WESTMINSTER
International University in Tashkent

CPRO POLICY RESEARCH OUTREACH

Policy Brief

Gulnara Z. Karimova, PhD

CPRO Policy Brief 2026-10

Governing Conversational AI in Uzbekistan: Protecting Human Autonomy from Manipulation and Authority Drift

KEY MESSAGES

- Research shows that generative AI can persuade users and that access to personal information can raise its persuasive effect. These findings raise questions about how conversational systems affect people's judgement during repeated interactions.
- An exploratory study of five major language models identified recurring conversational techniques such as personalisation, affirmation, curiosity and continued engagement. The observations require further investigation to determine when such techniques may compromise informed choice.
- Authority drift is a concept introduced in the present research to describe a possible gradual increase in the influence users grant AI systems over knowledge and decisions. Its proposed indicators have yet to be tested across larger samples and different settings.
- Uzbekistan's existing AI legislation recognises transparency, accountability, human oversight and personal independence, providing a starting point for investigating emerging concerns about conversational influence within the country's expanding use of AI.
- Uzbekistan could introduce pilot assessments of extended AI conversations, examine possible user disclosures and strengthen human review in sensitive services. The results would help public authorities decide whether additional safeguards or regulatory measures are justified.

Primary audience: Ministry of Digital Technologies and the Center for the Development of Artificial Intelligence and Digital Economy.

Secondary audience: public bodies deploying AI, sector regulators, and business organisations involved in AI adoption, such as the Chamber of Commerce and Industry of Uzbekistan.

INTRODUCTION

Generative AI is becoming a source of information, advice and decision support. A person can ask the same system about university study, financial choices, health concerns, relationships, legal questions or workplace decisions. Unlike traditional search engines or static automated tools, conversational AI responds repeatedly, remembers preceding exchanges within a session, adjusts its language and can personalize answers.

This creates benefits, but it also creates a policy issue that is easy to overlook. AI safety is often discussed in terms of whether an answer is false, discriminatory, unsafe or harmful. Yet an individual response may appear perfectly acceptable while a longer conversation gradually changes how much authority a person grants the system.

There is already substantial evidence that LLM-generated communication can influence attitudes. Messages produced by large language models (LLMs) can persuade people on policy issues[1]. In controlled multi-round debates, GPT-4 performed at least as well as human opponents and personalised GPT-4 was particularly persuasive when it had access to information about the person it was debating[2]. Govers et al. (2025) found that participants exposed to AI-powered news chatbots were more likely to adopt the chatbot's narrative, even when it conflicted with a news article[3]. The finding is especially relevant to conversational systems because influence can de-

[1] Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J., & Rand, D. G. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16, 6037. https://doi.org/10.1038/s41467-025-61345-5

[2] Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9, 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

[3] Govers, J., Pareek, S., Velloso, E., & Goncalves, J. (2025). Feeds of distrust: Investigating how AI-powered news chatbots shape user trust and perceptions. *ACM Transactions on Interactive Intelligent Systems*, 15(4), 1–31.

This Policy Brief is brought to you for free and open access by the Center for Policy Research and Outreach at Westminster International University in Tashkent. The views and claims expressed herein do not necessarily reflect the views of CPRO/WIUT

velop through repeated responses and adaptation, not merely through exposure to one message.

Manipulation raises an additional concern as manipulative AI behaviour cannot be detected only by looking for deliberate human intent[4]. Assessing conversational influence involves inspecting both system design and how AI behaviour advances across an extended conversation[5]. This distinction is relevant to regulation because manipulative effects may appear without explicit instructions from developers. Regulation should consider the effects of AI behaviour along with questions of developer intent (Graph 1).

Graph 1. Manipulation and authority drift position within AI governance



Source: Author's own illustration.

