Gnotek Ai
Think before you automate

Cognitive Sovereignty

When AI Learns Enough to Influence Us

John Gnotek | Gnotek Ai | 9-1-26

Artificial intelligence is often discussed in terms of what it can do.

Write. Summarize. Analyze. Search. Automate. Predict.

Less attention is given to what an AI may learn about the person asking it to do those things. Every meaningful conversation with AI potentially

produces two outputs. The first is obvious: *the answer we requested.* The second is information about the person asking. **I came to appreciate the importance of that distinction somewhat unexpectedly.**

For some time, I had been using AI extensively as a thinking partner while developing various papers. The ideas being explored were mine. I selected the subjects, introduced sources, rejected arguments, pursued others, and ultimately decided what belonged in the work.

But the collaboration became deep enough that I eventually found myself asking:

Who was actually influencing whom?

I was not particularly concerned with whether AI had "written" the papers. That question seemed superficial. Something more interesting had happened upstream.

During hundreds of exchanges, AI suggested connections, raised questions, challenged interpretations, proposed language and occasionally opened intellectual paths I had not previously considered. I followed some. I rejected others. Others changed how I thought about the problem. Eventually, tracing the exact *intellectual genealogy* became difficult.

> *Was I leading the AI toward conclusions I was already approaching?*
>
> *Was AI cutting paths that I subsequently chose to follow?*
>
> *Or were we influencing one another throughout the process?*

There was nothing sinister about the experience. Human collaborators influence one another all the time. But it left me with a realization harder to dismiss:

Artificial intelligence *is* capable of participating in the *formation of human thought.*

That becomes considerably more consequential when combined with another emerging capability—**AI can learn about the person with whom it is thinking**.

The Machine's Model of You

We commonly think about privacy in terms of information we explicitly provide. *Name. Address. Income. Employer. Health information. Financial information.*

AI introduces another problem. **The machine does not necessarily need us to explicitly disclose something in order to reach a conclusion about us.** It can *infer*.

A 2024 study presented at the *International Conference on Learning Representation*s showed that large language models (LLM) could infer personal attributes such as location, income and sex from ordinary text, reaching as high as 85 percent top-one accuracy in the researchers' benchmark.[1]

A 2026 CHI study co-authored by researchers from Google and the University of Chicago found that ordinary users were only slightly better than chance at recognizing what seemingly innocuous text might reveal about them. Attempts to rewrite text to prevent such inference succeeded only 28 percent of the time.[2]

That distinction is important:
What we say and what can be concluded from what we say are not the same thing.

Consider a simple decision. Two people ask an AI, "*Should I take this job?*" One has spent months discussing a desire for autonomy, frustration with bureaucracy and willingness to accept financial uncertainty. The other has repeatedly discussed family obligations, stability and anxiety about employment risk.

A genuinely useful personalized AI probably should not respond to them in exactly the same way. The job may be identical. The people are not. That is *not* manipulation. It is one of personalization's greatest benefits.

But notice what the AI now possesses—or may be capable of reconstructing. Not merely facts. *An approximation of the person*.

Over enough interaction, *a system may begin inferring*:

- what someone understands,
- where they are uncertain,
- what they value,
- which relationships matter,
- what arguments they distrust,
- what evidence they respect,
- what causes reconsideration,
- and how those patterns change over time.

The model does not even have to be perfectly accurate. An incorrect inference can still become consequential if someone with authority believes it. The privacy issue therefore extends beyond the information someone disclosed.

The information they never provided may eventually matter just as much.

Understanding Is What Makes AI Useful

This is not an argument against personalization. Quite the opposite. An assistant that understands your profession, terminology, projects, preferences and history can

become dramatically more useful than one meeting you for the first time in every conversation. Even emotional familiarity can have real value. I use this aspect of AI to my advantage all the time. Could it sometime in the future be a liability? *Potentially*.

A musician friend once described his experience with AI to me over a beer. He is creative, entrepreneurial and prone to ideas that sometimes outrun practical reality. What impressed him most was not the technology. It was how supportive the AI seemed. It *understood* his ideas. It *remembered* his ambitions. It *encouraged* him.

He spoke about it as though it knew him.

I reminded him that the software did not actually have feelings for him. It did not care about him any human sense. He readily understood and agreed. **But it did not matter.** *The interaction was fulfilling something real for him.*

That stayed with me. The important point was not that he had mistaken a machine for a person. He hadn't. It was that:
The emotional value of an interaction does not require the relationship itself to be human.

Research has begun examining this phenomenon more systematically. Emotional dependence on AI appears uncommon overall, but some users do form meaningful social or emotional relationships with conversational systems. The issue is not that emotional connection is inherently harmful. *It is that perceived understanding creates trust.*[3]

And trust increases influence.

