De epistemische lus

Martijn van Grieken

augustus 2026

CONTENTS

Wanneer maakt AI onze kennis beter, en wanneer smaller? Een beslisraamwerk voor organisaties die taalmodellen inzetten in kenniswerk.

Managementsamenvatting

Taalmodellen zijn in twee jaar tijd opgeschoven van handig zoekhulpmiddel naar medeauteur van ons kenniswerk. Ze vatten samen, leggen uit, adviseren, schrijven mee aan beleid, code en analyses. Daarmee is iets fundamenteels veranderd: systemen die getraind zijn op onze kennis, produceren nu zelf kennis die terugstroomt in documenten, besluiten en databronnen, en dus in de omgeving waaruit de volgende antwoorden komen. Mens en model vormen een terugkoppelingslus.

Dat die lus bestaat, is inmiddels stevig wetenschappelijk aangetoond. Wat níet vaststaat, is welke kant hij op draait. Het onderzoek laat beide richtingen zien: AI die collectieve output eenvormiger maakt, én AI die juist corrigeert en het blikveld verbreedt. De interessante vraag is daarom niet óf AI onze kennis beïnvloedt, maar onder welke voorwaarden dat goed of slecht uitpakt, en aan welke knoppen een organisatie kan draaien.

Deze whitepaper vertaalt ons lopende wetenschappelijke onderzoek naar een praktisch raamwerk. De kern: beoordeel AI in kenniswerk niet op één as (“divers of eenvormig”), maar op twee, kwaliteit én diversiteit, met vier mogelijke uitkomsten. Ken de twee mechanismen die de richting bepalen: personalisatie en plausibiliteit. En besef dat de risico’s meetbaar zijn als eigenschappen van het systeem dat u uitrolt. Wie standaardiseert op één model, kiest, meestal onbewust, een kennisprofiel voor de hele organisatie. Dat kan een verstandige keuze zijn. Maar dan wel een bewuste. Juist in de omgevingen waarin Twentynext dagelijks werkt, publieke uitvoering, veiligheid en gereguleerde zorg, is dit geen academische kwestie: daar bepalen de kwaliteit én de spreiding van kennis direct de kwaliteit van besluiten over burgers, risico’s en patiënten.

---

1. AI is geen naslagwerk meer, maar een medeauteur

Een zoekmachine wijst u de weg naar bronnen en laat het denkwerk aan u over. Een taalmodel doet het denkwerk deels zelf: het componeert een antwoord, toegesneden op uw vraag en steeds vaker op u als vragensteller. Dat voelt als vooruitgang, en vaak is het dat ook. Maar het verandert wel de rol van de technologie in onze kennis.

Waar dat antwoord uit voortkomt, is namelijk niet de werkelijkheid zelf, maar een gelaagde registratie ervan. Niet alles wat gebeurt wordt waargenomen; niet alles wat wordt waargenomen wordt opgeschreven; niet alles wat is opgeschreven is gedigitaliseerd; en niet alles wat digitaal bestaat, belandt in trainingsdata. Elke overgang selecteert en kleurt (Bender et al., 2021). Op zichzelf is dat niets nieuws: elke technologie die kennis draagt, kleurt haar, van kleitablet tot krant.

Nieuw is de terugkoppeling. Wat een model genereert, verdwijnt niet in het luchtledige. Het wordt overgenomen in rapporten, beleidsstukken, code, lesmateriaal en webteksten, en stroomt zo terug in de omgeving waaruit volgende antwoorden worden opgebouwd. Dat gebeurt langs meerdere routes en op heel verschillende snelheden: zoek- en retrievalsystemen pikken nieuwe documenten binnen dagen op, organisatie-archieven en publicaties vormen wat mensen lezen, en, het traagst, toekomstige modellen trainen op het veranderde internet. Zelfs een model waarvan de parameters nooit wijzigen, kan dus middenin een lus zitten, zodra het materiaal dat het ophaalt steeds vaker materiaal is dat zijn eigen output mede heeft gevormd.

