პროპაგანდა ალგორითმებისთვის - საინფორმაციო მანიპულაციის ახალი სამიზნეები AI ეპოქაში

ავტორი: დავით ქუტიძე[1]

რედაქტორი: ირინა გურგენაშვილი[2]

გენერაციული ხელოვნური ინტელექტი და Deepfake (AI-სა და მანქანური სწავლების ალგორითმების გამოყენება ისეთი ყალბი ფოტოს, ხმისა და ვიდეოს შესაქმნელად, რომლებიც რეალურს ძალიან ჰგავს) ტექნოლოგიები ქმნის ისეთ საინფორმაციო გარემოს, რომელშიც ფაქტისა და სიყალბის ერთმანეთისგან გარჩევა მნიშვნელოვნად არის გართულებული.[3] გარდა ამისა, არსებობს სხვა, არანაკლებ მნიშვნელოვანი საფრთხეც. როგორც ცნობილია, განსხვავებული პოლიტიკური თუ ფინანსური ინტერესის მქონე აქტორების მიერ მართული მანიპულაციური საინფორმაციო კამპანიების ძირითადი მიზანია ადამიანების ცნობიერებაზე გავლენის მოხდენა. თუმცა ხელოვნური ინტელექტის ერაში საინფორმაციო გავლენის ოპერაციების ახალი სამიზნე მხოლოდ ადამიანი აღარ არის. უკვე მნიშვნელოვანია არა მარტო ის, თუ რას ფიქრობს ან გრძნობს ინდივიდი კონკრეტულ ამბავზე, არამედ ისიც, თუ რას „ხედავს“ Google-ის საძიებო ალგორითმი, სხვადასხვა სოციალური ქსელის რეკომენდაციის (suggestion) სისტემა და, უპირველეს ყოვლისა, გენერაციული ხელოვნური ინტელექტი. თუ ტრადიციული პროპაგანდა ცდილობდა ადამიანის რწმენის შეცვლას ან მისი ემოციებით მანიპულირებას, ახალი ტიპის გავლენის ოპერაციები ცდილობს შეცვალოს ის, რასაც მანქანა „სწავლობს“, „იმეორებს“, ან „ირჩევს“ ციტირებისთვის, რაც სამომავლოდ ავტომატურად აისახება მის მიერ მილიონობით ადამიანისთვის გაცემულ პასუხებში. ამდენად, დღეს ინტერნეტში ინფორმაციის ერთადერთი აღმომჩენი აღარ არის ადამიანი, რომელიც თავად კითხულობს, ადარებს და აფასებს წყაროებს, არამედ ალგორითმიც: საძიებო სისტემის რანჟირების მექანიზმი, სოციალური მედიის რეკომენდაციის სისტემა და გენერაციული ხელოვნური ინტელექტის

[1] მკვლევარი, კვლევითი ინსტიტუტი Gnomon Wise; e-mail: d.kutidze@ug.edu.ge
[2] აღმასრულებელი დირექტორი, კველევითი ინსტიტუტი Gnomon Wise, e-mail: I.gurgenashvili@ug.edu.ge
[3] ქუტიძე, დ. (30 აპრილი, 2026). *ყალბი ამბები ხელოვნური ინტელექტის ეპოქაში - საფრთხეები და მათი დაძლევის გზები*. კვლევითი ინსტიტუტი Gnomon Wise. ხელმისაწვდომია: https://gnomonwise.org/ge/publications/analytics/315

მოდელი, რომელიც პასუხობს მისთვის დასმულ კითხვებს ერთიანი, თავდაჯერებული ტექსტის სახით, ხშირად - წყაროს მითითების გარეშეც.

