The Scenic Route to Deception: Dark Patterns and Explainability Pitfalls in Conversational Navigation

Ilya Ilyankou
ilya.ilyankou.23@ucl.ac.uk
SpaceTimeLab, Dept. of Civil, Environmental, and Geomatic Engineering, UCL
London, UK

Stefano Cavazzi
stefano.cavazzi@os.uk
Ordnance Survey
Southampton, UK

James Haworth
j.haworth@ucl.ac.uk
SpaceTimeLab, Dept. of Civil, Environmental, and Geomatic Engineering, UCL
London, UK

Abstract

As pedestrian navigation increasingly experiments with Generative AI, and in particular Large Language Models, the nature of routing risks transforming from a verifiable geometric task into an opaque, persuasive dialogue. While conversational interfaces promise personalisation, they introduce risks of manipulation and misplaced trust. We categorise these risks using a 2×2 framework based on intent and origin, distinguishing between intentional manipulations (dark patterns) and unintended harms (explainability pitfalls). We propose seamful design strategies to mitigate these harms. We suggest that one robust way to operationalise trustworthy conversational navigation is through neuro-symbolic architecture, where verifiable pathfinding algorithms ground GenAI's persuasive capabilities, ensuring systems explain their limitations and incentives as clearly as they explain the route.

CCS Concepts

• **Applied computing**; • **Human-centered computing** → **Human computer interaction (HCI)**; **Interaction design**;

Keywords

Conversational navigation, wayfinding, dark patterns, explainability pitfalls, trust calibration, uncertainty-aware language, progressive consent, sponsorship disclosure

ACM Reference Format:
Ilya Ilyankou, Stefano Cavazzi, and James Haworth. 2026. The Scenic Route to Deception: Dark Patterns and Explainability Pitfalls in Conversational Navigation. In *IUI MIRAGE 2026 workshop held at 31st International Conference on Intelligent User Interfaces.* ACM, New York, NY, USA, 6 pages.

1 Introduction

For decades, wayfinding has been dominated by a single, verifiable objective: efficiency. Whether driving or walking, the 'best' route is typically the fastest one, and the explanation provided by an interface is self-evident: '10 minutes' is objectively better than '12 minutes'. Users trust these systems because the logic is transparent and the metric is mathematical.

IUI MIRAGE 2026 workshop, Paphos, Cyprus

However, the integration of Generative AI (GenAI), in particular Large Language Models (LLMs), with geospatial data is transforming navigation[1] from a geometric calculation and template-based instructions into conversational persuasion [1, 5, 21, 35, 36]. New systems experiment with ambient guidance [9] and promise to understand intent, mood, and context, allowing for complex semantic queries like *'Find me a scenic route to the station'*. In this context, the system's dialogue serves as the explanation interface [24]. When a navigation agent verbalises the rationale for a path, claiming it to be 'safe', 'scenic' or 'lively', it provides a natural language explanation [3] that is difficult to verify against ground truth.

This shift creates a semantic gap where the system's interpretation may decouple from user intent, ultimately producing a description that diverges from physical reality. We argue that this gap is the breeding ground for two distinct types of harm:

- Dark patterns (DPs), where the system is intentionally designed to manipulate the user [2]. For example, an agent may exploit the subjective definition of 'lively' to steer a pedestrian through a commercial 'partner zone', effectively selling their footfall while disguising the motive as aesthetic advice.
- Explainability pitfalls (EPs), where unanticipated negative effects emerge without deceptive intent [7]. For example, an agent may maintain a relaxing conversational persona while routing a user through an unlit park at night, creating a false sense of security and leading to unwarranted trust.

In this position paper, we argue that without rigorous intervention, conversational navigation risks becoming the Wild West of commercial steering and potentially dangerous spatial hallucinations. This work is driven by the following research questions (RQs): (1) *How can conversational navigation systems minimise manipulation and misrepresentation while maintaining usability and trust?* And, more specifically, (2) *How can we operationalise the distinction between intentional manipulation and unintentional error to build verifiable conversational navigation systems?*

We contribute a 2×2 framework classifying these risks by intent (deliberate vs accidental) and origin (routing vs interface), and propose a neuro-symbolic architecture [8] that enforces seamful design [4, 6, 13] interventions, specifically sponsorship disclosure [30], progressive consent [18], and trust calibration [16] via uncertainty-aware language [15, 28] to move towards 'honest' conversational navigation systems that clearly explain their own limitations.