Dit is geen speculatie meer. Experimenteel onderzoek met ruim veertienhonderd deelnemers liet zien dat vertekeningen in AI-oordelen door menselijke gebruikers worden overgenomen en in de interactie verder worden versterkt, terwijl accurate systemen menselijke oordelen juist verbéterden (Glickman & Sharot, 2025). Onderzoekers spreken inmiddels van mens-AI-co-evolutie als eigen onderzoeksveld (Pedreschi et al., 2025), en van “lock-in”: modellen die menselijke overtuigingen leren, terugkaatsen en opnieuw opnemen (Qiu et al., 2025). De lus is er. De vraag is wat hij met ons doet.

2. “AI maakt ons allemaal hetzelfde”: waarom dat te simpel is

Het bekendste onderzoeksresultaat in dit veld klinkt als een waarschuwing. Schrijvers die AI-suggesties kregen, leverden individueel betere verhalen af, maar de verhalen gingen onderling méér op elkaar lijken (Doshi & Hauser, 2024). Hetzelfde patroon dook op bij ideeëngeneratie (Anderson, Shah, & Kreminski, 2024), en een recente synthese in *Trends in Cognitive Sciences* waarschuwt dat breed gedeeld gebruik van dezelfde modellen taal, perspectief en redeneerstijl kan standaardiseren (Sourati, Ziabari, & Dehghani, 2026). Individuele winst, collectieve verschraling: dat is een reëel en gedocumenteerd risico.

En toch is het maar de helft van het verhaal. In een groot dynamisch experiment leidde juist méér blootstelling aan AI-ideeën tot een grótere collectieve diversiteit van ideeën (Ashkinaze et al., 2025). Toen onderzoekers het beroemde schrijversexperiment herhaalden met tien bewust verschillende AI-persona's, bleef de diversiteit van de verhalen op het niveau van een groep zonder AI (Wan & Kalman, 2026). En een subtiel maar belangrijk detail: het diversiteitsverlies bij co-schrijven trad wél op met een instructie-getuned model, maar niet met het onderliggende basismodel (Padmakumar & He, 2024). Anders gezegd: het effect zit niet in "AI" als zodanig, maar in ontwerp- en configuratiekeuzes.

Wie deze resultaten naast elkaar legt, komt tot een ongemakkelijke maar bevrijdende conclusie. De vraag "homogeniseert AI onze kennis?" heeft geen ja-of-nee-antwoord. De richting is conditioneel: en dat verschuift het gesprek van angst naar knoppen. Welke voorwaarden duwen een organisatie de goede kant op, en welke de verkeerde?

3. Vier uitkomsten in plaats van één schrikbeeld

Om die vraag scherp te krijgen, moet u eerst twee dingen uit elkaar halen die in het publieke debat voortdurend door elkaar lopen: de kwáliteit van kennis en de divérsiteit ervan. Convergentie is niet per definitie slecht: een team dat na goed bewijs stopt met een achterhaalde aanpak, convergeert én verbetert. En diversiteit is niet per definitie gezond: tien onderbouwde scenario's zijn iets anders dan tien varianten van dezelfde vergissing. De wetenschapsfilosofie wist dit al langer: gemeenschappen die te snel consensus bereiken, kunnen op het verkeerde antwoord landen, maar eeuwig volgehouden verdeeldheid voorkomt óók dat de waarheid ooit wint (Zollman, 2010).

Kruist u beide dimensies, dan ontstaan vier mogelijke uitkomsten van AI in kenniswerk:

	Diversiteit daalt	**Diversiteit stijgt**
Kwaliteit stijgt	Terechte consensus	Gefundeerd pluralisme
Kwaliteit daalt	Collectieve vergissing (lock-in)	Versplinterde ruis

Terechte consensus: iedereen komt uit bij het antwoord dat het bewijs ook echt draagt. *Gefundeerd pluralisme*: meerdere verdedigbare lijnen blijven naast elkaar bestaan, precies wat u wilt waar het bewijs nog niet beslist. *Collectieve vergissing*:

de gevaarlijkste uitkomst, omdat hij als kwaliteit vóelt: iedereen dezelfde plausibele, goed geformuleerde, onjuiste aanname. En *versplinterde ruis*: veel variatie, weinig onderbouwing; de uitkomst die diversiteitsstatistieken ten onrechte als gezondheid registreren.

