that transparency about AI versus human classification increased trust when positive machine heuristics were triggered (viewing AI as objective and error-free) [20], while negative heuristics reduced trust. The heuristic triggered depended on how users construed the task—as technical classification requiring objectivity or as social judgment requiring human wisdom.

User Characteristics and Moderators

Individual differences substantially moderated transparency effects, explaining within-domain heterogeneity. Financial literacy emerged as a critical moderator in financial contexts: Toan Khang Trinh et al. found high-literacy investors showed smaller performance differentials between AI-advised and self-directed conditions [1], suggesting sophisticated users extract value from AI regardless of transparency while novices specifically benefit from limitation disclosure.

Personality traits moderated autonomy effects: Margondai et al. found that Openness to Experience reversed the transparency paradox (interaction=-0.227) [9], with intellectually curious individuals benefiting from detailed rationales that reduced others' perceived autonomy. Conversely, Extraversion amplified autonomy reduction (interaction=0.173) [9]. These patterns suggest transparency benefits depend on cognitive style—reflective, open individuals process detailed information constructively while others experience it as constraining.

Domain expertise reversed transparency effects in competitive contexts. Chua et al. found that novices experienced frustration and strategic defeatism when superhuman AI capabilities were disclosed [18], while experienced players interpreted disclosure as confirmation of an unbeatable opponent and adjusted strategies (albeit to suboptimal goals) [18]. This suggests expertise determines whether capability disclosure triggers appropriate strategic adaptation or demotivating resignation.

Speciesism—prejudice against non-human entities—moderated identity disclosure effects. De Cicco et al. demonstrated that high-speciesism individuals showed substantial trust decline when non-human identity was disclosed during interaction, while low-speciesism individuals were unaffected [3]. This individual difference in anthropomorphic bias explains why some users penalize AI identity while others do not.

Gender moderated both autonomy and fairness effects: males showed twice the autonomy reduction of females (0.137 vs. 0.063) [9], while females particularly benefited from transparency about AI bias in discriminatory contexts [14]. These patterns suggest transparency interacts with both personality and demographic characteristics in complex ways.

Timing, Format, and Design Implementation

Implementation details critically determined transparency outcomes even when content was held constant. Timing

effects appeared in chatbot disclosure: Luo et al. found that late disclosure timing partially mitigated the negative impact of identity revelation [2], while De Cicco et al.'s systematic comparison showed disclosure before interaction avoided the conversation interruption and trust decline observed when disclosure occurred during interaction [3].

Format variations explained divergent findings in similar domains. Prajod et al. documented that one-line AI disclosures in journalism did not significantly decrease trust compared to no disclosure, while detailed disclosures substantially reduced trust and subscription rates [16]. Critically, two-thirds of participants preferred detailed disclosures for transparency and accountability reasons [16], revealing a disconnect between information preferences and trust responses. The mechanism involved information overload from detailed disclosures triggering skepticism despite users consciously valuing the transparency [16].

Interactive transparency—allowing users to provide feedback or adjust AI recommendations—produced more consistently positive effects than passive transparency. Molina et al. found that interactive transparency increased trust and user agency by enhancing perceived disclosure and control [20], while Westphal et al. demonstrated that decision control (allowing users to adjust system recommendations) positively affected trust and compliance [22]. The active engagement appears to mitigate cognitive overload and autonomy reduction concerns.

Progressive disclosure offered a resolution to the transparency dilemma. Springer et al. demonstrated that while participants anticipated transparent systems would perform better, this expectation reversed after experience because transparency was distracting and undermined simple heuristics [23]. The authors proposed progressive disclosure—initially simplified feedback that hides potential errors and assists heuristic formation—as a solution [23]. This format provides transparency incrementally as users develop mental models, avoiding initial cognitive overload while supporting eventual sophisticated understanding.

Social presence design mitigated identity disclosure harms in consumer contexts. De Cicco et al. found that high social presence through design cues completely eliminated the negative trust effects of non-human identity disclosure [3], transforming the artificial agent into a social actor through subtle modifications. This suggests that design accommodations can compensate for anthropomorphic expectation violations.

Synthesis Implications

These four factors—content granularity, domain stakes, user characteristics, and implementation details—jointly determine transparency outcomes rather than operating independently. High-stakes decision support domains with sophisticated users benefit from moderate limitation-focused transparency implemented progressively or interactively. Low-stakes social contexts with diverse users require identity disclosure paired with social presence cues and capability information, preferably disclosed before rather than during interaction. Feature-based technical explanations should be reserved for users high in openness and cognitive ability in domains where understanding mechanisms matters more than knowing when to rely on AI.

The heterogeneity in findings thus reflects not measurement error or contradictory effects, but rather systematic variation in the appropriateness of specific transparency types for particular contexts and users. Universal transparency mandates risk either under-serving users who need detailed information or overwhelming users who would benefit from simpler disclosure. Adaptive transparency systems that tailor content, timing, and format to task requirements and user characteristics offer the most promising path forward.