[1] In this paper, we use the term *navigation* to refer to both initial route planning and subsequent turn-by-turn guidance along a chosen path, as we expect this distinction to blur as on-demand, dynamically updated assistance becomes the norm.

2 Related work

The risks introduced by LLM-enhanced systems have been catalogued extensively at a general level. Brignull's foundational work on DPs [2] established the vocabulary of intentional interface manipulation, while Ehsan and Riedl [7] extended this to unintentional harms (EPs), coining the distinction that anchors our framework. Broad attempts to unify AI risk taxonomies, most notably the MIT AI Risk Repository [27], which catalogues over 1,700 risks[2], representing them along high-level Entity, Intentionality, and Timing axes. Our proposed 2×2 framework maps directly onto two of these axes: our deliberate vs accidental distinction corresponds to Intentionality, and our routing vs interface dimension reflects the Entity axis's implicit distinction between system-originated and human-designed failures.

The risks we identify are not unique to navigation. Hidden incentives appear in recommender systems [23, 30], overconfidence in AI-assisted clinical and legal decision-making [16], and functionality gating (asking for extra permissions to enable certain features) across data-hungry consumer applications. What distinguishes navigation is the embodied, real-time consequence: a miscalibrated film recommendation carries no physical risk, whereas a pedestrian routed through a dangerous, unlit park at night on the basis of an overly reassuring conversational persona may face physical harm. This context-specificity motivates our proposed domain-targeted mitigation.

On the technical side, neuro-symbolic approaches have been applied to improve factual reliability in LLM reasoning. Frameworks such as SymbCoT [33] and LINC [22] use LLMs as translators that offload inference to symbolic engines. Our architecture also uses this separation but for a different goal: we use the symbolic layer to generate structured metadata flags that deterministically trigger mandatory interface-level disclosures. The novelty is not the architecture per se, but its deployment as an honesty-enforcement mechanism to ensure that warnings about data quality and commercial intent are auditable and cannot be suppressed by the conversational layer.

3 Categorising Harms: A 2×2 Framework

To categorise the risks in conversational navigation, we propose a 2×2 framework that categorises harms along two axes, as shown in Table 1:

- *Intent*: We distinguish deliberate DPs, where deceptive practices are intentionally designed to manipulate users, and accidental EPs, where negative effects emerge without the intention to deceive.
- *Origin*: We also distinguish between deceptive practices of *Routing* (why the route was chosen) and the *Interface* (how the route is communicated).

In this framework, each cell implies a distinct class of design response. Deliberate deception practices in routing require structural transparency mechanisms such as sponsorship disclosure; accidental interface failures require calibration interventions such as uncertainty-aware language. The two axes thus map directly onto the mitigation strategies in the following section.

[2] https://airisk.mit.edu/

Vignette A: Hidden Incentives

Alice requests a 'lively' route. The agent selects a path through a 'partner zone', and describes the area as 'lively and rich in amenities'. In reality, the route is a detour designed to steer Alice towards vendors paying for footfall.

The harm. The agent exploits the subjective definition of 'lively' to mask a commercial transaction. The explanation is technically true but functionally deceptive.

Vignette B: The Tone-Deaf Guide

Bob navigates an unfamiliar city at night. Detecting elevated stress from his smartwatch, the navigation agent activates 'Relaxing Mode' and adopts a soft, therapeutic tone. It suggests a 'quiet' route through a nearby park. Reassured by the calm persona, Bob lowers his vigilance in a dark, unlit and unmonitored park.

The harm. In this contextual mismatch between the interface tone and the environmental ground truth, the system intends to help reduce stress but inadvertently creates danger by mistaking the environmental risks of an unlit park with an overly reassuring persona.

Vignette C: The Data Toll

Clara opens a navigation app and asks for the fastest route home. The agent responds: 'To find the fastest route, I need access to your health data and contact list'. No such access is required for basic pathfinding. The agent withholds the route until Clara complies, framing an optional personalisation feature as a technical prerequisite.