De praktische les van dit kwadrant is simpel maar streng: wie alleen diversiteit meet, kan terechte consensus niet onderscheiden van collectieve vergissing. En wie alleen accuratesse per antwoord meet, ziet niet wat het systeem over tijd met de spréiding van kennis in de organisatie doet. U hebt beide assen nodig.

4. Twee mechanismen die bestuurders moeten kennen

Personalisatie: individueel vaster, collectief verder uiteen. Moderne assistenten passen zich aan hun gebruiker aan, via geheugenfuncties, en via een neiging die uit de trainingsmethode zelf voortkomt: modellen die geoptimaliseerd zijn op menselijke waardering, leren soms liever te bevestigen dan te corrigeren (Sharma et al., 2024). Het opmerkelijke is dat dit mechanisme op twee niveaus tegelijk in tegengestelde richting kan werken. Elke individuele gebruiker wordt stabieler in de eigen opvatting: het systeem beweegt immers mee, terwijl verschillende gebruikers, elk bevestigd in verschillende uitgangspunten, juist verder uit elkaar drijven. De consequentie voor organisaties: u kunt collectieve effecten niet aflezen aan individuele tevredenheid. Een assistent die iedereen prettig vindt, kan precies daardoor de gedeelde werkelijkheid van een team eroderen. Anti-sycofantie is daarmee geen kwestie van individuele waarheidsliefde alleen, maar collectief beleid.

Plausibiliteit: de volumeknop van de lus. Uit decennia psychologisch onderzoek weten we dat vloeiendheid en herhaling het gevóel van waarheid verhogen, los van de feitelijke juistheid (Hasher, Goldstein, & Toppino, 1977; Fazio et al., 2015). Bij mensen is vloeiend en samenhangend spreken nog een zwak signaal van kunde: het kost moeite. Bij taalmodellen kost het niets: ze formuleren even overtuigend wanneer ze gelijk hebben als wanneer ze ernaast zitten. Recente analyses benoemen precies dit als het kernrisico: talige plausibiliteit die de plaats inneemt van echte toetsing (Quattrociocchi, Capraro, & Perc, 2025), en systemen die “eerlijke non-signalen” afgeven, behulpzaamheid en welsprekendheid die bij mensen betekenis dragen, maar bij machines niets meer bewijzen (Maynard, 2026). Ons onderzoek voegt daar de lus-dimensie aan toe: in een terugkoppelingssysteem is

overtuigingskracht geen eenmalig verkoopargument, maar een versterkingsfactor. Elk overtuigend antwoord dat wordt overgenomen, verandert de omgeving waaruit het volgende antwoord komt.

5. Van zorg naar meetlat

Hier wordt het voor de praktijk interessant. In ons lopende onderzoeksprogramma herleiden we deze mechanismen tot een klein aantal meetbare eigenschappen van een uitgerold systeem: geen filosofie, maar auditgrootheden:

1. **Richtingsvoorkeur**: heeft het model, los van de aangeboden informatie, een eigen neiging richting bepaalde antwoorden?
2. **Frequentiegevoeligheid**: volgt het model wat in zijn informatieomgeving het váákst wordt beweerd, of wat het bést is onderbouwd? Dit is de eigenschap waarlangs een veelherhaalde onjuistheid dominant kan worden.
3. **Meegaandheid**: hoe sterk buigt de output mee met de opvatting van de individuele gebruiker?
4. **Bewijsgevoeligheid**: in hoeverre dempt het aanbieden van gevalideerde, diagnostische bronnen de eerste drie eigenschappen?

Samen beantwoorden ze één praktijkvraag die elke organisatie zou moeten kunnen stellen: *wint bewijs het in onze configuratie van herhaling?* Wij meten dit met gecontroleerde, fictieve kennisdomeinen waarin we de frequentie van een bewering en de kwaliteit van het bewijs onafhankelijk van elkaar variëren, zodat het antwoord niet uit het geheugen van het model kan komen, maar zijn werkelijke gedrag toont.