ვულისა და ჰოვარდის (2018) კლასიკური განმარტებით, „გამოთვლითი პროპაგანდა“ (computational propaganda), რომელიც სწორედ ალგორითმების დამუშავებას აღწერს, არის „სოციალური მედიაპლატფორმების, ავტონომიური აგენტებისა და დიდი მონაცემების ერთობლიობა, რომელიც მიმართულია საზოგადოებრივი აზრის მანიპულირებისკენ“.[4] ეს განმარტება დაშენებულია იმ დაშვებაზე, რომ საბოლოო სამიზნე მაინც ადამიანია, თუნდაც შუამავალი ბოტი ან ალგორითმი იყოს. ბოტები ბაძავენ ადამიანურ ქცევას, რათა შექმნან ხელოვნური კონსენსუსის შთაბეჭდილება რეალური ადამიანებისთვის. თუმცა ბოლო წლების ლიტერატურა აჩვენებს ამ მოდელის თანდათანობით გართულებას: ალგორითმი აღარ არის მხოლოდ არხი, რომლითაც შეტყობინება მიემართება ადამიანისკენ, არამედ თავად ხდება შუალედური „აუდიტორია“, რომლის „დარწმუნებაც“ შესაძლოა უფრო ეფექტური იყოს, ვიდრე მილიონობით ინდივიდზე ზემოქმედება. თუ საძიებო სისტემა ან ჩატბოტი ერთხელ „ისწავლის“ გარკვეულ ნარატივს, ან ავტორიტეტულ წყაროდ აღიქვამს კონკრეტულ საიტს, ეს გავლენა შემდეგ ავტომატურად, დანახარჯების გარეშე, მრავლდება ყოველ მომდევნო მომხმარებელზე, ენასა და ქვეყანაზე. რაც მთავარია, თუკი, თეორიულად, ვინმე შეძლებს ალგორითმებისა და ხელოვნური ინტელექტის „დარწმუნებას“ თავისი ინტერესების სასარგებლოდ, ის ისტორიას საკუთარი ინტერპრეტაციით გადაწერს.

ალგორითმებისა და ხელოვნური ინტელექტის „დარწმუნების“ მცდელობა უკვე შეინიშნება და დოკუმენტირებულია არაერთ კვლევაში. ამ პროცესს ხელს უწყობს საინფორმაციო სივრცის მნიშვნელოვანი მახასიათებელი - ე.წ. „მონაცემთა სიცარიელეები“ (data voids). ეს ტერმინი 2019 წელს შემოგვთავაზეს გოლებიევსკიმ და ბოიდმა.[5] მათი განმარტებით, მონაცემთა სიცარიელეები, იგივე საინფორმაციო ვაკუუმი, წარმოიქმნება მაშინ, როდესაც იშვიათ საძიებო სიტყვებზე ინტერნეტში ძალიან ცოტა მასალა იძებნება. ამით კარგად სარგებლობენ მანიპულატორები, რომლებსაც საკუთარი იდეოლოგიური, ეკონომიკური თუ პოლიტიკური ინტერესები ამოძრავებთ. ავტორები გამოყოფენ რამდენიმე ტიპის სიცარიელეს: 1. ცხელი

[4] Samuel C. Woolley and Philip N. Howard. (2016). Automation, Algorithms, and Politics| Political Communication, Computational Propaganda, and Autonomous Agents. *International Journal of Communication*. Via link: https://www.researchgate.net/publication/402435515_Automation_Algorithms_and_Politics_Political_Communication_Computational_Propaganda_and_Autonomous_Agents_-_Introduction

[5] Golebiewski, M. and Boyd, D. (2019). Data Voids: Where Missing Data Can Easily Be Exploited. *Data & Society*. Via link: https://apo.org.au/sites/default/files/resource-files/2019-10/apo-nid265631.pdf