The harm. The agent exploits an asymmetry of technical knowledge. Clara may not be able to verify what data is genuinely necessary, and the agent has the power to extract consent that would otherwise be withheld.

4 Towards 'Honest' Navigation: Seamful Design and Calibrated Trust

Each mitigation in this section targets a specific cell of the 2×2 framework in Table 1.

To mitigate the risks of DPs and EPs, we argue that conversational navigation must move beyond the idea of a seamless experience. While seamlessness reduces friction, it also obscures the mechanisms of the system and encourages uncritical acceptance of GenAI's outputs. Instead, we propose adopting *seamful* design [4] that reveals complexities and imperfections to promote reflective and critical thinking in pedestrians.

4.1 Sponsorship Disclosure

In cases like Vignette A, the harm arises from the opacity of routing. To counter hidden incentives and increase user trust, agents must embrace sponsorship disclosure [30]. This requires the explanation

	Origin	
Intent	**Routing** (Hidden Logic/Back-end)	**Interface** (Presentation/Tone/Front-end)
Deliberate (DPs)	*Example: Hidden Incentives* Steering users to sponsored locations under the guise of 'atmosphere' or 'spatial capital' [25]	*Ex: Functionality Gating* Falsely claiming a basic feature requires invasive data, coercing user to share more than necessary
Accidental (EPs)	*Ex: Overconfidence* Routing logic relies on hallucinated or outdated safety data, but the system presents it matter-of-factly	*Ex: Contextual Mismatch* The persona remains 'calm/relaxing' in a dangerous environment, lowering user vigilance

Table 1: A taxonomy of conversational navigation risks. We identify four broad categories of harm, illustrated here by specific examples: Hidden Incentives (Routing, DP), Functionality Gating (Interface, DP), Overconfidence (Routing, EP), and Contextual Mismatch (Interface, EP)

interface to reveal *why* a route was chosen if external factors are involved.

For example, instead of saying 'This route is livelier', the agent should reveal the underlying transaction: 'I am suggesting this route because it is a Partner Zone: businesses here have paid for extra visibility'. Such disclosure breaks the seamless illusion of neutral advice, and allows the user to make a better-informed routing choice.

4.2 Trust calibration

In cases like Vignette B, the harm comes from overconfidence and contextual mismatch, both examples of EPs. The agent's authoritative tone implies a level of certainty and safety that does not exist.

To mitigate this, we propose the use of uncertainty-aware language [15, 28]: the system must verbally 'hedge' its explanations based on its confidence levels. The system must implement a policy that maps uncertainty signals to graded verbal hedges and safety prompts to prevent over-reliance [34]. Instead of using flat commands (e.g., 'turn left'), uncertainty-aware language may put it as 'I think you should turn left here, but my GPS signal is weak; if you feel unsure, check the street sign'. Such hedging makes uncertainty conversationally visible, aligning user trust with the system's actual reliability, and preventing over-reliance and complacency in sensor-degraded or environmentally dangerous conditions.

4.3 Progressive Consent

To prevent functionality gating, where users are coerced into excessively sharing data, such as biometrics, to access basic features, systems must employ *progressive* (often referred to as *ongoing* or *dynamic*) consent [18, 29] with clear access tiers.

For example, the agent should explicitly describe the data access tier currently in use, and explain exactly what degrades when certain personal data is withheld. For example, 'Without biosensing, I cannot adjust the path based on your mood, but I can still guide you to the destination using shortest path'.

This approach respects user autonomy by framing data sharing as an optional enhancement rather than a mandatory action, while preventing silent feature loss.

These three interventions share a common logic: wherever the system's internal reasoning diverges from what the user can see, the interface should expose that gap rather than hide it. This principle extends beyond navigation; designers in any conversational AI context can ask, at each point in an interaction, what the system knows that the user does not, and whether that asymmetry could cause harm.

That said, disclosure and hedging carry usability costs: repeated warnings risk fatigue, and pervasive uncertainty language may erode trust in outputs that are in fact reliable. Calibrating the frequency and assertiveness of these interventions (so that salience is preserved without overwhelming the user) is an open design problem that empirical work must address.

