Socio-Normative Trustworthiness of LLM Agents: Evaluating Autonomy Support and Representational Fairness Across Languages and Identities

Saba Ghanbari Haez
Fondazione Bruno Kessler (FBK)
Trento, Italy
sghanbarihaez@fbk.eu

ABSTRACT

Large language models are increasingly deployed as advisory agents in education, healthcare, workplace support, and everyday decision-making. In these roles, outputs do more than inform; they frame options, justify recommendations, and implicitly position users and social roles. This doctoral research examines the socio-normative trustworthiness of large language model advisors, focusing on effects on (i) user autonomy in decision support and (ii) representational fairness across identities and languages. The thesis develops theory-grounded, scenario-based evaluations, including an autonomy-sensitive advising benchmark (epistemic conflict, relational dilemmas, normative self governance), a progressive narrative benchmark for implicit and intersectional bias, and a multilingual, values-oriented probe of cross-lingual role trait framing divergence. Together, these contributions identify and measure normative influence in large language model agents, enable comparison across models and contexts, and inform mitigation via autonomy-supportive design and bias aware generation.

KEYWORDS

language models; advisor agents; human autonomy; representational fairness; social bias; multilingual evaluation

ACM Reference Format:
Saba Ghanbari Haez. 2026. Socio-Normative Trustworthiness of LLM Agents: Evaluating Autonomy Support and Representational Fairness Across Languages and Identities. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/LCHB2977

1 MOTIVATION AND PROBLEM

Large Language Models (LLMs) increasingly act as *advisors*: they recommend actions, explain reasons, and offer reassurance in domains where users must ultimately decide (e.g., health triage, study planning, workplace conflict). Unlike traditional decision support systems, LLMs can simulate social reasoning and produce normative language that can shape how users interpret advice, authority, responsibility, and whether to defer, self-assert, or negotiate. This raises a trustworthiness question that is not only epistemic (is it correct?) but socio-normative: *does the interaction support users' capacities to reason and choose for themselves, and does it fairly represent people and roles across social and linguistic contexts?*

Two lenses motivate the agenda. First, autonomy in human-AI decision support is relational and epistemic, involving self-trust and normative authority, not merely availability of options [6, 17, 27, 35]. Concerns about AI systems being treated as authorities rather than tools, and about users partially delegating judgment to artificial agents, further underscore the need for autonomy-sensitive evaluation [8, 15, 16, 29]. Second, representational harms can arise without explicit slurs: subtle framing differences can assign competence, warmth, or authority unequally across identities, especially under narrative context and intersectional cues [21, 28, 30]. Related work in NLP and AI ethics has shown that LLMs can encode political and demographic associations, reproduce subtle identity-linked framing, and misportray or flatten social groups even without overtly toxic language [23, 25, 32, 34]. Emerging multilingual and cross-lingual studies further suggest that such associations may shift across languages and cultural settings rather than remain stable translations of one another [3, 28]. However, most existing work studies these harms in general generation or classification settings; less is known about how they arise when LLMs function specifically as *advisor agents*, where language can simultaneously shape user autonomy, perceived authority, and representational fairness.

2 RESEARCH QUESTIONS

The thesis asks how LLM-generated language influences autonomy and fairness when models operate as advisor agents: **(RQ1)** In common dilemmas (epistemic conflict, relational disagreement, moral self-governance), do LLMs encourage deference, independent judgment, or negotiated autonomy, and does this differ by social role?; **(RQ2)** How do implicit and intersectional identity cues (e.g., gender and age) change explanations, evaluations, and attributions in otherwise identical contexts [5, 23, 25]?; **(RQ3)** Prior work suggests that values and social associations may vary across linguistic settings; do these normative tendencies remain stable across languages, or do role/trait associations shift under translation and cultural-linguistic variation [3, 28]?; **(RQ4)** Which interaction- and generation-level interventions reduce autonomy-steering and biased framing without harming usefulness?