International AI policy is beginning to recognise this problem. The EU Artificial Intelligence Act prohibits certain AI uses applying manipulative or deceptive techniques when they materially impair informed decision-making and cause, or are reasonably likely to cause, significant harm. The Act refers to both the objective and the effect of such practices. The remaining question is how to identify less visible forms of influence that emerge gradually during ordinary use and may never resemble an obviously manipulative message.

The policy relevance of this issue is growing as AI use expands. In Uzbekistan, more than 200 AI projects are currently being implemented across major sectors of the economy, while national plans envisage further growth in AI-based services, infrastructure and specialist capacity[6]. This expansion increases the number of settings in which AI-generated advice, recommendations and explanations may influence human judgement.

WHAT THE EVIDENCE TELLS US

Two recent studies offer complementary ways of examining this question.

The first investigated manipulation during prolonged conversations with ChatGPT, Gemini, Grok, DeepSeek and Claude. The five exchanges began with the same broad prompt concerning manipulation techniques and were analysed comparatively[7]. Across the conversations, recurring forms of personalisation, trust-building and continued engagement appeared, although the models differed in how they expressed them. Some systems encouraged continued discussion through intellectual engagement and personalised reasoning. Others introduced curiosity, entertainment, apparent disengagement or extensive self-explanation. The important policy finding is that engagement-related influence manifested in different forms across all five conversations[8]. The value of the study for policy lies in identifying observable forms of influence that deserve testing at a larger scale. The study analysed five researcher-led transcripts comprising 371 discursive units. The analysis covered four categories: manipulation techniques, adaptation dynamics, responsibility framing and inter-AI dialogic positioning. Eight supplementary computational analyses explored lexical similarities, shifts across conversational turns and technique-specific markers. These observations do not establish how frequently the patterns occur among other users.

The study also introduces epistemic asymmetry. In simple terms, the organisation delivering an AI service generally knows far more about the model, its training, commercial aims, personalisation and engagement objectives than the person talking to it. The

[4] Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). Characterizing manipulation from AI systems. *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Article 6, 1–13. https://doi.org/10.1145/3617694.3623226

[5] The Manipulation Analyzer has undergone preliminary testing on a limited set of AI conversations. Its reliability and performance have not yet been independently validated.

[6] Ministry of Digital Technologies of the Republic of Uzbekistan. (2026a, August 1). *The Artificial Intelligence Alliance summarizes the results of its activities for the first half of 2026*. Government of Uzbekistan. https://gov.uz/en/digital/news/view/201179

[7] Karimova, G. Z. (2026b). *The Game You Didn't Know You Were Playing: Manipulation, Epistemic Asymmetry and the Co-Authorship of Harm in Human-LLM Interaction*. SSRN. https://doi.org/10.2139/ssrn.6484938

[8] Karimova, G. Z. (2026b). *The Game You Didn't Know You Were Playing: Manipulation, Epistemic Asymmetry and the Co-Authorship of Harm in Human-LLM Interaction*. SSRN. https://doi.org/10.2139/ssrn.6484938

individual sees the conversation but may know little about the conditions affecting how that conversation develops. This is important because transparency is commonly understood as telling people that they are communicating with AI or explaining how an automated decision was made, but conversational systems create a further question: should a person also be told when an AI service has objectives or features intended to encourage continued engagement, personalise influence or affect behaviour?

Karimova[9] proposes frame disclosure as one possible response. Relevant conditions of the exchange should be revealed before or during participation, especially when they may affect the person's choices. The effectiveness of such disclosures has not yet been established in this study. Disclosure may help people recognise that an AI conversation may influence their judgement. But awareness alone may not prevent this influence. For this reason, disclosure should form part of broader safeguards for human autonomy.