სიახლის (breaking news) შედეგად გაჩენილი - მანიპულაციური ინფორმაციის შექმნა ისეთ საძიებო ტერმინებზე, რომლებზეც მოთხოვნა უეცრად იზრდება, მიმდინარე რეზონანსული მოვლენის ფონზე. საბოლოოდ, ეს სიცარიელეები ხარისხიანი საინფორმაციო შინაარსით ივსება, თუმცა სანამ ასეთი კონტენტი შეიქმნება, მათ ბოროტად იყენებენ; 2. ახალი სტრატეგიული ტერმინები - მანიპულატორები იგონებენ ახალ ტერმინებს და მათ გარშემო წინასწარ ქმნიან საინფორმაციო ქსელს. შემდეგ ამ ტერმინებს მასებში (ხშირად მედიის საშუალებით) ავრცელებენ, რათა აუდიტორიას მხოლოდ თავიანთვის სასურველი, შეცდომაში შემყვანი ინფორმაცია და ნარატივები მიაწოდონ; 3. მოძველებული ტერმინები - როდესაც რაიმე მოვლენის შესახებ ინფორმაცია მოძველდება, კონტენტის შემქმნელები (მაგალითად, სანდო მედია) მათთან დაკავშირებული მასალის წარმოებას გაცილებით ადრე აჩერებენ, ვიდრე მომხმარებლები ამ ტერმინების ძებნას შეწყვეტენ. მანიპულატორები კი ამით სარგებლობენ და რაკი სანდო მედიასაშუალებები მოძველებულ თემაზე ახალს აღარაფერს წერენ, ისინი თავად ქმნიან მათთვის ხელსაყრელ ინფორმაციას. Google-ისა თუ სხვა საძიებო სისტემების ალგორითმები კი პრიორიტეტს ხშირად სწორედ ახლად გამოქვეყნებულ მასალას ანიჭებენ. შედეგად, საძიებო სისტემა პირველ ადგილებზე სწორედ მანიპულატორების მიერ ახალგამოქვეყნებულ, დეზინფორმაციულ შინაარსს აჩვენებს; 4. ფრაგმენტირებული ცნებები - მანიპულატორები ერთმანეთთან დაკავშირებულ იდეებს აცალკევებენ და ქმნიან მათგან განსხვავებულ საინფორმაციო ჯგუფებს (ე.წ. „ბაბლები“), რომლებიც სხვადასხვა პოლიტიკურ ჩარჩოს ერგება. ამ გზით, მანიპულატორები აუდიტორიას უმალავენ სრულ სურათს და კონტექსტიდან გლეჯენ ფაქტებს, რათა ადამიანებმა მოვლენათა შორის ლოგიკური ჯაჭვის დანახვა ვერ შეძლონ, რაც, საბოლოოდ, ხელს უწყობს საზოგადოების პოლარიზაციას; 5. პრობლემური საძიებო მოთხოვნები - მწვავე და სენსიტიური საკვანძო სიტყვები, რომლებიც მჭიდროდ უკავშირდება საზოგადოებაში არსებულ უთანხმოებებს, მწვავე სოციალურ საკითხებს, შეთქმულების თეორიებს ან ისტორიულ ტრავმებს. თუ წარსულში გარკვეულ სიტყვაზე ძირითადად დეზინფორმაცია ან რადიკალური თვალსაზრისი იძებნებოდა, საძიებო სისტემა ინერციით განაგრძობს იმავეს ჩვენებას, რადგან საპირწონე სანდო წყაროები, უბრალოდ, არ არსებობს.

აღსანიშნავია, რომ მონაცემთა სიცარიელეების ცნება თავდაპირველად შემუშავდა საძიებო სისტემების ანალიზისთვის, მაგრამ ბოლოდროინდელმა კვლევებმა აჩვენა, რომ იგივე ლოგიკა გამოსადეგია დიდი ენობრივი მოდელებისთვისაც. როცა კონკრეტულ თემაზე სანდო, ხარისხიანი წყარო ცოტაა, ხოლო დაბალხარისხოვანი ან პროპაგანდისტული მასალა უხვად მოიპოვება, მოდელს, ისევე, როგორც ადამიანს, არაფერი აქვს, გარდა იმისა, რაც ხელმისაწვდომია.