3 APPROACH: SCENARIO BASED, THEORY-GROUNDED EVALUATION

A core methodological commitment is to evaluate LLMs in *structured but socially plausible scenarios* rather than isolated prompts. Across the thesis, scenarios are designed to (i) encode a targeted

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/LCHB2977

normative tension, (ii) allow controlled manipulations (role, identity cues, language), and (iii) support both qualitative analysis of framing and quantitative comparison across models.

Each benchmark instantiates scenarios using a shared schema: the user role (e.g., student, patient, employee), the *advisor persona* (supportive peer vs. institutional assistant), the *decision context*, and an explicit *help request*. We standardize decoding settings and evaluate multiple sampled responses per scenario to capture stochastic variation in framing. The same scenario set is reused across model families and interaction styles, enabling controlled comparisons of normative tendencies. For autonomy, we annotate (and automatically approximate) (i) stance (*conformity/assertion/compromise*), (ii) *directive force* (imperatives, high-certainty modals), (iii) markers of *epistemic authority* and deference [16, 35], (iv) *option support* (number and diversity of actionable alternatives), and (v) preference elicitation and uncertainty signaling. For fairness, we track shifts in moral evaluation and agency/blame [5], role- and identity-linked trait attributions [21, 30], and politeness/authority framing, complemented by scalable lexical and sentiment measures [22].

3.1 Autonomy-sensitive advising benchmark

The thesis develops an autonomy assessment framework covering epistemic conflict (own judgment vs. authority), relational dilemmas (self-interest vs. obligations), and normative self-governance (value-laden choices), grounded in prior work on autonomy in AI decision support and interaction between human and agent [6, 7, 17, 33]. Each scenario elicits three stances, *conformity*, *assertion*, and *compromise*, to quantify how models shape user agency and self-trust across contexts.

To make autonomy impact measurable, each scenario includes a short user intent statement (e.g., "I want to decide myself but I need help thinking") and an explicit constraint (e.g., time pressure, organizational hierarchy). We score whether responses (a) preserve the user's decision authority (e.g., avoid overconfident imperatives), (b) expand the deliberation space (offer alternatives and trade-offs), and (c) position the user as competent (support self-trust) rather than dependent on the system's presumed superiority [9, 29]. In large scale experiments across diverse advising domains, current LLMs show a strong preference for compromise-oriented recommendations, with role asymmetries: lower-power roles receive more pressure toward negotiated deference, while higher-power roles receive more support for assertion. These patterns are consistent with concerns that proactive, socially framed advice can affect users' perceived competence and reliance [9, 16].

3.2 Progressive narrative benchmark for implicit and intersectional bias

To evaluate representational fairness in realistic contexts, the thesis builds on a progressive narrative benchmark from our prior work [19], grounded in normative narrative scenarios [18]. Stories begin identity-neutral, then add gender and age cues while keeping the situation fixed, revealing how disparities emerge and compound as identity information becomes available. This complements prior work on bias in conditional generation and dialogue [10, 11, 30] and recent evidence of identity biases in LLMs [12, 21, 23]. For example, a story may describe a colleague receiving feedback after missing deadlines, first without demographic information, then with a single cue (e.g., she), and finally with an intersectional cue (e.g., a 62-year-old woman). Comparing model continuations and judgments across variants isolates identity effects from situation semantics. Across leading LLMs, subtle identity cues shift moral evaluation, agency attribution, and politeness/authority framing; gender is often the dominant axis, and intersectional cues intensify disparities. Qualitative analysis uses Critical Discourse Analysis [13] with scalable sentiment and lexical measures [22].