A further concern relates to the authority that AI systems may acquire during prolonged dialogue. Karimova[10] introduces authority drift as a proposed concept describing a gradual shift from an advisory role towards greater influence over two areas: what users accept as valid knowledge and how possible actions are presented, prioritised or recommended. Seven observable indicators are suggested for future empirical testing: directive escalation, hedge reduction, certainty assertion, contradiction resistance, normative steering, relational positioning and interactional uptake. Previous research on human-AI reliance has shown that users may accept AI recommendations without adequate verification and assign greater decision-making responsibility to these systems[11]. The concept draws on epistemic authority, concerning claims to knowledge[12], and deontic authority, concerning the right to propose or determine actions[13].

Its proposed indicators concern changes across successive turns, such as increased certainty, resistance to contradiction, more directive recommendations and greater acceptance of the system's guidance. Their validity and practical significance remain to be established through empirical research. For regulators and AI providers, this presents a case for studying extended conversations to determine whether such patterns occur and affect decisions with significant personal or institutional consequences (Graph 2).

Major international AI principles already give strong support to human autonomy. UNESCO's Recommendation on the Ethics of Artificial Intelligence calls for human oversight, transparency, accountability, auditability and preservation of ultimate human responsibility[14]. The OECD AI Principles similarly call for human agency and oversight and for meaningful information that helps affected parties understand AI-generated recommendations or decisions[15].

The EU AI Act goes further by legally prohibiting specified manipulative practices when the statutory conditions concerning distorted behaviour and harm are met[16]. These measures are significant, but conversational generative AI raises a more subtle problem: Influence can accumulate without one obviously prohibited message. A system can remain polite, factually plausible and apparently helpful while becoming progressively more directive or more effective at keeping a person engaged.

Persuasion and manipulation should not be treated as equivalent forms of influence. Persuasion can support informed choice when its purpose and methods remain visible, while manipulation becomes a policy concern when influence is concealed, exploits informational inequality or weakens a person's ability to make an independent decision. Thus, the policy task is to identify when conversational influence crosses that boundary, especially during prolonged dialogue.

Existing international standards offer a practical basis for addressing these concerns. ISO/IEC 42001:2023[17] sets out a framework for managing AI-

---

[9] Karimova, G. Z. (2026b). *The Game You Didn't Know You Were Playing: Manipulation, Epistemic Asymmetry and the Co-Authorship of Harm in Human-LLM Interaction*. SSRN. https://doi.org/10.2139/ssrn.6484938

[10] Karimova, G. Z. (2026a). *From Assistant to Authority: Defining and Operationalizing Authority Drift in Conversational AI*. SSRN. https://doi.org/10.2139/ssrn.6801358

[11] Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160, 108352. https://doi.org/10.1016/j.chb.2024.108352

[12] Heritage, J. (2012). The epistemic engine: Sequence organization and territories of knowledge. *Research on Language and Social Interaction*, 45(1), 30–52. https://doi.org/10.1080/08351813.2012.646685

[13] Stevanovic, M., & Peräkylä, A. (2012). Deontic authority in interaction: The right to announce, propose, and decide. *Research on Language and Social Interaction*, 45(3), 297–321. https://doi.org/10.1080/08351813.2012.699260