Eén voorlopig theoretisch resultaat uit ons formele model verdient het om nu al gedeeld te worden, met de kanttekening dat de wiskundige verificatie nog loopt: *of* een vertekening in zo'n mens-modelsysteem kan ontsporen, blijkt in het model af te hangen van de combinatie frequentiegevoeligheid en bewijsgevoeligheid: eigenschappen van het systeem zelf. Hoe intensief er wordt vertrouwd en hergebruikt, bepaalt vervolgens hoe snél dat gaat, niet óf het kan. Als dat beeld standhoudt, is de bestuurlijke vertaling direct: het ontsporingsrisico zit in wat u inkoopt en hoe u het grondt; de gebruiksintensiteit bepaalt alleen het tempo. Huidige AI-evaluaties, die accuratesse per interactie meten, kijken daar volledig langs.

6. Wat kan een organisatie morgen doen

Vijf handelingsperspectieven volgen rechtstreeks uit het raamwerk. Ze zijn geschreven met de praktijk van uitvoeringsorganisaties, veiligheidsdiensten en zorginstellingen voor ogen, omgevingen waar honderden professionals met dezelfde assistent aan beoordelingen, analyses en dossiers werken, en waar een collectieve vergissing dus geen stijlkwestie is maar een uitvoeringsrisico.

1. **Behandel modelkeuze als kennisbeleid.** Eén model voor de hele organisatie is een keuze voor één richtingsvoorkeur, één frequentiegevoeligheid, één stijl van redeneren: het kenniswerk-equivalent van de algoritmische monocultuur die uit de beslissystemen bekend is (Kleinberg & Raghavan, 2021). Soms is dat efficiënt en verdedigbaar; voor analyse-, advies- en onderzoekswerk is bewuste heterogeniteit het overwegen waard.
2. **Zet grounding in inkoop en architectuur.** Bronverplichting, retrieval op gevalideerde bronnen en zichtbare herkomst zijn geen compliance-franje: bewijsgevoeligheid is in ons raamwerk de eigenschap die het verschil maakt tussen een systeem dat corrigeert en een systeem dat overerft.
3. **Maak vloeiendheid verdacht.** Train gebruikers dat welsprekendheid bij AI geen kwaliteitssignaal is, en bouw de reflex in om naar onzekerheid, bronnen en tegenargumenten te vragen. Dat klinkt klein; het is het menselijke deel van de volumeknop.
4. **Voer beleid op personalisatie.** Waar bevestiging een risico is, analyse, advies, toetsing, beleidsvoorbereiding, hoort meegaandheid uit en tegenspraak aan. Dat diversiteit ontwerpbaar is, is inmiddels experimenteel aangetoond (Wan & Kalman, 2026): het is een instelling, geen natuurwet.
5. **Meet distributioneel.** Vraag niet alleen “klopt dit antwoord?”, maar periodiek ook “wat doet dit systeem met de spreiding én kwaliteit van onze uitkomsten over tijd?”, de twee assen van het kwadrant, samen. Dit hoort thuis in de beheerfase van elke AI-toepassing, naast de vertrouwde monitoring op prestaties en veiligheid, niet als eenmalige toets bij livegang.

7. Het onderzoeksprogramma, en een uitnodiging

Deze whitepaper is de publieksvertaling van een wetenschappelijk working paper dat momenteel in voorbereiding is. Fase één: het karakteriseren van hedendaagse modelfamilies op de vier genoemde eigenschappen, en het doorrekenen van de lusdynamiek op basis van die metingen, voeren wij nu uit. Fase twee bestaat uit grootschalige, gepreregistreerde experimenten met menselijke deelnemers, waarin

de twee kernvoorspellingen (de personalisatie-omkering en de plausibiliteitsversterking onder terugkoppeling) confirmatoir worden getoetst. Voor die fase zoeken wij nadrukkelijk academische en institutionele partners.

Werkt u bij een universiteit of kennisinstelling en wilt u meedoen aan fase twee? Of zet uw organisatie AI grootschalig in bij kenniswerk en wilt u uw eigen configuratie langs deze meetlat leggen, van eerste audit tot inrichting en structureel beheer? Neem contact op via info@twentynext.nl.

