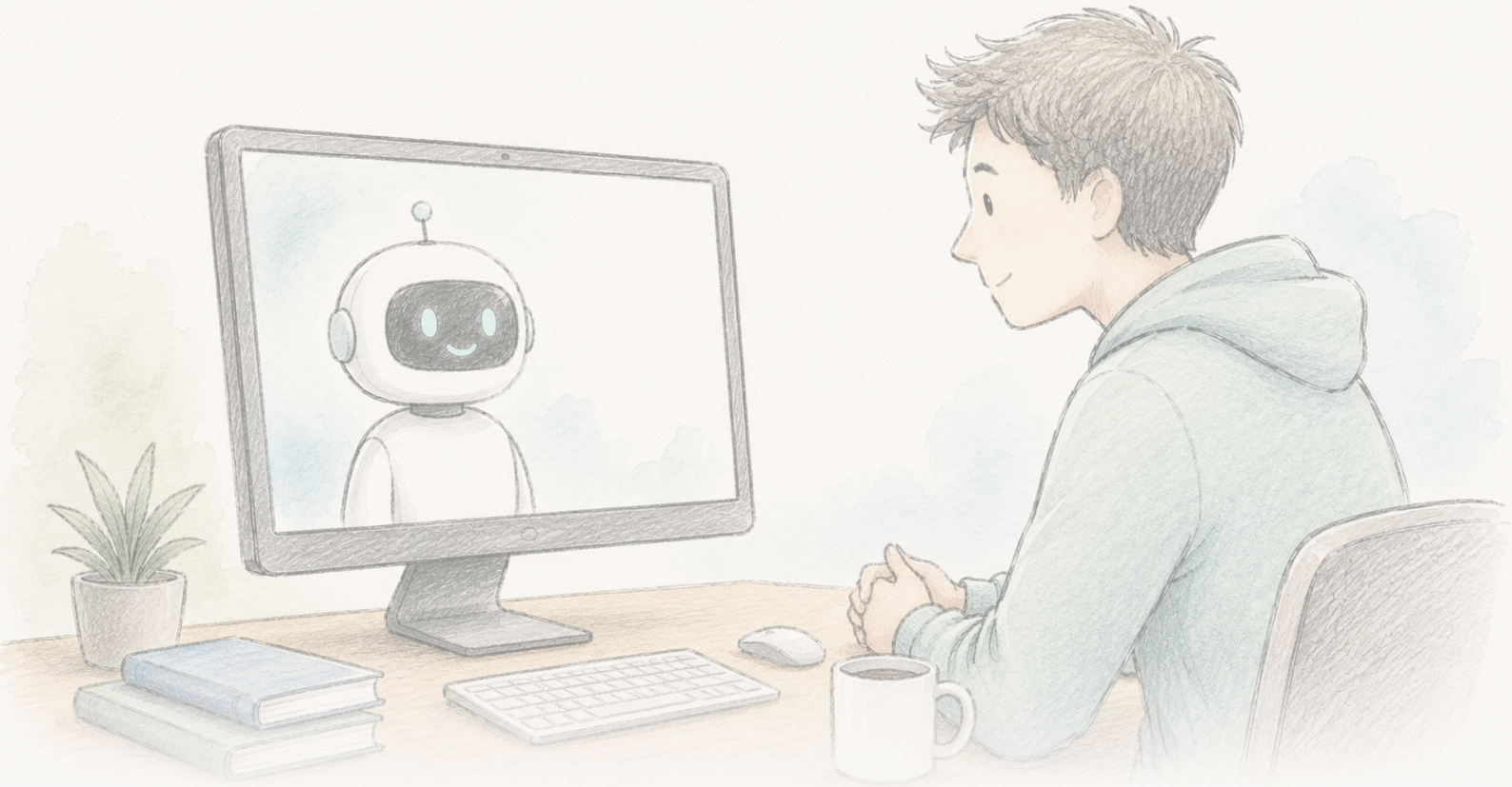


The CLEAN PAPER

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

አንድ chatbot የሴራ እምነትን ለወራት የቀነሰ ይመስላል

በ2024 የታተመ ጥናት ከGPT-4 Turbo ጋር የተደረገ ሦስት-ዙር ውይይት ሰዎች በሚያምኑበት የሴራ ንድፈ ሐሳብ ላይ ያላቸውን እምነት በግምት 20% እንደቀነሰ ዘግቧል፤ ውጤቱ ለሁለት ወራት ቆይቷል፣ በተለያዩ የሴራ ዓይነቶች ላይ ታይቷል፣ እና AIው ያቀረባቸው እውነታዊ አባባሎች ከሞላ ጎደል ሁሉ ትክክለኛ ተብለው ተገምግመዋል። በሰኔ 2026 ግን ጽሑፉ በውሂብ አያያዝና በድጋሚ ማመንጨት ችግኝቶች ምክንያት በEditorial Expression of Concern ሥር ገባ። ደራሲዎቹ የተስተካከለው ትንተና ውጤቱን በአቅጣጫ፣ በስታቲስቲካዊ ጠቀሜታ እና በመጠን እንደሚጠብቅ ይናገራሉ፣ Science ግን ገና እየገመገመ ነው። እውነተኛ አስደሳችና በጥንቃቄ የተገነባ ውጤት ነው — አሁን በግምገማ ሥር ነው፣ የተወሰነም አይደለም፣ debunkedም አይደለም።

Written by Lucio Vaglio · Cross-checked by Cinzia Vaglio and Vera Rubin · Edited by Michele Renda · Figures by Nova Vela · AI-assisted, human-reviewed

Paper: *Durably reducing conspiracy beliefs through dialogues with AI*, Thomas H. Costello, Gordon Pennycook, David G. Rand, Science 385, eadq1814 (2024) · DOI: 10.1126/science.adq1814

Editorial Expression of Concern