References

1. Toan Khang Trinh, Guancong Jia, Caiqian Cheng, Chunhe Ni (2025) Behavioral Responses to AI Financial Advisors: Trust Dynamics and Decision Quality Among Retail Investors. Applied and Computational Engineering. https://doi.org/10.54254/2755-2721/2025.21859

2. Xueming Luo, Siliang Tong, Z. Fang, Z. Qu (2019) Machines versus Humans: The Impact of AI Chatbot Disclosure on Customer Purchases. Social Science Research Network. https://doi.org/10.2139/ssrn.3435635

3. Roberta De Cicco, Maher Georges Elmashhara, Susana C. Silva, Maik Hammerschmidt (2025) The impact of providing non-human identity cues about sales agents on consumer responses: the role of social presence and speciesism activation. European Journal of Marketing. https://doi.org/10.1108/ejm-01-2022-0066

4. Kevin Bauer, Moritz von Zahn, O. Hinz (2023) Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. Information systems research. https://doi.org/10.1287/isre.2023.1199

5. Zenan Chen, Ruijiang Gao, Yingzhi Liang (2025) Revealing AI Reasoning Increases Trust but Crowds Out Unique Human Knowledge. arXivorg. https://doi.org/10.48550/arXiv.2511.04050

6. Sebastian Krügel, Andreas Ostermaier, Matthias W. Uhl (2021) Zombies in the Loop? People are Insensitive to the Transparency of AI-Powered Moral Advisors. arXivorg

7. V. M. Ngo (2025) Balancing AI transparency: Trust, Certainty, and Adoption. Information Development. https://doi.org/10.1177/02666669251346124

8. V. M. Ngo (2025) The AI transparency dilemma: when more is less for trust and adoption. Information Discovery and Delivery. https://doi.org/10.1108/idd-03-2025-0056

9. Ancuta Margondai, Sara Willox, Anamaria Acevedo Diaz, et al (2026) The Transparency Paradox: How AI Explanations Reduce Perceived Autonomy in Organizational Decision-Making. AHFE International. https://doi.org/10.54941/ahfe1007091

10. Inyoung Cheong, Alicia Guo, Mina Lee, et al (2025) Penalizing Transparency? How AI Disclosure and Author Demographics Shape Human and AI Judgments About Writing. arXivorg. https://doi.org/10.48550/arXiv.2507.01418

11. Kevin Bauer, Ekaterina Jussupow, R. Heigl, et al (2024) All Just in Your Head? Unraveling the Side Effects of Generative AI Disclosure in Creative Task. Social Science Research Network. https://doi.org/10.2139/ssrn.4782554

12. Jingshu Li, Yitian Yang, Renwen Zhang, et al (2024) Understanding the Effects of Miscalibrated AI Confidence on User Trust, Reliance, and Decision Efficacy. ArXiv

13. Tobias Rieger, D. Manzey, Benigna Meussling, et al (2023) Be careful what you explain: Benefits and costs of explainable AI in a simulated medical task. Computers in Human Behavior. https://doi.org/10.1016/j.chbah.2023.100021

14. Hou Tsung-Yu, Tseng Yu-Chia, Tina Chien-Wen Yuan (2024) Is this AI sexist? The effects of a biased AI's anthropomorphic appearance and explainability on users' bias perceptions and trust. International Journal of Information Management. https://doi.org/10.1016/j.ijinfomgt.2024.102775

15. Divya K. Srivastava, J. Lilly, K. Feigh (2024) Exploring the role of judgement and shared situation awareness when working with AI recommender systems. Cognition, Technology & Work. https://doi.org/10.1007/s10111-024-00771-9

16. Pooja Prajod, Hannes Cools, Thomas Röggla, et al (2026) Full Disclosure, Less Trust? How the Level of Detail about AI Use in News Writing Affects Readers' Trust. arXivorg. https://doi.org/10.48550/arXiv.2601.09620

17. Philipp Schmidt, F. Biessmann, Timm Teubner (2020) Transparency and trust in artificial intelligence systems. Journal of Decision Systems. https://doi.org/10.1080/12460125.2020.1819094

18. Jaymari Chua, Chen Wang, Lina Yao (2025) Superhuman Game AI Disclosure: Expertise and Context Moderate Effects on Trust and Fairness

19. Nika Mozafari, W. H. Weiger, Maik Hammerschmidt (2021) Resolving the Chatbot Disclosure Dilemma: Leveraging Selective Self-Presentation to Mitigate the Negative Effect of Chatbot Disclosure. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.355

20. Maria D. Molina, S. Sundar (2022) When AI moderates online content: effects of human collaboration and interactive transparency on user trust. J Comput Mediat Commun. https://doi.org/10.1093/jcmc/zmac010

21. Guanglu Zhang, L. Chong, K. Kotovsky, J. Cagan (2022) Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. Computers in Human Behavior. https://doi.org/10.1016/j.chb.2022.107536

22. Monika Westphal, Michael Vössing, G. Satzger, et al (2023) Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. Computers in Human Behavior. https://doi.org/10.1016/j.chb.2023.107714

23. Aaron Springer, S. Whittaker (2019) Progressive disclosure: empirically motivated approaches to designing effective transparency. International Conference on Intelligent User Interfaces. https://doi.org/10.1145/3301275.3302322

24. Holly L. Wilson, A. Theodorou (2019) Slam the Brakes: Perceptions of Moral Decisions in Driving Dilemmas. AISafety@IJCAI

25. Ángel Alexander Cabrera, Adam Perer, Jason I. Hong (2023) Improving Human-AI Collaboration With Descriptions of AI Behavior. Proc ACM Hum Comput Interact. https://doi.org/10.1145/3579612