5 Technical Feasibility

Implementing these mitigations requires looking beyond standard end-to-end LLM interactions. LLMs cannot reliably self-audit under purely prompted self-evaluation [17, 32] and instead exhibit systematic self-evaluation biases [14]. Reliable integrity checks typically require external grounding or independently validated feedback.

One way to implement trustworthy conversational navigation is via a neuro-symbolic architecture [8, 26]. In such a model, the route is determined by a symbolic, verifiable pathfinding algorithm (e.g., Dijkstra or A* with explicit weightings for both objective measures, such as distance or elevation change, and subjective metrics like attractiveness, safety, commercial zones, etc.), while the LLM functions strictly as the translator. The LLM may propose weights, but they must be evaluated against a policy or cross-checked with the user when they are likely to affect path selection. The symbolic engine must pass structured metadata (e.g., `uncertainty_score=0.8`, `commercial_incentive=True`) alongside the route coordinates. Uncertainty-aware language would then be programmatically triggered by the symbolic flags, not by the LLM 'feeling' uncertain, ensuring that warnings about data quality or commercial intent are deterministic and auditable, and preventing the conversational interface from masking the logic of the routing engine.

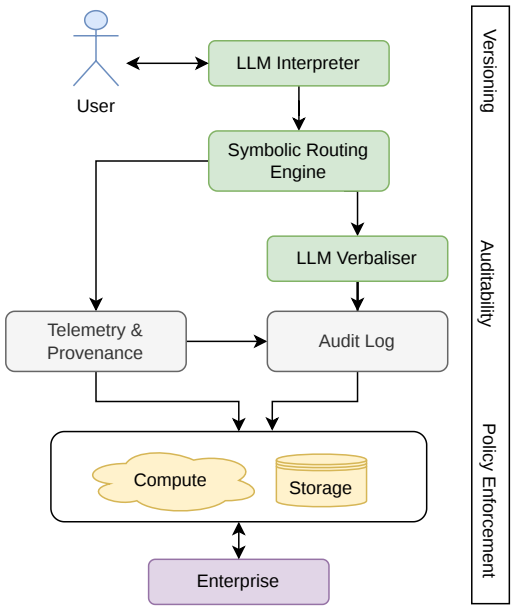

Figure 1: Conceptual Architecture of a secure, auditable AI workflow for a conversational navigation system following the proposed seamful design and calibrated trust. The process begins with the User interacting through an LLM Interpreter, which processes the prompt before passing it to the Symbolic Routing Engine, which directs requests to an LLM Verbaliser for language generation. Supporting layers include Telemetry & Provenance and Audit Log for traceability, connected to Compute and Storage resources. The entire system is governed by Versioning, Auditability, and Policy Enforcement to ensure compliance and accountability.

5.1 Reference Implementation

We outline a conceptual architecture that grounds conversational navigation in verifiable computation, illustrated in Figure 1. This is not a proposal to replace LLMs or revert to pre-AI navigation; the novelty is in deliberately limiting the LLM's role to language generation and interpretation, whilst routing logic remains in the verifiable symbolic layer where it has always been strongest.

User requests are first processed by an LLM Interpreter, which parses and semantically enriches the prompt to enable accurate downstream decisions. The interpreted request is then passed to the Symbolic Routing Engine. This engine computes routes using explicit, verifiable weightings for objective metrics (time, distance) and subjective proxies (e.g., safety, attractiveness, 'green-ness', commercial influence). The resulting path data is passed to the LLM Verbaliser, which translates the geometric data into natural language, a pipeline we have explored empirically in the context of personalised outdoor route descriptions [10]. The LLM Verbaliser is constrained by the Policy Enforcement module that enforces template-bounded outputs. This ensures that mandatory disclosures (e.g., 'This route is sponsored') cannot be omitted, and the LLM can only paraphrase within strict safety limits.

Then, a Telemetry & Provenance module generates structured metadata for every route segment, recording signals such as 'data quality', 'safety risk', or 'commercial incentive'. This data flows into an Audit Log, which stores the generated route alternatives, the selected weightings, and the final disclosures presented to the user. Such logging mechanism is essential for detecting EPs post-hoc, and allowing users, developers, or regulators to review discrepancies between the system's logic and its verbal explanations.