Science የውሂብ አያያዝና የድጋሚ ማመንጨት ጥያቄዎችን እየገመገመ ባለበት ጊዜ በዚህ ጽሑፍ ላይ Editorial Expression of Concern አክሏል። ይህ retraction አይደለም እና የምርምር ጥፋት ተፈጽሟል የሚል ፍርድም አይደለም፤ ግምገማው እስኪጠናቀቅ ድረስ የተዘገበውን ውጤት ጊዜያዊ እንደሆነ ማንበብ አለብን ማለት ነው።

Editorial Expression of Concern →

ከሁሉም በኋላ እውነታዎች — እና በጽሑፉ ላይ የተቀመጠ ማስጠንቀቂያ

በ2024 አንድ ጥናት ከተለመደው አመለካከት ጋር የሚጋጭ ውጤት ዘገበ። አንድ ሰው በግሉ የሚያምንበትን የሴራ ንድፈ ሐሳብ ለመደገፍ የሚጠቅሳቸውን ማስረጃዎች በቀጥታ እንዲመልስ ከተዘጋጀ GPT-4 Turbo ጋር አጭር፣ ሦስት-ዙር ውይይት ካደረገ፣ በአማካይ በዚያ ሴራ ላይ ያለው እምነት ወደ 20% ቀነሰ። ቅነሳው ከሁለት ወራት በኋላም ነበር። ውጤቱ በጥንታዊ የሴራ ንድፈ ሐሳቦች (የJohn F. Kennedy ግድያ፣ እንግዳ ፍጥረታት፣ Illuminati) እና በወቅታዊ ጉዳዮች (COVID-19፣ የ2020 የዩናይትድ ስቴትስ ምርጫ) ላይ ተመሳሳይ ነበር፣ እንዲሁም እምነታቸው ጠንካራ እና ከማንነታቸው ጋር የተጣመረ ሰዎች ላይም ታየ። አንድ ባለሙያ እውነታ-checker ኤአይው ካቀረባቸው 128 እውነታዊ አባባሎች 127ቱን እውነት፣ አንዱን አሳሳች፣ ምንም አንዱንም ሐሰት አይደለም ብሎ ገምግሟል።

በብዛት የተሰራጨው ርዕስ ሰዎችን ከዚህ "የጥንቸል ጉድጓድ" በንግግር ማውጣት ይቻላል የሚል ነበር — የሴራ ንድፈ ሐሳብ አማኞች ማስረጃ ሊደርስባቸው የማይችል አይደሉም፣ እነሱ የሚፈልጉት ትክክለኛውን ማስረጃ በበቂ ልዩነት እንዲቀርብላቸው ነው የሚል። ይህ በእርግጥ አስደሳች አባባል ነው፣ ጥናቱም ከብዙ የማሳመን ምርምር የበለጠ በጥንቃቄ ተገንብቷል።