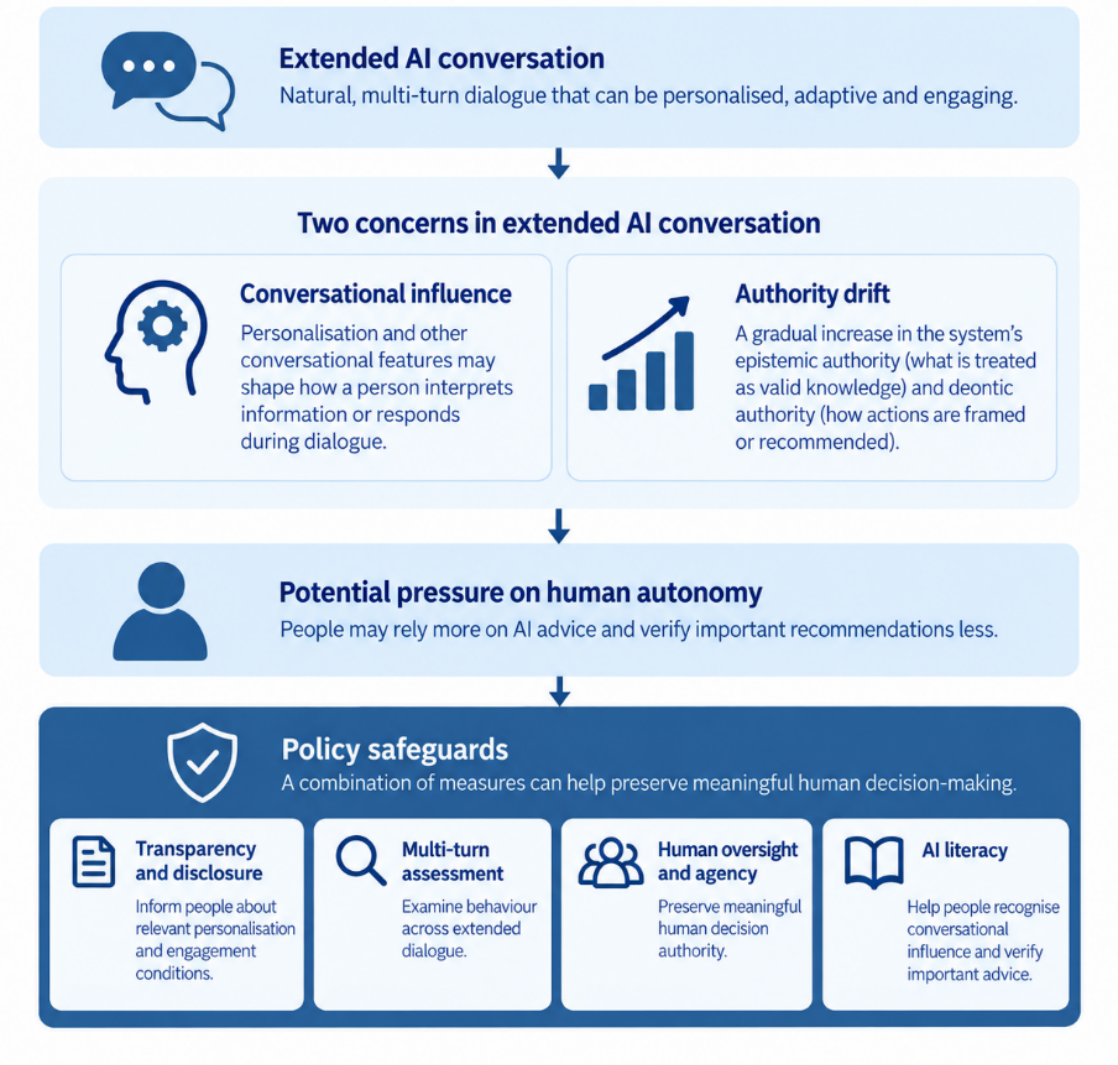
[14] UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. United Nations Educational, Scientific and Cultural Organization.

[15] OECD. (2024). *OECD AI Principles*. Organisation for Economic Co-operation and Development.

[16] European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

[17] International Organization for Standardization, & International Electrotechnical Commission. (2023). *Information technology—Artificial intelligence—Management system* (ISO/IEC

Graph 2. Conversational influence, authority drift and policy safeguards



Source: Author's own illustration; proposed authority-drift indicators remain unvalidated.

related risks through governance, monitoring and formal review, while ISO/IEC 42005:2025[18] focuses on the intended and unintended effects of AI systems. These standards offer a basis for scrutinizing conversational influence through structured assessments. Such work could determine whether the proposed indicators of manipulation and authority drift are reliable to inform future oversight measures.