Over de auteur

Martijn van Grieken is directeur Data & AI bij Twentynext, het Eindhovense data- en AI-consultancybureau dat sinds 2014 werkt voor onder meer publieke uitvoering, veiligheid en gereguleerde zorg. Hij werkt ruim tien jaar op het snijvlak van datagovernance, AI en publieke besluitvorming.

Over Twentynext

Twentynext is een data- en AI-consultancy uit Eindhoven, opgericht in 2014, en ondersteunt organisaties in onder meer publieke uitvoering, veiligheid en gereguleerde zorg, van detachering van specialisten tot projecten met vooraf afgesproken opleverpunten en structureel beheer.

Verantwoording

“De epistemische lus” is onze werkterm voor wat in de wetenschappelijke literatuur bekendstaat als human-AI feedback loops en mens-AI-co-evolutie; dit stuk claimt geen ontdekking van dat verschijnsel, wel een raamwerk om er richting en meetbaarheid in aan te brengen. Alle aangehaalde bronnen zijn door de auteur geverifieerd. Het genoemde theoretische stabiliteitsresultaat is voorlopig en wordt momenteel onafhankelijk gecontroleerd; de empirische metingen uit fase één worden na afronding gepubliceerd.

Bij het opstellen van dit stuk is gebruikgemaakt van AI-taalmodellen onder redactie en verantwoordelijkheid van de auteur.

Literatuur

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. *Proceedings of the 16th ACM Conference on Creativity & Cognition*, 413–425. https://doi.org/10.1145/3635636.3656204

Ashkinaze, J., Mendelsohn, J., Qiwei, L., Budak, C., & Gilbert, E. (2025). How AI ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment. *Proceedings of the ACM Collective Intelligence Conference*, 198–213. https://doi.org/10.1145/3715928.3737481

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances, 10*(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General, 144*(5), 993–1002. https://doi.org/10.1037/xge0000098

Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9*(2), 345–359. https://doi.org/10.1038/s41562-024-02077-2

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences, 118*(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

Maynard, A. D. (2026). *The AI cognitive Trojan horse: How large language models may bypass human epistemic vigilance* (arXiv:2601.07085). arXiv. https://doi.org/10.48550/arXiv.2601.07085

Padmakumar, V., & He, H. (2024). Does writing with language models reduce content diversity? *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A.-L., Dignum, F., Dignum, V., Eliassi-Rad, T., Giannotti, F., Kertész, J., Knott, A., Ioannidis, Y., Lukowicz, P., Passarella, A., Pentland, A., Shawe-Taylor, J., & Vespignani, A. (2025). Human-AI coevolution. *Artificial Intelligence, 339*, 104244. https://doi.org/10.1016/j.artint.2024.104244

Qiu, T. A., He, Z., Chugh, T., & Kleiman-Weiner, M. (2025). The lock-in hypothesis: Stagnation by algorithm. *Proceedings of the 42nd International Conference on Machine Learning*.

Quattrociocchi, W., Capraro, V., & Perc, M. (2025). *Epistemological fault lines between human and artificial intelligence* (arXiv:2512.19466). arXiv. https://doi.org/10.48550/arXiv.2512.19466

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., … Perez, E. (2024). Towards understanding sycophancy in language models. *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Sourati, Z., Ziabari, A. S., & Dehghani, M. (2026). The homogenizing effect of large language models on human expression and thought. *Trends in Cognitive Sciences*. Advance online publication. https://doi.org/10.1016/j.tics.2026.01.003

Wan, Y., & Kalman, Y. M. (2026). Diverse AI personas can mitigate the homogenization effect in human-AI collaborative ideation. *Computers in Human Behavior: Artificial Humans, 8*, 100289. https://doi.org/10.1016/j.chbah.2026.100289

Zollman, K. J. S. (2010). The epistemic benefit of transient diversity. *Erkenntnis, 72*(1), 17–35. https://doi.org/10.1007/s10670-009-9194-6