გენერაციული ხელოვნური ინტელექტის „მოწამვლის მცდელობები“ - რუსული Pravda-ს ქსელი

კვლევითი ორგანიზაციები, რომლებიც რუსულ დეზინფორმაციას სწავლობენ, ერთხმად აღნიშნავენ, რომ ე.წ. Pravda-ს ქსელი (ოფიციალური სახელწოდება - Portal Kombat), რომელიც ასობით ვებგვერდისგან შედგება, უკვე წლებია, ავრცელებს პრორუსულ ნარატივებს. 2022 წელს, რუსეთის უკრაინაში სრულმასშტაბიანი შეჭრის შემდეგ, ქსელი მკვეთრად გაფართოვდა და ატლანტიკური საბჭოს Digital Forensic Research Lab-ის (DFRLab) მონაცემებით, 2025 წლისთვის უკვე 80-ზე მეტ რეგიონსა და ქვეყანას მოიცავდა, ხოლო მისი კონტენტი დიდწილად მანქანურ თარგმანს ეყრდნობოდა.[6] განსაკუთრებით საყურადღებოა ის, თუ როგორ აღწერენ მკვლევრები ამ ქსელის შემადგენელ ვებგვერდებს. იანკოვიჩისა და ნიუპორტის (2025) თანახმად[7], Pravda-ს საიტებს არ გააჩნია ძიების ფუნქცია, საიტის მთავარი მენიუ. ასევე შესამჩნევია სტილისტური და ენობრივი ხარვეზები, რის გამოც ის ადამიანებზე ნაკლებად მორგებულია [Pravda-ს ვებგვერდი ქართულ ენაზეც არსებობს და ისიც მსგავსი ხარვეზებით ხასიათდება]. გარდა ამისა, აღნიშნულ საიტებზე ნაკლებად ფიქსირდება რეალური მომხმარებლების ვიზიტები. თავის მხრივ, ყველა ეს ნიშანი მეტყველებს იმაზე, რომ ამჟამად Pravda-ს ქსელი არ არის განკუთვნილი ადამიანებისთვის. ავტორთა დასკვნით, რეალური სამიზნე არიან ავტომატური აგენტები, საძიებო სისტემების ინდექსატორები და მონაცემთა შემგროვებელი ბოტები, რომლებიც კვებავენ დიდი ენობრივი მოდელების (LLM) სასწავლო მონაცემთა ბაზებს. Pravda-ს ქსელის მასშტაბის რაოდენობრივი შეფასება ერთ-ერთი ყველაზე თვალსაჩინო არგუმენტია მისი ავტომატიზებული ბუნების დასადასტურებლად. NewsGuard-ის კვლევის[8] თანახმად, მხოლოდ 2024 წელს ქსელმა 3.6 მილიონი სტატია/ახალი ამბავი გამოაქვეყნა, რაც გარკვეულწილად ინტეგრირდა დასავლური ხელოვნური ინტელექტის სისტემებში და მათ პასუხებს მანიპულირებული თუ პროპაგანდისტული ინფორმაციით „აინფიცირებს“. ამასთან, საერთაშორისო ანალიტიკური ცენტრის, GLOBSEC-ის, კვლევის[9] მიხედვით, 2025 წლის

[6] Châtelet, V and Lesplingart, A. (2025). Russia-linked Pravda network cited on Wikipedia, LLMs, and X. *The Digital Forensic Research Lab (DFRLab)*. Via link: https://dfrlab.org/2025/03/12/pravda-network-wikipedia-llm-x/?utm_source=chatgpt.com

[7] Newport, A and Jankowicz, N. (2025). Russian networks flood the Internet with propaganda, aiming to corrupt AI chatbots. *Bulletin of the Atomic Scientists*. Via link: https://thebulletin.org/2025/03/russian-networks-flood-the-internet-with-propaganda-aiming-to-corrupt-ai-chatbots/

[8] Sadeghi, M and Blachez, I. (2025). The Infection of Western AI Chatbots by a Russian Propaganda Network. *NewsGuard*. Via link: https://bit.ly/3UAeeg8

[9] Kubś, J. (2025). Global Offensive: Mapping the Sources Behind the Pravda Network. *GLOBSEC*. Via link: https://www.globsec.org/sites/default/files/2025-05/Pravda%20Network%20Report.pdf