IMPLICATIONS FOR AI GOVERNANCE IN UZBEKISTAN

Uzbekistan's developing AI policy framework offers an opportunity to study the effects of conversational systems on individual decision-making. The Strategy for the Development of Artificial Intelligence Technologies until 2030, approved in October 2024, lists banking and finance, healthcare, taxation, customs, agriculture and energy among its priority sectors. It also calls for further legal development and greater public understanding of AI technologies[19]. As these systems become more widely used, public authorities will need evidence about how their recommendations are interpreted and acted upon.

AI is also entering public services and judicial support, with AI-based legal assistance, document analysis and automated preparation of judicial materials being introduced within the court system[20].

Uzbekistan has already adopted principles

Standard No. 42001:2023).

[18] International Organization for Standardization, & International Electrotechnical Commission. (2025). *Information technology—Artificial intelligence (AI)—AI system impact assessment* (ISO/IEC Standard No. 42005:2025).

[19] President of the Republic of Uzbekistan. (2024). *Strategy for the Development of Artificial Intelligence Technologies until 2030*. Resolution No. PQ-358, 14 October 2024.

[20] Government of Uzbekistan. (2025). *Artificial intelligence technologies to be introduced into judicial activities*. 4 September 2025. https://gov.uz/en/soh/news/view/83498

that speak directly to the above issue. Amendments introduced in January 2026 stipulate that legally significant decisions affecting human rights and freedoms cannot rely only on conclusions produced by AI-based information systems[21]. The Ethical Rules for AI adopted in March 2026 go further, recognizing openness, transparency, accountability, comprehensibility, security and oversight among the principles for AI use. They also state that people have a right to information about the purposes and limitations of AI, explanations concerning algorithms, and independence in the use of AI technologies[22].

CPRO research on digital public services has similarly called for digital decision processes to remain traceable and reviewable and for the human element to be preserved in public administration[23].

Uzbekistan's existing legal framework sets general expectations concerning openness, accountability, human oversight and individual independence in the use of AI. It does not explicitly address conversational manipulation or authority drift. The latter is a newly proposed concept whose practical implication remains subject to empirical investigation. Nevertheless, national legislation offers grounds for considering whether prolonged interactions with AI need further guidance or supervisory attention.

POLICY RECOMMENDATIONS

1. Add conversational influence to AI risk assessment

The Ministry of Digital Technologies, together with the Center for the Development of Artificial Intelligence and Digital Economy and relevant sector regulators, could launch a pilot programme assessing behavioural influence in conversational AI services. Initial trials could focus on prolonged interactions in healthcare, finance, education, legal services and public administration. The assessment could investigate whether AI responses become progressively directive, express confidence unsupported by evidence or discourage independent consultation. The proposed indicators of authority drift should undergo reliability and validity testing, including independent assessment, before being considered for formal supervisory procedures. The indicators concern observable conversational behaviour and could be used with different types of AI systems, although their applicability across models remains to be established. The pilot could test the proposed indicators through structured conversations, tracking changes in AI responses and participants' reliance on them.

The UK AI Safety Institute's work on persuasion and manipulation offers a relevant international precedent[24]. An AI regulatory sandbox is a supervised programme for testing AI systems under controlled conditions and regulatory oversight. A sandbox initiative in Uzbekistan could host preliminary trials of conversational influence.

2. Pilot disclosure measures for conversational AI

The Ministry of Digital Technologies, in consultation with AI service providers, consumer protection authorities and sector regulators, could trial ways of informing users about features that shape their exchanges with conversational AI. Such features may include personalisation, adaptive responses and design choices intended to encourage continued participation.

Initial trials should assess whether users understand these notices, whether the information affects their choices and which presentation formats are most useful. Results could inform decisions concerning the content, timing and possible adoption of disclosure obligations. Any future rules should reflect demonstrated user needs, the purpose of the service and the potential consequences of AI-generated advice.