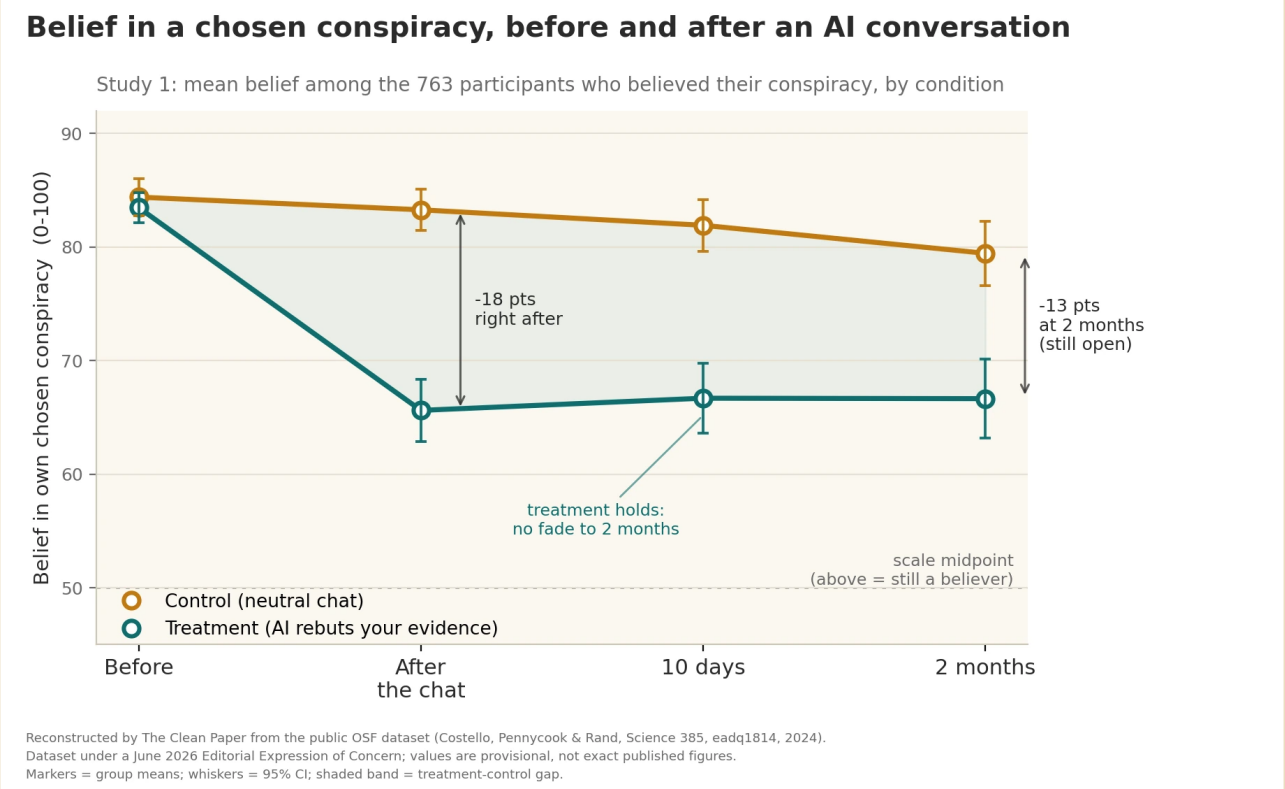
ነገር ግን በሰኔ 2026 *Science* በጽሑፉ ላይ Editorial Expression of Concern አወጣ። ይህ ብዙ እንደገና የሚተረኩ ታሪኮች የሚዘሉት ክፍል ነው፣ እና አሁን ውጤቱ ምን ያህል ክብደት ሊሸከም እንደሚችል የሚወስነውም ይህ ነው።

> **Expression of Concern ምንድን ነው — እና ምን አይደለም?**
>
> Editorial Expression of Concern አንድ ጆርናል ጥያቄዎች እንደተነሱ ነገር ግን ምርመራቸው ገና እንዳልተጠናቀቀ አንባቢዎችን ለማሳወቅ በጽሑፍ ላይ የሚያክለው መደበኛ ማስጠንቀቂያ ነው። Retraction **አይደለም** — ጽሑፉ አልተሰረዘም — እና በራሱ የምርምር ጥፋት መኖሩን የሚወስን ፍርድም አይደለም። ትርጉሙ፡- መዝገቡ ሊቀየር ስለሚችል ይህን ጽሑፍ ከዚህ ገደብ ጋር አንብቡት። እዚህ፣ ጉዳዮቹ ከደረሱባቸው በኋላ ደራሲዎቹ መርምረው ዝርዝሩን ለ*Science* ሪፖርት አድርገዋል፣ የተስተካከለ ትንተናም አቅርበዋል።

