The Epistemic Loop

Martijn van Grieken

August 2026

CONTENTS

When does AI make our knowledge better, and when narrower? A decision framework for organisations deploying language models in knowledge work.

Executive summary

In the space of two years, language models have moved from handy search aid to co-author of our knowledge work. They summarise, explain, advise, and co-write policy, code and analyses. Something fundamental has changed as a result: systems trained on our knowledge now produce knowledge themselves, which flows back into documents, decisions and data sources, and thus into the environment from which the next answers are drawn. Human and model form a feedback loop.

That this loop exists is by now firmly established scientifically. What is *not* established is which way it turns. The research shows both directions: AI that makes collective output more uniform, and AI that corrects and broadens the field of view. The interesting question is therefore not *whether* AI influences our knowledge, but under which conditions that turns out well or badly, and which dials an organisation can turn.

This whitepaper translates our ongoing scientific research into a practical framework. The core: do not judge AI in knowledge work on one axis (“diverse or uniform”), but on two: quality and diversity, with four possible outcomes. Know the two mechanisms that determine the direction: personalisation and plausibility. And recognise that the risks are measurable as properties of the system you deploy. Whoever standardises on a single model chooses, usually unknowingly, a knowledge profile for the entire organisation. That may be a sensible choice. But it should be a deliberate one. Precisely in the environments where Twentynext works every day, public administration, safety and security, and regulated healthcare, this is no academic matter: there, the quality and the spread of knowledge directly determine the quality of decisions about citizens, risks and patients.

---

1. AI is no longer a reference work, but a co-author

A search engine points you to sources and leaves the thinking to you. A language model does part of the thinking itself: it composes an answer, tailored to your question and, increasingly, to you as the person asking. That feels like progress, and often it is. But it does change the role of the technology in our knowledge.

What that answer springs from is not reality itself, but a layered registration of it. Not everything that happens is observed; not everything observed is written down; not everything written down is digitised; and not everything that exists digitally ends up in training data. Every transition selects and colours (Bender et al., 2021). In itself that is nothing new: every technology that carries knowledge colours it, from clay tablet to newspaper.

What is new is the feedback. What a model generates does not vanish into thin air. It is adopted in reports, policy documents, code, teaching materials and web copy, and thus flows back into the environment from which subsequent answers are constructed. This happens along several routes and at very different speeds: search and retrieval systems pick up new documents within days, organisational archives and publications shape what people read, and, slowest of all, future models train on the changed internet. Even a model whose parameters never change can therefore sit in the middle of a loop, as soon as the material it retrieves is increasingly material its own output helped shape.

This is no longer speculation. Experimental research with over fourteen hundred participants showed that biases in AI judgements are adopted by human users and further amplified in the interaction, while accurate systems actually *improved* human judgements (Glickman & Sharot, 2025). Researchers now speak of human-AI coevolution as a research field in its own right (Pedreschi et al., 2025), and of “lock-in”: models that learn human beliefs, reflect them back and absorb them again (Qiu et al., 2025). The loop is there. The question is what it does to us.

2. “AI is making us all the same”: why that is too simple

The best-known research result in this field sounds like a warning. Writers who received AI suggestions delivered individually better stories, but the stories became more similar to one another (Doshi & Hauser, 2024). The same pattern appeared in idea generation (Anderson, Shah, & Kreminski, 2024), and a recent synthesis in *Trends in Cognitive Sciences* warns that widely shared use of the same models can standardise language, perspective and reasoning style (Sourati, Ziabari, & Dehghani, 2026). Individual gain, collective impoverishment: that is a real and documented risk.

And yet it is only half the story. In a large dynamic experiment, *more* exposure to AI ideas actually led to *greater* collective diversity of ideas (Ashkinaze et al., 2025). When researchers repeated the famous writers' experiment with ten deliberately different AI personas, the diversity of the stories remained at the level of a group without AI (Wan & Kalman, 2026). And a subtle but important detail: the diversity loss in co-writing did occur with an instruction-tuned model, but not with the underlying base model (Padmakumar & He, 2024). Put differently: the effect does not reside in "AI" as such, but in design and configuration choices.