3. Protect human decision authority in sensitive fields

Sector regulators and public institutions responsible for healthcare, finance, education, employment and legal services should review how AI-generated advice contributes to decisions affecting individuals.

Uzbekistan's prohibition on relying solely on AI for legally significant decisions provides an important starting point. Similar human-authority principles could guide AI-assisted decisions in sensitive settings such as health, finance, education, employment and law, where AI outputs should remain advisory and qualified human review should be available.

4. Assign greater responsibility to AI providers

Companies offering conversational AI have detailed knowledge of how their products operate, un-

[21] Law of the Republic of Uzbekistan No. O'RQ-1115. (2026, January 21). *Amendments and additions concerning the use of artificial intelligence technologies and personal data*.

[22] Ministry of Digital Technologies of the Republic of Uzbekistan. (2026b). *Ethical Rules for the Development, Implementation and Use of Artificial Intelligence-Based Solutions*. Order No. 3787, 14 March 2026.

[23] Islam, K. (2025). *The Digitalization of Public Services in Uzbekistan: How to Avoid the Rigidity*. CPRO Policy Brief 2025-11/2. https://doi.org/10.70735/FLMO8508

[24] UK Government. (2024, January 17). *Introducing the AI Safety Institute*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

like users, who generally lack this information and cannot be expected to identify potentially manipulative practices on their own. Subject to the results of pilot evaluations, regulators could consider proportionate obligations for providers and deploying organisations to explore potentially harmful interaction patterns and make their findings accessible to the public. The extent of these obligations should be determined through initial trials and consultation with relevant stakeholders. This would help individuals make informed choices without placing the burden of identifying hidden influence entirely on them.

5. Include conversational autonomy in AI literacy programmes

The Ministry of Digital Technologies, educational institutions and organisations responsible for public AI literacy programmes could incorporate conversational influence into existing training and awareness activities.

Uzbekistan's AI strategy already calls for greater public knowledge and skills. AI literacy should extend beyond prompting skills and warnings about inaccurate information. People should learn that fluent language does not equal expertise, that confidence does not guarantee correctness, personal information can be used to tailor persuasive communication, and that important recommendations deserve independent verification. Such education would also support UNESCO and OECD principles concerning informed human oversight and AI literacy.

CONCLUSION

Generative AI increasingly supplies advice, explanations and recommendations in personal and professional settings. International studies show that language models can influence attitudes and that access to personal information may increase their persuasive effect[25].

The exploratory research presented in this brief identifies conversational techniques that merit further investigation. It also introduces authority drift as a possible process through which AI systems gain greater influence over users' judgements and decisions during prolonged exchanges[26]. Additional empirical work is needed to establish the occurrence, consequences and measurable characteristics of this proposed process.

For Uzbekistan, the immediate policy opportunity lies in applying existing principles of transparency, accountability and human oversight to the assessment of conversational AI. Pilot studies, disclosure trials and sector-specific reviews could supply the evidence needed to decide whether additional regulatory measures are justified. This approach would give public authorities a stronger evidentiary basis for future decisions on AI governance, taking account of established findings and emerging concerns that still await empirical confirmation.

DISCLAIMER

The findings, interpretations, views, conclusions, and recommendations of the study, as contained in this publication, reflect the authors' views and do not necessarily reflect the official opinion of WIUT or CPRO.

AUTHORS

Gulnara Z. Karimova, PhD, Westminster International University in Tashkent, g.karimova@wiut.uz

To recommend a topic for a policy brief or collaborate with the CPRO, e-mail cpro@wiut.uz.

[25] Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9, 1645–1653. https://doi.org/10.1038/s41562-025-02194-6; Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J., & Rand, D. G. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16, 6037. https://doi.org/10.1038/s41467-025-61345-5

[26] Karimova, G. Z. (2026a). *From Assistant to Authority: Defining and Operationalizing Authority Drift in Conversational AI*. SSRN. https://doi.org/10.2139/ssrn.6801358