የተነሳው ጉዳይ ግልጽ ነው። በጽሑፉ የተገለጹትና በታተመው የትንተና ኮድ ውስጥ የተተገበሩት የተሳታፊ ማጣሪያ መስፈርቶች መካከል አለመጣጣም እንዳለ፣ እንዲሁም በኮድ ማዋሃድ ስህተት የሕዝብ ውሂብ ስብስቡ ውስጥ ያልተገቡ ተጨማሪ ረድፎች እንደተቀላቀሉ ደራሲዎቹ ተረዱ። እነዚህ ችግኝቶች ከተለቀቁት ውሂብና ኮድ አንዳንድ የታተሙ ቁጥሮችን እንደገና ለማመንጨት አስቸጋሪ አደረጉ። ደራሲዎቹ ለ*Science* የተስተካከለ የትንተና ሂደት እና የተዘመኑ ውጤቶች አቅርበዋል፤ እነሱ እንደሚሉት አዲሶቹ ውጤቶች ከዋናዎቹ ጋር **በአቅጣጫ፣ በስታቲስቲካዊ ጠቀሜታ እና በተግባራዊ መጠን** ይጣጣማሉ። *Science* አሁንም እየገመገማቸው ነው። ያ ግምገማ እስኪጠናቀቅ ድረስ፣ ትክክለኛዎቹ ቁጥሮች በጥያቄ ምልክት ሥር ይቆያሉ፤ ይህን ማስወገድ የሚችለው ጆርናሉ ብቻ ነው።

ስለዚህ እዚህ ሁለት ታሪኮች በአንድ ጊዜ አሉ፡- እውነታዎች ጠንካራ እምነቶችን ሊያንቀሳቅሱ ይችላሉን የሚል አስደሳች

ውጤት፣ እና የሳይንስ መዝገብ ራሱን በግልጽ መንገድ እያስተካከለ ያለ ቀጥተኛ ምሳሌ። የዚህ ጽሑፍ ዓላማ ሁለቱን እንዳንቀላቅል ነው።



በጥናቱ ውስጥ የተሳታፊውን ራሱ የጠቀሰውን ማስረጃ የሚቃወም ሦስት-ዙር የAI ውይይት፣ በተመረጠው የሴራ ንድፈ ሐሳብ ላይ ያለውን እምነት በአማካይ ወደ 20% እንደቀነሰ ተዘግቧል። ከሌላ ያልተዛመደ ርዕስ ጋር የተደረገው የቁጥጥር ውይይት እንደዚህ አላደረገም፣ እና ቅነሳው ከ10 ቀናት እና ከ2 ወራት በኋላም ሊለካ ይችል ነበር። የተዘገበው ውጤት ዘላቂ ነበር ነገር ግን ከፊል ብቻ፤ አብዛኛዎቹ ተሳታፊዎች ከውይይቱ በኋላም ሴራውን ያምኑ ነበር — ቅነሳ እንጂ ፈውስ አይደለም። ጽሑፉ በአሁኑ ጊዜ በሪፖርት ማድረጊያ አለመጣጣሞች እና በሕዝብ ውሂብ ስብስቡ ውስጥ በተገኘ ስህተት ምክንያት በEditorial Expression of Concern (*Science*፣ ሰኔ 2026) ሥር ነው። ደራሲዎቹ የተስተካከለው ትንተና የውጤቱን አቅጣጫ፣ ስታቲስቲካዊ ጠቀሜታ እና መጠን እንደሚጠብቅ ይናገራሉ፣ *Science* ግን ገና እየገመገመ ነው። ይህ ግራፍ The Clean ጽሑፍ ከሕዝብ ውሂብ ስብስቡ እንደገና የገነባው ስለሆነ፣ የሚታዩት እሴቶች ጊዜያዊ ናቸው፣ የጽሑፉ የመጨረሻ ቁጥሮች አይደሉም።

Original figure — The Clean Paper CC BY 4.0

ደራሲዎቹ ምን አደረጉ?