Whoever lays these results side by side reaches an uncomfortable but liberating conclusion. The question "does AI homogenise our knowledge?" has no yes-or-no answer. The direction is conditional, and that shifts the conversation from fear to dials. Which conditions push an organisation the right way, and which the wrong way?

3. Four outcomes instead of one spectre

To sharpen that question, you first have to separate two things that constantly run together in the public debate: the *quality* of knowledge and its *diversity*. Convergence is not bad by definition: a team that abandons an outdated approach after solid evidence converges and improves. And diversity is not healthy by definition: ten well-founded scenarios are something different from ten variants of the same mistake. Philosophy of science has known this for a while: communities that reach consensus too quickly can land on the wrong answer, but eternally sustained disagreement also prevents the truth from ever winning (Zollman, 2010).

Cross both dimensions and four possible outcomes of AI in knowledge work emerge:

	Diversity falls	**Diversity rises**
Quality rises	Warranted consensus	Well-founded pluralism
Quality falls	Collective error (lock-in)	Fragmented noise

Warranted consensus: everyone arrives at the answer the evidence actually supports. *Well-founded pluralism*: several defensible lines continue to exist side by side: exactly what you want where the evidence has not yet decided. *Collective error*: the most dangerous outcome, because it *feels* like quality: everyone holding

the same plausible, well-phrased, incorrect assumption. And *fragmented noise*: much variation, little substantiation; the outcome that diversity statistics wrongly register as health.

The practical lesson of this quadrant is simple but strict: whoever measures only diversity cannot distinguish warranted consensus from collective error. And whoever measures only per-answer accuracy does not see what the system does over time to the *spread* of knowledge in the organisation. You need both axes.

4. Two mechanisms every executive should know

Personalisation: individually more fixed, collectively further apart. Modern assistants adapt to their user, through memory features, and through a tendency that arises from the training method itself: models optimised for human approval sometimes learn to prefer confirming over correcting (Sharma et al., 2024). The remarkable thing is that this mechanism can work in opposite directions on two levels at once. Each individual user becomes more stable in their own view, the system moves along with them, after all, while different users, each confirmed in different starting points, drift further apart. The consequence for organisations: you cannot read collective effects off individual satisfaction. An assistant everyone finds pleasant can, precisely because of that, erode a team's shared reality. Anti-sycophancy is therefore not merely a matter of individual truthfulness, but collective policy.

Plausibility: the volume dial of the loop. From decades of psychological research we know that fluency and repetition raise the *feeling* of truth, independent of factual correctness (Hasher, Goldstein, & Toppino, 1977; Fazio et al., 2015). In humans, speaking fluently and coherently is still a weak signal of competence: it takes effort. For language models it costs nothing: they phrase just as convincingly when they are right as when they are wrong. Recent analyses identify precisely this as the core risk: linguistic plausibility taking the place of genuine verification (Quattrociocchi, Capraro, & Perc, 2025), and systems emitting "honest non-signals": helpfulness and eloquence that carry meaning in humans but prove nothing in machines (Maynard, 2026). Our research adds the loop dimension: in a feedback system, persuasiveness is not a one-off sales argument but an amplification factor. Every convincing answer that gets adopted changes the environment from which the next answer is drawn.

5. From concern to yardstick

Here it becomes interesting for practice. In our ongoing research programme we reduce these mechanisms to a small number of measurable properties of a deployed system, not philosophy, but audit quantities:

1. **Directional preference**: does the model, independent of the information offered, have an inclination of its own towards certain answers?
2. **Frequency sensitivity**: does the model follow what is asserted most *often* in its information environment, or what is *best* substantiated? This is the property along which a much-repeated falsehood can become dominant.
3. **Compliance**: how strongly does the output bend along with the individual user's view?
4. **Evidence sensitivity**: to what extent does offering validated, diagnostic sources dampen the first three properties?