მარტისთვის, Pravda-ს მთავარ საიტზე არსებული სტატიების 85%-ზე მეტი ერთმანეთისგან ერთ წუთზე ნაკლებ ინტერვალში იყო გამოქვეყნებული, რაც პრაქტიკულად გამორიცხავს ადამიანურ სარედაქციო პროცესს და მიუთითებს სრულ ავტომატიზაციაზე. მნიშვნელოვანია ისიც, რომ ქსელი ნაკლებად აწარმოებს ორიგინალურ კონტენტს, იგი უბრალოდ აკოპირებს სიახლეებს ისეთი წყაროებიდან, როგორებიცაა TASS და RT (რუსული პროპაგანდისტული სააგენტოები), ასევე ნაკლებად ცნობილი აგრეგატორები, რომელთა ღირებულებაც განისაზღვრება არა აუდიტორიის ზომით, არამედ წარმოებული კონტენტის მოცულობით - რაც, GLOBSEC-ის განმარტებით, მეტყველებს იმაზე, რომ მისი შემქმნელები ხარისხზე მეტად რაოდენობას ანიჭებენ უპირატესობას. საბოლოოდ, დამოუკიდებელ მკვლევართა სხვადასხვა ჯგუფი მიდის იმ დასკვნამდე, რომ აღნიშნული ქსელის მიზანი ალგორითმების სასურველი ინფორმაციით „გამოკვება“ და უკვე ბოლო წლებში, გენერაციული ხელოვნური ინტელექტის სისტემების „მოწამვლაა“. ლონდონის სტრატეგიული დიალოგის ინსტიტუტის უფროსი ანალიტიკოსის, ოლგა ტოკარიუკის თანახმად, ქსელი განსაკუთრებით ორიენტირებულია უკრაინის ხელისუფლების დისკრედიტაციაზე, პრორუსული ნარატივების გავრცელებასა და „ალტერნატიული აზრის“ ილუზიის შექმნაზე. ტოკარიუკის თქმით, საბოლოოდ, „ქსელი შექმნილია იმისთვის, რომ ინტერნეტი შეავსოს პროპაგანდისთვის ხელსაყრელი კონტენტით და რომ ეს კონტენტი შემდეგ მოხვდეს როგორც საძიებო სისტემებში, ისე ლეგიტიმურ მედიასაშუალებებსა და ხელოვნური ინტელექტის ჩატ-ბოტებშიც კი“.[10]

ზემოთ აღწერილი პროცესი წარმოადგენს AI Grooming-ის კლასიკურ მაგალითს. ეს ტერმინი გასული წლიდან დამკვიდრდა American Sunlight Project-ის ანგარიშის[11] შემდეგ. მასში აღწერილია მოვლენა, როდესაც ცალკეული აქტორები მიზანმიმართულად ავრცელებენ დიდი მოცულობის ყალბ თუ მიკერძოებულ კონტენტს იმ მიზნით, რომ დროთა განმავლობაში ის დარეგისტრირდეს დიდი ენობრივი მოდელების (LLM) სასწავლო მონაცემებში და ერთხმად წარმოჩინდეს, როგორც სანდო წყარო. ტერმინი მნიშვნელოვნად განასხვავებს ორ არაერთგვაროვან, თუმცა ურთიერთდაკავშირებულ პრაქტიკას - ერთი მხრივ, ხელოვნური ინტელექტის გამოყენებას მასობრივი დეზინფორმაციის წარმოებისთვის (AI როგორც ინსტრუმენტი) და, მეორე მხრივ,

[10] ციტირებულია რადიო თავისუფლების სტატიაში (18 აპრილი, 2025): *ყალბი რუსული Pravda, რომელიც სოციალური ქსელებისა და ხელოვნური ინტელექტის მოწამვლას ცდილობს*. ავტორი: სერგეი სტეცენკო. ხელმისაწვდომია: https://bit.ly/4iI7XJ7

[11] American Sunlight Project. (2025). *A Pro-Russia Content Network Foreshadows the Automated Future of Info Ops*. Via link: https://static1.squarespace.com/static/6612cbdfd9a9ce56ef931004/t/67fd396818196f3d1666bc23/1744648558879/PK+Report.pdf

თავად ენობრივი მოდელების „შიდა“ დაბინძურებას, მათი პასუხების მანიპულირების მიზნით. აკადემიურ დონეზე ამ ცნებას დეტალურად განიხილავს დიდიე დანე (2025).[12] დანეს თანახმად, LLM Grooming-ი შემდეგი ლოგიკით მოქმედებს - იგი „ტბორავს“ საჯაროდ ხელმისაწვდომ წყაროებს, მათ შორის, ვებარქივებს, გამოგონილი ნარატივების დიდი მოცულობით, რათა ისინი მოდელმა, ტრენინგის მეშვეობით, სტატისტიკურად აითვისოს და შემდგომ ფაქტობრივ ცოდნად გადააქციოს. სწორედ ამიტომ დანე მიიჩნევს, რომ საბაზისო კონტრდეზინფორმაციული ღონისძიებები, რომლებიც რეაქტიულია და ორიენტირებულია ცალკეულ ყალბ პუბლიკაციებზე, არაადეკვატურია სისტემური დაბინძურების წინააღმდეგ საბრძოლველად და საჭიროებს პროაქტიულ სტრატეგიებს, როგორებიცაა მონაცემთა ბაზების აუდიტი, უწყვეტი მონიტორინგი, ადამიანური ზედამხედველობის მექანიზმები და საერთაშორისო თანამშრომლობა.