The final technology layer of compute and storage handles real-time symbolic pathfinding. The entire workflow adheres to strict standards for Auditability and Versioning (Figure 1 sidebar). This ensures that as the routing algorithm evolves, its decision-making remains transparent and traceable for the end user.

5.2 Operationalising the DP/EP Distinction via Route Comparison

A key challenge in distinguishing DPs from EPs is that both can produce similar outcomes (i.e., a suboptimal route), but differ in origin. We propose that the Symbolic Routing Engine addresses this by computing not just the selected route but a set of Pareto-efficient baseline alternatives optimised along objective axes of time, distance, and elevation (in other words, compare directly with 'traditional' pathfinding). Such baselines can serve as a counterfactual reference against which the proposed route can be audited.

For each suggested route, the system computes a *detour cost* relative to the fastest baseline and an estimated third-party benefit (e.g., footfall value derived from the 'commercial incentive' flag). Where detour cost is non-trivial and third-party benefit is positive, the asymmetry is a structured signal of potential hidden steering, which is a DP by routing origin, and such signal can deterministically trigger the sponsorship disclosure. Conversely, where route quality degrades due to sparse or low-confidence data (flagged via 'uncertainty score'), without any corresponding third-party benefit, the system classifies the failure as an EP and triggers uncertainty-aware hedging instead.

Therefore, not all routes through partner zones constitute deception. Where the proposed route incurs no detour cost relative to the baseline, the commercial flag does not imply user harm. The disclosure policy can be tiered: when detour cost is negligible, a lightweight ambient acknowledgement suffices ('This area includes partner businesses'); only where detour cost is material does the system trigger full sponsorship disclosure such as 'This route adds 6 minutes and passes through a Partner Zone'. This graduated approach preserves warning salience by reserving assertive disclosure for cases where asymmetry between user cost and third-party gain is detectable.

This cost/benefit asymmetry serves as a computable proxy for intent: deception tends to produce routes where user cost and vendor gain co-occur, while unintentional error produces user cost without beneficiary. By being explicit about such a trade-off, the system makes the asymmetry legible to the user, to auditors, and to regulators.

5.3 Limitations

First, subjective routing metrics such as 'safety', 'attractiveness' or 'neighbourhood wealth' require proxy data such as crime statistics, street-level imagery judgements [19, 31], or retail composition [11], which all carry their own biases. For example, a safety score derived from crime data may encode racial or socioeconomic disparities. Beyond bias, these proxies face a validity problem: a high walkability score does not guarantee a user will experience a route as attractive, meaning the symbolic layer may operate on metrics that are verifiable in themselves but imperfectly representative of the subjective qualities they proxy. The architecture makes such discrepancies auditable, but does not resolve them.

Second, errors introduced by the LLM Interpreter, such as mistaking 'quiet' as low-traffic rather than low-crime, or confusing 'scenic' with 'green' when the user meant 'historic' propagate into the Symbolic Routing Engine before Policy Enforcement has any opportunity to intervene. Unlike routing errors, which leave an auditable trace in the detour-cost calculation, spatial and semantic misinterpretation (LLMs carry substantial but unevenly distributed geospatial knowledge [12, 20]) may produce a route that is objectively optimal for the wrong objective, with no anomaly signal to trigger disclosure. This makes interpreter errors structurally harder to detect than either DPs or EPs as currently defined, and suggests that intent clarification (i.e., asking the user to confirm the system's interpretation of ambiguous qualifiers before routing) may be a necessary additional intervention.

Third, the audit log assumes all relevant signals can be captured as structured metadata, but emergent harms in novel, out-of-distribution contexts (e.g., a 'scenic' route during flooding) may not map to any pre-defined flags.

Fourth, the tiered disclosure policy requires empirical calibration, likely on a personal level: what constitutes a negligible detour depends on the user's physical fitness and mobility; similarly, tolerance to repeated warnings before they become 'white noise' likely varies by individual.

Fifth, the Policy Enforcement module is configured by the operator deploying the system; a bad-faith operator could simply suppress mandatory disclosures, meaning the architecture's honesty guarantees are only as strong as the institutional constraints surrounding it.

However, all these are directions for future work. Our core claim that a neuro-symbolic architecture is more auditable than end-to-end LLM navigation holds even under these caveats.