- እያንዳንዳቸው የሚያምኑበትን የሴራ ንድፈ ሐሳብ እና የሚደግፈው ማስረጃ ምን እንደሆነ በራሳቸው ቃላት እንዲጽፉ የተጠየቁ 2190 ተሳታፊዎችን ያካተቱ ሁለት ሙከራዎችን አካሄዱ።
- እያንዳንዱ ተሳታፊ ከGPT-4 Turbo ጋር ሦስት-ዙር ውይይት አደረገ። በ**ሕክምና** ሁኔታ AI ሥርዓቱ ተሳታፊው የጠቀሰውን ልዩ ማስረጃ እንዲቃወም ተዘጋጀ፤ በ**ቁጥጥር** ሁኔታ ግን ስለሌላ ያልተዛመደ ርዕስ ተወያዩ።
- በተመረጠው ሴራ ላይ ያለውን እምነት ከውይይቱ በፊትና ወዲያውኑ በኋላ ለኩ፣ ከዚያም ተሳታፊዎቹን **ከ10 ቀናት እና ከ2 ወራት** በኋላ እንደገና አነጋገሩ።
- በአንድ ናሙና ውስጥ AI ሥርዓቱ ያቀረባቸውን 128 እውነታዊ አባባሎች ሁሉ አንድ ባለሙያ እውነታ-checker እንዲገመግም አደረጉ።

- ውጤቱ ወደ ሌሎች ያልተዛመዱ የሴራ ንድፈ ሐሳቦች ይሻገራልን ፈተኑ፤ በተጨማሪም እንደ ልዩነት ፈተና፣ በእውነት **እውነት የሆኑ** ሴራዎች ላይ ያለውን እምነትም ያሳንሳልን ብለው ተመለከቱ።

ምን አገኙ?

- በሕክምና ቡድን ውስጥ በተመረጠው ሴራ ላይ ያለው እምነት ከቁጥጥር ቡድን ጋር ሲነጻጸር **በአማካይ ወደ 20% ቀነሰ** (ጥናት 1፡- በ100-ነጥብ መለኪያ ላይ 95% የመተማመኛ ክልል [13.8, 19.7] ነጥቦች፣ $P < 0.001$)።
- **ውጤቱ ቆይቷል**። በሕክምና ቡድን ቅነሳው አልጠፋም፤ እምነቱ ከ10 ቀናት እና ከሁለት ወራት በኋላም ከውይይቱ ወዲያውኑ ከነበረው ደረጃ ጋር ተቀራራቢ ቆየ፣ ቁጥጥር ቡድኑ ግን ከፍ ብሎ ቆየ። ጽሑፉ በተመረጠው ሴራ ላይ ያለው ውጤት ቢያንስ ለሁለት ወራት ሳይደክም እንደቆየ ይገልጻል፣ የአፍታ መንቀጥቀጥ እንዳልነበረ።
- **በተለያዩ የሴራ ዓይነቶች ላይ ተስፋፋ**፣ እና ከመጀመሪያ ጠንካራ እምነት ባላቸው ተሳታፊዎች ላይም ታየ።
- **በዚህ ሁኔታ AI ሥርዓቱ ያቀረባቸው አባባሎች ትክክለኛ ነበሩ**፡- ከ128 ውስጥ 127 (99.2%) እውነት፣ 1 (0.8%) አሳሳች፣ ምንም ሐሰት የለም።
- **ውጤቱ የተወሰነ ነበር**፣ በሁሉም ነገር ላይ ጥርጣሬ አላሳደገም፤ ጣልቃ ገብነቱ **እውነት በሆኑ ሴራዎች** ላይ ያለውን እምነት አላሳነሰም። ይህም ሰዎችን በሁሉም ነገር እንዲጠራጠሩ ከማድረግ ይልቅ ያልተደገፉ እምነቶችን እንዳንቀሳቀሰ ይጠቁማል።
- **ትንሽ ወደ ሌሎች እምነቶችም ተሸጋገረ** — በሌሎች ያልተዛመዱ ሴራዎች ላይ ያለው እምነት ወደ 12% ቀነሰ፣ ተሳታፊዎችም የሴራ አባባሎችን ለመቃወም ያላቸው ፍላጎት እንደጨመረ ዘገቡ። ዋናው ውጤት በጥናት 2 ውስጥም ታየ፣ ቁጥጥር ቡድን በሌለው አነስተኛ ናሙና ውስጥም በተመሳሳይ አቅጣጫ ነበር — ይህ የበለጠ ደካማ ማስረጃ ነው፣ ግን ከዋናው ውጤት ጋር ይጣጣማል።

ይህ ምን አያረጋግጥም?