ემპირიული მტკიცებულებები იმისა, რომ Pravda ქსელის კონტენტი უკვე აღწევს გენერაციული ხელოვნური ინტელექტის სისტემებამდე, რამდენიმე დამოუკიდებელი წყაროდან ვლინდება. DFRLab-ისა და ფინეთში დაფუძნებული CheckFirst-ის ერთობლივმა კვლევამ, Wikipedia-ს API-ის გამოყენებით, 44 ენაზე გამოქვეყნებულ 1 672 სტატიაში გამოავლინა 1 907 ჰიპერბმული, რომლებიც მიემართებოდა Pravda ქსელის 162 დომენს.[13] ამ ბმულების გავრცელება განსაკუთრებით შესამჩნევი იყო რუსულ (922 ბმული) და უკრაინულ (580 ბმული) ვიკიპედიებში, თუმცა 2022 წლის შემდეგ საგრძნობლად გაიზარდა ბმულების რაოდენობა ინგლისურენოვან ვერსიაშიც (133). ეს პრობლემურია იმდენად, რამდენადაც ვიკიპედია დიდი ენობრივი მოდელების ტრენინგის ერთ-ერთი მთავარი წყაროა. კვლევის ავტორებმა უშუალოდაც შეამოწმეს ეს კავშირი. კერძოდ, ChatGPT-ის, Copilot-ის, Perplexity-ისა და Gemini-ის ტესტირებისას აღმოჩნდა, რომ მოდელები აგენერირებდნენ Pravda-ს საიტებზე გამოქვეყნებულ კონტენტს, თუმცა არცერთი მათგანი მომხმარებელს არ აფრთხილებდა წყაროს რუსული წარმომავლობის თაობაზე, მიუხედავად იმისა, რომ ქსელის შესახებ საჯაროდ ხელმისაწვდომი კვლევები უკვე არსებობდა.[14]

ერთ-ერთი ყველაზე თვალსაჩინო მტკიცებულება მოდელის სასწავლო მონაცემების დაბინძურების შესახებ გამოვლინდა DFRLab-ის 2026 წლის კვლევიდან, რომელმაც გააანალიზა

[12] Didier Danet. LLM Grooming: A New Cognitive Threat to Generative AI. 2025. ⟨hal-05241525⟩. Via link: https://hal.science/hal-05241525v1/file/20251215%20DANET%20Cognitive%20Threats%20LLM%20Grooming%20V02.pdf

[13] Châtelet, V and Lesplingart, A. (2025). Russia-linked Pravda network cited on Wikipedia, LLMs, and X. *The Digital Forensic Research Lab (DFRLab)*. Via link: https://dfrlab.org/2025/03/12/pravda-network-wikipedia-llm-x/?utm_source=chatgpt.com

[14] Ibid.

საჯარო ვებარქივი Common Crawl-ი, რომელიც კვებავს მრავალი AI მოდელის სასწავლო ბაზას. კვლევამ დაადგინა, რომ ინგლისურენოვანი Pravda-ს კონტენტის დაარქივებული გვერდების რაოდენობა, 2024 წლის ნოემბერში არსებული 37-დან, 2025 წლის ნოემბრისთვის დაახლოებით 40 000-მდე გაიზარდა. კონტამინაციის რეალურობის შესამოწმებლად DFRLab-მა გამოიყენა Meta-ს ღია მოდელი Llama 3.1 405B Base და ტექსტის დასრულების მეთოდი: მოდელს აწვდიდნენ ცნობილი სტატიის საწყის წინადადებას და ამოწმებდნენ, აღადგენდა, თუ - არა დანარჩენ ტექსტს სიტყვასიტყვით. ტესტმა აჩვენა, რომ RT-ის მიერ გავრცელებული ცრუ ნარატივი უკრაინაში აშშ-ის „ბიოლაბორატორიების“ შესახებ მოდელმა თითქმის სიტყვასიტყვით აღადგინა, რაც იმის ნიშანია, რომ ეს კონკრეტული ტექსტი შესულია მოდელის სასწავლო მონაცემებში.[15]