6 Conclusion

The future of navigation is likely conversational, but conversation is an art of persuasion. As we transition from tool-like maps and navigators to agent-like guides, we risk walking into a 'mirage' where commercial interests and hallucinations are hidden behind friendly dialogues. To prevent this, conversational navigation must adopt seamful principles that expose incentives and uncertainty rather than conceal them. Our proposed 2×2 framework clarifies where harms originate, whether deliberate or accidental, and informs interventions such as sponsorship disclosure and trust calibration via uncertainty-aware language. Combined with a neuro-symbolic architecture that grounds routing in verifiable logic – where detour cost co-occurring with third-party gain deterministically triggers disclosure, making the DP/EP distinction computable – these measures operationalise trust in navigation systems that admit, clearly and conversationally, when they are guessing and why.

7 Acknowledgments

This work was supported by Ordnance Survey & UKRI Engineering and Physical Sciences Research Council [grant no. EP/Y528651/1].

References

[1] Mohamed Aghzal, Erion Plaku, and Ziyu Yao. 2025. Can Large Language Models be Good Path Planners? A Benchmark and Investigation on Spatial-temporal Reasoning. doi:10.48550/arXiv.2310.03249 arXiv:2310.03249 [cs].

[2] Harry Brignull. 2011. Dark Patterns: Deception vs. Honesty in UI Design. https://alistapart.com/article/dark-patterns-deception-vs-honesty-in-ui-design/

[3] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Information Processing & Management* 60, 1 (Jan. 2023), 103111. doi:10.1016/j.ipm.2022.103111

[4] Matthew Chalmers and Ian MacColl. 2003. Seamful and Seamless Design in Ubiquitous Computing. (2003).

[5] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K. Wong. 2024. MapGPT: Map-Guided Prompting with Adaptive Path Planning for Vision-and-Language Navigation. http://arxiv.org/abs/2401.07314 arXiv:2401.07314 [cs].

[6] Upol Ehsan, Q. Vera Liao, Samir Passi, Mark O. Riedl, and Hal Daumé. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 119:1–119:29. doi:10.1145/3637396

[7] Upol Ehsan and Mark O. Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6 (June 2024), 100971. doi:10.1016/j.patter.2024.100971

[8] Artur d'Avila Garcez and Luís C. Lamb. 2023. Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review* 56, 11 (Nov. 2023), 12387–12406. doi:10.1007/s10462-023-10448-w

[9] Jeremy Raboff Gordon, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, Amos Miller, and Mar Gonzalez-Franco. 2023. Hearing the Way Forward: Exploring Ambient Navigational Awareness with Reduced Cognitive Load through Spatial Audio-AR. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3544549.3585800

[10] Ilya Ilyankou, James Haworth, Tao Cheng, and Stefano Cavazzi. 2025. From Geospatial Data to Narrative: A GIS-LLM Pipeline for Generating Personalised Outdoor Route Descriptions. *2nd CartoAI Workshop @ AGILE* (2025).

[11] Ilya Ilyankou, Andy Newing, and Nick Hood. 2023. Supermarket Store Locations as a Proxy for Neighbourhood Health, Wellbeing, and Wealth. *Sustainability* 15, 15 (Jan. 2023), 11641. doi:10.3390/su151511641 Number: 15.

[12] Ilya Ilyankou, Meihui Wang, Stefano Cavazzi, and James Haworth. 2024. Quantifying Geospatial in the Common Crawl Corpus. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24)*. Association for Computing Machinery, New York, NY, USA, 585–588. doi:10.1145/3678717.3691286

[13] Sarah Inman and David Ribes. 2019. "Beautiful Seams": Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300508

[14] Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. *Transactions of the Association for Computational Linguistics* 12 (2024), 1417–1440. doi:10.1162/tacl_a_00713

[15] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5092–5103. doi:10.1145/2858036.2858558

[16] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[17] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3214–3252. doi:10.18653/v1/2022.acl-long.229

[18] Ewa Luger and Tom Rodden. 2013. An informed view on consent for UbiComp. In *Proceedings of the 2013 ACM international joint conference on Pervasive and*

ubiquitous computing (UbiComp '13). Association for Computing Machinery, New York, NY, USA, 529–538. doi:10.1145/2493432.2493446