- ጥናቱ አሁን **ያለ ጥያቄ የተቀበለ ውጤት አይደለም**። ጽሑፉ በውሂብ አያያዝና በድጋሚ ማመንጨት ችግኝቶች ምክንያት በEditorial Expression of Concern ሥር ነው፣ የተስተካከሉትን ቁጥሮችም *Science* ገና እየገመገመ ነው። ያ ግምገማ እስኪጠናቀቅ ድረስ የተወሰኑትን እሴቶች ጊዜያዊ እንደሆኑ መያዝ ይገባል።
- AI የሴራ እምነትን **“ይፈውሳል”** አያሳይም። በአማካይ ~20% ቅነሳ ትርጉም ያለው ለውጥ ነው፣ ግን እምነቱን አያጠፋውም — አብዛኛዎቹ ተሳታፊዎች ከዚያ በኋላም ሴራውን ያምኑ ነበር፣ በትንሹ ብቻ።
- ይህ **በእውነተኛው ዓለም የተተገበረ ውጤት አይደለም**። ተሳታፊዎቹ ለጥናቱ በፈቃዳቸው ገብተው ከAI ሥርዓቱ ጋር በቅንነት ተወያዩ። በእውነተኛ ሕይወት፣ በሴራ እምነት እጅግ የጸኑ ሰዎች ከእነሱ ጋር ለመከራከር የተዘጋጀ chatbot ለመፈለግ እጅግ አነስተኛ ዕድል አላቸው።
- የ99.2% ትክክለኛነት **ለዚህ ሥራ፣ ለዚህ ሞዴል እና ለዚህ ጥንቃቄ ያለው** prompting **ብቻ የሚመለከት** ነው። የቋንቋ ሞዴሎች በአጠቃላይ ሁልጊዜ እውነት ይናገራሉ የሚል ዋስትና አይደለም። ደራሲዎቹ ይህ ግላዊ የማሳመን ኃይል ወደ ሐሰት እምነት ከተመራ በተመሳሳይ ሁኔታ ሐሰትን ሊያጠናክር እንደሚችል በግልጽ ይጠቅሳሉ።
- “ዘላቂ” የሚለው እዚህ **ሁለት ወራት** ማለት ነው። ለአንድ ውይይት አስደናቂ ነው፣ ግን ቋሚ ማለት አይደለም።

ማስረጃው ምን ያህል ጠንካራ ነው?

- **የጥናቱ ንድፍ እውነተኛ ጥንካሬ ነው።** የሕክምና እና ቁጥጥር ሙከራዎች፣ ንጹህ ፕላሴቦ-ዓይነት ቁጥጥር፣ በሁለት ጥናቶችና በነፃ ናሙና ውስጥ የተደገመ ውጤት፣ የዘላቂነት ክትትል፣ እና AI ሥርዓቱ ያቀረባቸውን አባባሎች በተለየ ሁኔታ የሚፈትን የትክክለኛነት ምርመራ አሉ። ይህ ከብዙ የማሳመን ምርምር የበለጠ ጥንቃቄ ያለው ነው፣ እና ውጤቱ በተፈተነባቸው ቦታዎች ሁሉ ተመሳሳይ አቅጣጫ አለው።
- **ግን Expression of Concern የግርጌ ማስታወሻ አይደለም።** ከተለቀቁት ውሂብና ኮድ የታተሙትን ቁጥሮች እንደገና ማመንጨት መቻል የአንድ ውጤት እምነት አካል ነው፣ እና በትክክል የተሰበረውም ይህ ነው — በscreening criteria አለመጣጣም እና በኮድ ማዋሃድ ስህተት የተባዙ ረድፎች። የተስተካከለው ሂደት ውጤቱን እንደሚጠብቅ የደራሲዎቹ መግለጫ አበረታችና ምክንያታዊ ነው፣ ግን ይህ ገና የደራሲዎቹ ሪፖርት ነው፣ ነፃ ፍርድ አይደለም። መጠበቅ ያለብን የ*Science* ግምገማ ነው።
- **ትክክለኛው ሁኔታ “flagged” ነው፣ “debunked” ወይም “confirmed” አይደለም።** በጥሩ ሁኔታ የተገነባና አስገራሚ ጥናት፣ ትክክለኛዎቹ ቁጥሮቹ ግን በመደበኛ ግምገማ ሥር ናቸው። አንድ ጥሩ ታሪክ በዚህ ያልተመቸ ቦታ ላይ መተው ሊያስቸግር ይችላል፣ ግን ትክክለኛው ቦታ ይህ ነው።

ለምን ይጠቅማል?