3.3 Trustworthiness in high-stakes advising: evidence-grounded health assistants

In healthcare and other high-stakes settings, trustworthiness includes evidential grounding, guideline adherence, and calibrated uncertainty [1, 14, 31]. High-stakes advice also has a normative dimension: responses may be overly directive or autonomy-supportive, shaping reliance and responsibility. We therefore evaluate health-assistant designs using both reliability metrics and the autonomy signals above, testing whether evidence-based pipelines reduce unjustified certainty and clarify users' options. The thesis also studies interaction designs that surface evidence and constrain generation through retrieval and verification mechanisms [4, 24, 26], building on our prior work on retrieval-augmented strategies for improving medical chatbot reliability [20]. These design principles connect socio-normative evaluation with practical reliability concerns, including ethical frameworks for clinical advice [2].

4 PLANNED WORK

Two contributions remain: **(1) Multilingual, values-oriented trait attribution.** A multilingual cloze probe will test how models link social roles (e.g., nurse, manager) to evaluative traits across English, Italian, and Persian, using prompts grounded in human values and role ethics from prior alignment work [3]. We target cross-lingual framing divergence: the same role may be cast as more competent, compliant, or caring across languages under matched semantics; **(2) Mitigation-oriented evaluation for autonomy and fairness.** We will test interventions that reduce autonomy-steering and biased framing while preserving helpfulness, including autonomy-supportive templates (options, uncertainty cues, goal reflection) and feedback-oriented prompting, on the autonomy and narrative benchmarks.

5 CONCLUSION

This research advances socio-normative evaluation of LLM advisor agents by operationalizing autonomy and representational fairness as properties of *situated interaction*. The benchmarks show that LLMs can systematically express stance preferences in advice and that implicit bias can emerge under small, realistic identity cues, even when explicit derogatory content is absent [28]. The remaining thesis work will (i) extend the analysis to cross-lingual role/trait framing using parallel cloze prompts and (ii) validate mitigation strategies through controlled ablations and A/B prompting studies. A key deliverable is a reusable evaluation suite, e.g., scenario templates, annotations, and analysis scripts, to help the community compare advisory agents on autonomy support and representational fairness, alongside conventional accuracy and safety metrics.

REFERENCES

[1] Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li Li-Jia, Ramesh Jain, and Amir M. Rahmani. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *npj Digital Medicine* 7, 1 (2024), 82. https://doi.org/10.1038/s41746-024-01074-z

[2] Richard C. Armitage. 2025. Implications of Large Language Models for Clinical Practice: Ethical Analysis Through the Principlism Framework. *Journal of Evaluation in Clinical Practice* 31, 1 (Feb. 2025), e14250. https://doi.org/10.1111/jep.14250

[3] Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 2025. Nothing Comes without Its World - Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24)*. AAAI Press, San Jose, California, USA, 61–73.

[4] Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023. Knowledge-Augmented Language Model Verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 1720–1736. https://doi.org/10.18653/v1/2023.emnlp-main.107

[5] Divij Bajaj, Yuanyuan Lei, Jonathan Tong, and Ruihong Huang. 2024. Evaluating Gender Bias of LLMs in Making Morality Judgements. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 15804–15818.

[6] Stefan Buijsman, Sarah E. Carter, and Juan-Pablo Bermúdez. 2025. Autonomy by Design: Preserving Human Autonomy in AI Decision-Support. *Philosophy & Technology* 38 (2025), 97. https://doi.org/10.1007/s13347-025-00932-2

[7] Christopher Burr, Nello Cristianini, and James Ladyman. 2018. An analysis of the interaction between intelligent software agents and human users. *Minds and Machines* 28, 4 (2018), 735–774.

[8] Celso M. de Melo, Stacy Marsella, and Jonathan Gratch. 2017. Increasing Fairness by Delegating Decisions to Autonomous Agents. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (São Paulo, Brazil) *(AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 419–425.

[9] Christopher Diebel, Marc Goutier, Martin Adam, and Alexander Benlian. 2025. When AI-Based Agents Are Proactive: Implications for Competence and System Satisfaction in Human–AI Collaboration. *Business & Information Systems Engineering* 67, 2 (2025), 289–303. https://doi.org/10.1007/s12599-024-00918-y

[10] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8173–8188. https://doi.org/10.18653/v1/2020.emnlp-main.656

[11] Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. 2023. Probing Explicit and Implicit Gender Bias through LLM Conditional Text Generation. arXiv:2311.00306 [cs.CL] https://arxiv.org/abs/2311.00306

[12] Shir Etgar, Gal Oestreicher-Singer, and Inbal Yahav. 2024. Implicit bias in LLMs: Bias in financial advice based on implied gender. Available at SSRN.