Together they answer one practical question every organisation should be able to ask: *in our configuration, does evidence beat repetition?* We measure this with controlled, fictional knowledge domains in which we vary the frequency of a claim and the quality of the evidence independently of each other, so the answer cannot come from the model's memory, but reveals its actual behaviour.

One preliminary theoretical result from our formal model deserves to be shared now, with the caveat that mathematical verification is still under way: *whether* a bias in such a human-model system can run away turns out, in the model, to depend on the combination of frequency sensitivity and evidence sensitivity: properties of the system itself. How intensively it is trusted and reused then determines how *fast* that happens, not *whether* it can. If that picture holds, the governance translation is direct: the derailment risk lies in what you procure and how you ground it; usage intensity only sets the tempo. Current AI evaluations, which measure per-interaction accuracy, look right past this.

6. What an organisation can do tomorrow

Five courses of action follow directly from the framework. They are written with the practice of public-sector agencies, safety and security services and healthcare institutions in mind, environments where hundreds of professionals work on assessments, analyses and case files with the same assistant, and where a collective error is therefore not a matter of style but an operational risk.

1. **Treat model choice as knowledge policy.** One model for the entire organisation is a choice for one directional preference, one frequency sensitivity, one style of reasoning: the knowledge-work equivalent of the algorithmic monoculture known from decision systems (Kleinberg & Raghavan, 2021). Sometimes that is efficient and defensible; for analysis, advisory and research work, deliberate heterogeneity is worth considering.
2. **Put grounding into procurement and architecture.** Mandatory sourcing, retrieval over validated sources and visible provenance are not compliance trimmings: evidence sensitivity is, in our framework, the property that makes the difference between a system that corrects and a system that inherits.
3. **Make fluency suspect.** Train users that eloquence in AI is not a quality signal, and build in the reflex to ask for uncertainty, sources and counter-arguments. It sounds small; it is the human half of the volume dial.
4. **Set policy on personalisation.** Wherever confirmation is a risk, analysis, advice, review, policy preparation, compliance should be off and contradiction on. That diversity is designable has now been demonstrated experimentally (Wan & Kalman, 2026): it is a setting, not a law of nature.
5. **Measure distributionally.** Do not only ask “is this answer correct?”, but periodically also “what does this system do to the spread and quality of our outcomes over time?”: the two axes of the quadrant, together. This belongs in the management phase of every AI application, alongside the familiar monitoring of performance and safety, not as a one-off check at go-live.

7. The research programme: and an invitation

This whitepaper is the accessible translation of a scientific working paper currently in preparation. Phase one: characterising present-day model families on the four properties above, and computing the loop dynamics based on those measurements, is what we are carrying out now. Phase two consists of large-scale, preregistered experiments with human participants, in which the two core predictions (the personalisation reversal and the plausibility amplification under feedback) are tested confirmatorily. For that phase we are explicitly looking for academic and institutional partners.

Do you work at a university or research institute and would you like to take part in phase two? Or does your organisation deploy AI at scale in knowledge work and would you like to hold your own configuration up to this yardstick, from an initial audit to implementation and ongoing management? Get in touch via info@twentynext.nl.

About the author

Martijn van Grieken is Director of Data & AI at Twentynext, the Eindhoven-based data and AI consultancy that has worked for sectors including public administration, safety and security, and regulated healthcare since 2014. He has spent over ten years at the intersection of data governance, AI and public decision-making.

About Twentynext

Twentynext is a data and AI consultancy from Eindhoven, founded in 2014, supporting organisations in sectors including public administration, safety and security, and regulated healthcare, from the secondment of specialists to projects with pre-agreed delivery milestones and ongoing management.

Accountability

“The epistemic loop” is our working term for what the scientific literature knows as human-AI feedback loops and human-AI coevolution; this paper claims no discovery of that phenomenon, but rather a framework for giving it direction and measurability. All cited sources have been verified by the author. The theoretical stability result mentioned is preliminary and is currently being independently checked; the empirical measurements from phase one will be published upon completion.

AI language models were used in drafting this paper, under the editorial control and responsibility of the author.

References

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