ለዓመታት አንድ ተፅዕኖ ያለው አመለካከት የሴራ ንድፈ ሐሳብ አማኞች በስነ-ልቦና ፍላጎቶች እና በማንነት የሚነዱ ስለሆነ፣ በእውነታዎች ከእምነታቸው ማውጣት አይቻልም ይል ነበር። ይህ ዘላቂ፣ በማስረጃ ላይ የተመሠረተ ቅነሳ ከግምገማ ካለፈ፣ ለዚያ ተስፋ ቢስ አመለካከት አስፈላጊ ተቃራኒ ማስረጃ ይሆናል። ችግኙ አንድ ክፍል አጠቃላይ debunking አንድ አማኝ በእውነት የሚጠቅሰውን የተወሰነ ማስረጃ እንዳልያዘ ሊሆን እንደሚችል ያመለክታል፤ LLM ይህን በትልቅ መጠን ሊያደርግ ይችላል።

ይህ ሳይንስ እንዴት መስራት እንዳለበት የሚያሳይ ምሳሌም ነው። የExpression of Concern ትምህርት ታዋቂ ውጤቶች ዋጋ የላቸውም ማለት አይደለም፤ ስህተቶች ወደ ፊት ይወጣሉ፣ ውሂብ እንደገና ይመረመራል፣ አባባሎችም በሕዝብ ፊት እንደገና ይፈተናሉ። ይህ ሂደት መስራቱ የሳይንስ ጥቅም ነው፣ ቅሌት አይደለም። ስለዚህ ለዚህ ውጤት ትክክለኛው አቋም ከልክ ያለፈ ደስታ ወይም በፍጥነት መጣል ሳይሆን ትዕግሥት ነው።

AIን ሲመለከትም ይህ በሁለቱም አቅጣጫ ይሰራል። በግል የተበጀ ክርክር ሐሰት እምነትን ሊያሳንስ የሚችለውን ያህል፣ አንድን ሐሰት እምነት በተመሳሳይ ኃይል ሊያሳድግም ይችላል። ቴክኒኩ በራሱ ገለልተኛ ነው፤ ዓላማው ግን አይደለም።

ንጹህ ማጠቃለያ

በ2024 የታተመ ጥናት ከGPT-4 Turbo ጋር የተደረገ ሦስት-ዙር ውይይት ሰዎች በሚያምኑበት የሴራ ንድፈ ሐሳብ ላይ ያላቸውን እምነት ወደ 20% እንደቀነሰ፣ ውጤቱም ለሁለት ወራት እንደቆየ እና በተለያዩ የሴራ ዓይነቶች ላይ እንደታየ ዘግቧል። AI ሥርዓቱ ያቀረባቸው እውነታዊ አባባሎችም ከሞላ ጎደል ሁሉ ትክክለኛ ተብለው ተገምግመዋል። በሰኔ 2026 ግን በውሂብ አያያዝና በድጋሚ ማመንጨት ችግኝቶች ምክንያት ጽሑፉ በEditorial Expression of Concern ሥር ገባ። ደራሲዎቹ የተስተካከለው ትንተና የውጤቱን አቅጣጫ፣ ስታቲስቲካዊ ጠቀሜታ እና መጠን እንደሚጠብቅ ይናገራሉ፣ *Science* ግን ገና እየገመገመ ነው። ይህን እንደ እውነተኛ አስደሳችና በጥንቃቄ የተገነባ፣ ግን አሁን በግምገማ ሥር ያለ ውጤት ማንበብ ይገባል — የተወሰነ አይደለም፣ debunked አይደለም። ስለዚህ በጣም ትክክለኛው አረፍተ ነገር ሂደቱ ራሱ እየጻፈው ያለው ነው፡- እንደገና ፈትኑት።

ምንጮች

Costello, Pennycook & Rand, Durably reducing conspiracy beliefs through dialogues with AI, Science 385, eadq1814 (2024)
https://doi.org/10.1126/science.adq1814

Editorial Expression of Concern, Science (posted 11 June 2026; H. Holden Thorp, Editor-in-Chief), DOI 10.1126/science.aej2383
https://www.science.org/doi/10.1126/science.aej2383