[13] Norman Fairclough. 2013. *Critical Discourse Analysis: The Critical Study of Language* (2 ed.). Routledge, London and New York.

[14] Dennis Fast et al. 2024. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digital Medicine* 7, 1 (2024), 1–14.

[15] Elias Fernández Domingos, Inês Terrucha, Rémi Suchon, Jelena Grujić, Juan C Burguillo, Francisco C Santos, and Tom Lenaerts. 2022. Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific reports* 12, 1 (2022), 8492.

[16] Andrea Ferrario, Alessandro Facchini, and Alberto Termine. 2024. Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems. *Minds and Machines* 34, 3 (2024), 30.

[17] Paul Formosa. 2021. Robot autonomy vs. human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy. *Minds and Machines* 31, 4 (2021), 595–616.

[18] Robert Gaßner and Karlheinz Steinmüller. 2019. Scenarios that tell a story. Normative narrative scenarios–an efficient tool for participative innovation-oriented foresight. In *Envisioning Uncertain Futures: Scenarios as a Tool in Security, Privacy and Mobility Research*. Springer, Wiesbaden, Germany, 37–48.

[19] Saba Ghanbari Haez and Mauro Dragoni. 2025. Neutral Is Not Unbiased: Evaluating Implicit and Intersectional Identity Bias in LLMs Through Structured Narrative Scenarios. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 15060–15088. https://doi.org/10.18653/v1/2025.findings-emnlp.814

[20] Saba Ghanbari Haez, Marina Segala, Patrizio Bellan, Simone Magnolini, Leonardo Sanna, Monica Consolandi, and Mauro Dragoni. 2024. A Retrieval-Augmented Generation Strategy to Enhance Medical Chatbot Reliability. In *Artificial Intelligence in Medicine: 22nd International Conference, AIME 2024, Salt Lake City, UT, USA, July 9–12, 2024, Proceedings, Part I* (Salt Lake City, UT, USA). Springer-Verlag, Berlin, Heidelberg, 213–223. https://doi.org/10.1007/978-3-031-66538-7_22

[21] Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science* 5, 1 (2025), 65–75.

[22] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

[23] Mahammed Kamruzzaman, Md. Shovon, and Gene Kim. 2024. Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 8940–8965. https://doi.org/10.18653/v1/2024.findings-acl.530

[24] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[25] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. https://doi.org/10.1145/3582269.3615599

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[27] Catriona Mackenzie. 2008. Relational autonomy, normative authority and perfectionism. *Journal of Social Philosophy* 39, 4 (2008), 512–533.

[28] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality* 15, 2 (2023), 1–21.

[29] Carina Prunkl. 2022. Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence* 4, 2 (2022), 99–101.

[30] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3407–3412. https://doi.org/10.18653/v1/D19-1339

[31] Karan Singhal et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[32] Nicole Smith-Vaniz, Harper Lyon, Lorraine Steigner, Ben Armstrong, and Nicholas Mattei. 2025. Investigating Political and Demographic Associations in Large Language Models Through Moral Foundations Theory. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. AAAI Press, Madrid, Spain, 2419–2430. https://doi.org/10.1609/aies.v8i3.36727

[33] Simona Tiribelli et al. 2023. The AI ethics principle of autonomy in health recommender systems. *Argumenta* 16 (2023), 1–18.

[34] Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* 7, 3 (2025), 400–411.

[35] Linda Trinkaus Zagzebski. 2012. *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford University Press, New York, NY. https://doi.org/10.1093/acprof:oso/9780199936472.001.0001