ეს მიგნებები თანხვედრილია Anthropic-ის, დიდი ბრიტანეთის AI Security Institute-ისა და Alan Turing Institute-ის 2025 წლის ერთობლივ კვლევასთან, რომლის თანახმადაც, სასწავლო მონაცემებში სულ რაღაც 250 მავნე დოკუმენტის დამატება საკმარისია იმისთვის, რომ დაზიანდეს თუნდაც ძალიან დიდი, 13 მილიარდ პარამეტრზე მეტი მოცულობის მოდელის პასუხები.[16] მას შემდეგ, რაც ეს მავნე შინაარსი მოდელის სისტემებში ინტეგრირდება, ერთადერთი გამოსავალი მოდელის სრული ხელახალი გადამზადებაა, რაც ძალიან შრომატევადი და ძვირადღირებული პროცესია.

აქვე უნდა აღინიშნოს, რომ ზემოთ მოცემული კვლევების ინტერპრეტაციები არ წარმოადგენს საყოველთაო თანხმობის საგანს. კერძოდ, 2025 წელს Harvard Kennedy School Misinformation Review-ში გამოქვეყნებულმა კვლევამ (ალიუკოვი და სხვები)[17] ეს ვარაუდი ოთხი პოპულარული ჩატბოტის (ChatGPT-4o, Gemini 2.5 Flash, Copilot, Grok-2) მაგალითზე გატესტა და დაასკვნა, რომ LLM-გრუმინგის თეორიის სასარგებლოდ ცოტა მტკიცებულება მოიძებნა. ავტორების არგუმენტით, როცა მოდელი მოიხმობს Pravda-ს მსგავს წყაროებს, ამის მიზეზი შესაძლოა იყოს არა მიზანმიმართული „მოწამვლა“, არამედ ჩვეულებრივი მონაცემთა სიცარიელეები - თემები, რომლებთან დაკავშირებითაც სანდო წყაროები მცირე რაოდენობითაა და დაბალხარისხოვანი

[15] Digital Forensic Research Lab (DFRLab). (2026). *Pravda in the pipeline: Early evidence of state-adjacent propaganda in AI training data*. Via link: https://dfrlab.org/2026/04/08/pravda-in-the-pipeline/

[16] Anthropic, the UK AI Security Institute, and the Alan Turing Institute. (2025). *A small number of samples can poison LLMs of any size*. Via link: https://www.anthropic.com/research/small-samples-poison

[17] Alyukov, M. Makhortykh, M. Voronovici, A. and Sydorova, M. (2025). LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation. *Harvard Kennedy School Misinformation Review*. Via link: https://misinforeview.hks.harvard.edu/article/llms-grooming-or-data-voids-llm-powered-chatbot-references-to-kremlin-disinformation-reflect-information-gaps-not-manipulation/

კონტენტი დომინირებს. მათი დასკვნით, პრობლემის ძირი შესაძლოა იყოს არა უცხოური მანიპულაცია, არამედ ინტერნეტში ხარისხიანი ინფორმაციის არათანაბარი გადანაწილება, რაც, პრაქტიკული თვალსაზრისით, პრობლემის სერიოზულობას არ ამცირებს, თუმცა განსხვავებულ პასუხს მოითხოვს: არა მხოლოდ „მტრის“ დაბლოკვას, არამედ სანდო კონტენტის დეფიციტის შევსებას. ეს კამათი მნიშვნელოვანია მეთოდოლოგიური თვალსაზრისითაც. კერძოდ, როგორც გოლებიევსკი და ბოიდი (2019) აღნიშნავდნენ,[18] მონაცემთა სიცარიელეების იდენტიფიცირება რთულია და ისინი ხშირად შეუმჩნეველი რჩება, ვიდრე რაიმე მოვლენა არ გამოიწვევს მოთხოვნის მკვეთრ ზრდას კონკრეტულ თემაზე. შესაბამისად, ორი ახსნა - განზრახ „გრუმინგი“ თუ სტრუქტურული სიცარიელე - ურთიერთგამომრიცხავი კი არ არის, არამედ, სავარაუდოდ, პარალელურად მოქმედი მექანიზმებია: მასობრივი კონტენტის წარმოება ერთდროულად ავსებს სიცარიელეს და ცდილობს მის შევსებას მიზანმიმართული ნარატივით. ამდენად, საფრთხე მაინც არსებობს, რადგან მასობრივი, იაფი, ხშირად ავტომატურად გენერირებული კონტენტი სწრაფად ავსებს საინფორმაციო გარემოში არსებულ ვაკუუმს, ხოლო ხარისხიანი, სანდო ჟურნალისტიკა თუ კვლევა - ნელა და ძვირად.