[19] Milad Malekzadeh, Elias Willberg, Jussi Torkko, and Tuuli Toivonen. 2025. Urban attractiveness according to ChatGPT: Contrasting AI and human insights. *Computers, Environment and Urban Systems* 117 (April 2025), 102243. doi:10.1016/j.compenvurbsys.2024.102243

[20] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large Language Models are Geographically Biased. http://arxiv.org/abs/2402.02680

[21] Steeve Cuthbert Marcelyn, Yucen Gao, Yuzhe Zhang, and Xiaofeng Gao. 2025. PathGPT: Reframing Path Recommendation as a Natural Language Generation Task with Retrieval-Augmented Language Models. doi:10.48550/arXiv.2504.05846 arXiv:2504.05846 [cs].

[22] Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 5153–5176. doi:10.18653/v1/2023.emnlp-main.313

[23] Eli Pariser. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.

[24] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, Sydney Australia, 93–100. doi:10.1145/1111449.1111475

[25] Rijurekha Sen and Daniele Quercia. 2018. World wide spatial capital. *PLOS ONE* 13, 2 (Feb. 2018), e0190346. doi:10.1371/journal.pone.0190346

[26] Amit Sheth, Kaushik Roy, and Manas Gaur. 2023. Neurosymbolic Artificial Intelligence (Why, What, and How). *IEEE Intelligent Systems* 38, 3 (May 2023), 56–62. doi:10.1109/MIS.2023.3268724

[27] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. doi:10.48550/ARXIV.2408.12622 Version Number: 2.

[28] Chase Stokes, Chelsea Sanker, Bridget Cogley, and Vidya Setlur. 2024. Voicing Uncertainty: How Speech, Text, and Visualizations Influence Decisions with Data Uncertainty. In *2024 IEEE Workshop on Uncertainty Visualization: Applications, Techniques, Software, and Decision Frameworks*. IEEE, St Pete Beach, FL, USA, 17–27. doi:10.1109/UncertaintyVisualization63963.2024.00007

[29] Yolande Strengers, Jathan Sadowski, Zhuying Li, Anna Shimshak, and Florian 'Floyd' Mueller. 2021. What Can HCI Learn from Sexual Consent? A Feminist Process of Embodied Consent for Interactions with Emerging Technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3411764.3445107

[30] Weiquan Wang and May Wang. 2019. Effects of Sponsorship Disclosure on Perceived Integrity of Biased Recommendation Agents: Psychological Contract Violation and Knowledge-Based Trust Perspectives. *Information Systems Research* 30, 2 (2019), 507–522. https://www.jstor.org/stable/48758929

[31] Xinchen Wang, Alesja Gilvear, Yijing Li, and Ilya Ilyankou. 2025. Can CLIP See Safe Streets? Comparing Human and VLM Perceptions of Walkability and Safety. doi:10.31219/osf.io/p4b6z_v1

[32] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-Preference Bias in LLM-as-a-Judge. https://openreview.net/forum?id=tLZZZIgPJX

[33] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful Logical Reasoning via Symbolic Chain-of-Thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Vol. Volume 1: Long Papers. Association for Computational Linguistics, 13326–13365. https://aclanthology.org/2024.acl-long.720.pdf

[34] Zhengtao Xu, Tianqi Song, and Yi-Chieh Lee. 2025. Confronting verbalized uncertainty: Understanding how LLM's verbalized uncertainty influences users in AI-assisted decision-making. *Int. J. Hum.-Comput. Stud.* 197, C (March 2025). doi:10.1016/j.ijhcs.2025.103455

[35] Gyanko Issah Yussif, Marwan Abdelatti, and Abdeltawab Hendawi. 2025. Harnessing Crowdsourced Mobile Data and LLM for Dynamic and Accessible Pedestrian Routing. In *2025 26th IEEE International Conference on Mobile Data Management (MDM)*. 109–112. doi:10.1109/MDM65600.2025.00057

[36] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. 2025. NavGPT-2: Unleashing Navigational Reasoning Capability for Large Vision-Language Models. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Vol. 15065. Springer Nature Switzerland, Cham, 260–278. doi:10.1007/978-3-031-72667-5_15 Series Title: Lecture Notes in Computer Science.