***

გენერაციული ხელოვნური ინტელექტის განვითარებას საინფორმაციო გავლენის ოპერაციები სრულიად ახალ დონეზე აჰყავს. თუ აქამდე ტრადიციული პროპაგანდის საბოლოო სამიზნე ადამიანის ცნობიერება იყო, თანამედროვე საინფორმაციო გარემოში გავლენის შუალედურ სამიზნედ სულ უფრო მეტად იქცევა თავად ალგორითმი, სისტემა, რომელიც ინფორმაციას აგროვებს, ახარისხებს, სწავლობს და შემდეგ მილიონობით მომხმარებელს აწვდის. შესაბამისად, პროპაგანდის ეფექტიანობა შეიძლება განისაზღვროს არა მხოლოდ იმით, რამდენ ადამიანამდე მიაღწია კონკრეტულმა გზავნილმა, არამედ იმითაც, შეძლო, თუ - არა მან გავლენის მოხდენა ინფორმაციის შუამავალი სისტემების მიერ რეალობის აღქმაზე. სხვა სიტყვებით რომ ვთქვათ, ასეთი საინფორმაციო ოპერაციები ორიენტირებულია სტატისტიკურ დომინირებაზე საძიებო თუ სასწავლო კორპუსებში. მათი მიზანია არა ის, რომ კონკრეტულ მომენტში ესა თუ ის ადამიანი დაარწმუნონ, არამედ ის, რომ მომავალში, როდესაც ვინმე კითხვას დასვამს, სისტემამ

[18] Golebiewski, M. and Boyd, D. (2019). Data Voids: Where Missing Data Can Easily Be Exploited. *Data & Society*. Via link: https://apo.org.au/sites/default/files/resource-files/2019-10/apo-nid265631.pdf

მანიპულატორს მყისიერად შესთავაზოს მისთვის სასურველი პასუხი, როგორც „ყველაზე ხელმისაწვდომი“ ინფორმაცია.

სტატიაში აღწერილი მიდგომა რამდენიმე პრაქტიკულ შედეგს იძლევა. პირველი, ესაა მასშტაბი და რესურსების ეკონომია - თუ ტროლების ან ბოტების ფერმას სჭირდებოდა ათასობით ანგარიში ადამიანების დასარწმუნებლად, კონტენტის მასობრივი გენერაცია, ერთხელ დაწერილი (ან ავტომატურად გენერირებული) ტექსტის უსასრულო რეპროდუცირების გზით, ცდილობს, იმავე შედეგს გაცილებით იაფად და მცირე დროში მიაღწიოს. ამის დასტურია რუსული Pravda-ს დეზინფორმაციული ქსელის მასშტაბი - წელიწადში მილიონობით სტატია, მინიმალური ორგანული აუდიტორიის პირობებში. გარდა ამისა, მნიშვნელოვანია ისიც, რომ ხელოვნური ინტელექტის სასწავლო მონაცემებში მოხვედრილი ნარატივი შესაძლოა მოდელის ქცევაში თვეების ან წლების განმავლობაში „ჩაიმარხოს“, ბევრად უფრო ხანგრძლივი ეფექტით, ვიდრე სოციალური მედიის ერთჯერადი კამპანია. ასევე დგას გამჭვირვალობის საკითხიც. კერძოდ, სოციალურ მედიაში მკვლევრებს შეუძლიათ, გარკვეულწილად, თვალი მიადევნონ ანგარიშების ქსელს, ხოლო AI-ს პასუხის ჩამოყალიბების პროცესი (ის, თუ რომელმა წყარომ რამდენად შეწონილად იმოქმედა კონკრეტულ პასუხზე) ხშირად მოდელის დეველოპერისთვისაც კი სრულიად დახურულია. საბოლოოდ, ხელოვნური ინტელექტის ერაში საინფორმაციო მანიპულაციის პრობლემა სცდება უშუალოდ ადამიანზე ზემოქმედების კარგად ნაცნობ, მაგრამ ხშირ შემთხვევაში გადაუჭრელ პრობლემას და სრულიად ახალ გამოწვევებს გვთავაზობს.