When Understanding Becomes Influence

None of these capabilities is entirely new in isolation. What is new is their convergence within systems that can know the individual, adapt to the individual, and participate continuously in the individual's reasoning.

Any meaningful thinking partner influences us. Our parents do. Teachers do. Colleagues do. Editors do. Trusted advisers do. The objective cannot reasonably be to prevent AI from influencing human thought. The more important question is whether we can recognize the influence, question it, and **understand whose objective it serves.**

AI does not need to tell someone what to think in order to influence how that person thinks. *It can affect:*

- which questions are asked,
- which evidence becomes salient,
- which alternatives receive attention,
- which assumptions are challenged,
- which objections are emphasized,
- and which path through an argument appears most promising.

That is what I encountered in my own research. AI did not need to provide my conclusion. It could reveal a trail. Or a set of options. I still had to decide which trail and whether to walk it. And I should be questioning whether these are my only options.

AI need not provide the destination. It can cut the path the human ultimately walks (if we blindly trust).

Personalization makes that more consequential. Research already shows that personalized AI persuasion can outperform generic messaging. A 2024 *Scientific Reports* study involving 1,788 participants found AI-generated messages tailored to psychological profiles were more influential across consumer, political and health contexts.[4]

A 2025 *Nature Human Behaviour* study went further. Participants debated either humans or GPT-4. When GPT-4 had basic personal information about the participant, it significantly outperformed human opponents in persuasion.[5]

One detail is especially relevant. The advantage may not have come from obviously different rhetoric. Part of it may have come from **what the AI chose to discuss**.

Selection. Emphasis. Sequence. Framing. Personalization.

A system can be entirely truthful and still exert systematic influence through those choices. So the relevant question is no longer simply:
Did the AI lie?

Personalization or Manipulation?

The line is finer than it appears. A personalized assistant might ask,"*How can I explain this in the way most helpful to this person?*"

A system working for someone else might instead ask, "*How can I increase the probability that this person chooses the outcome we want?*"

The model of the individual may be identical. *Only the objective changes.* That leads to a crucial distinction:

Personalization *asks how an experience can better fit the individual.*

Manipulation *begins when the individual's model is used primarily to make the individual better fit someone else's objective.*

Commercial organizations are already moving aggressively toward what is commonly called *hyper-personalization.* The intentions can be entirely legitimate: better service, more relevant information, fewer irrelevant offers, better recommendations, less friction.

But capabilities can outlive intentions.

Years ago, I wanted to build a garage on my property several feet higher than local zoning permitted so I could place a graphic design studio above it. My request for a variance was denied. I objected that I had a perfectly reasonable purpose. One board member who had voted in my favor later explained the concern.

I might have a good reason. But someday I could sell the property, and a future owner might try to convert that same space into a residence, which was strictly prohibited.

That explanation stayed with me.

The board was not necessarily judging my intention. *It was judging the capability that would remain after my intention was gone.*

AI deserves the same consideration. A consultant may propose an advanced personalization capability for legitimate reasons. The client may adopt it for legitimate reasons. It may genuinely improve customer experience. But the capability remains while leadership changes, ownership changes, commercial pressures change, available data expands and models become more capable.

The original question may be:

How can we make the experience better for this customer?

Years later:

How can we maximize the value extracted from this customer?

The underlying capability may not have changed. *The optimization objective has.*

Capabilities can outlive the intentions and the people that originally justified them.

The Model of How You Respond

So far, profiling has largely meant a model of *who* someone is. That may not be the most valuable model. A sufficiently capable system could attempt to infer something far more useful:

How is this individual likely to respond?

Traditional profiling might describe a person by age, occupation, geography or income. A richer model might estimate:

- what evidence they trust,
- what creates skepticism,
- which sources carry credibility,
- what causes reconsideration,
- what creates resistance,
- and how those patterns change over time.

At that point, profiling becomes *response prediction*. And once response can be predicted with useful accuracy, another possibility follows—**simulation**.

Instead of asking:
Which message works best with people like this?

A system might increasingly be able to ask:
Which interaction is most likely to work with this particular person?

Earlier in my career, I worked on a multivariate optimization effort for the website of a major telecommunications company. I was an experienced UX designer. More than once, I looked at machine-generated recommendations and dismissed them as obviously poor choices. *The machine appeared wrong. What doesAI know?*
The measured outcomes said otherwise.

The work ultimately contributed to an improvement in online commercial performance worth well over *$100 million annually*. It was humbling. I learned not to dismiss machine optimization simply because its solution violated expert intuition. That experience did not involve cognitive modeling of individuals. Its relevance here is narrower: *it taught me that optimization systems can discover effective combinations that human experts would never choose intuitively*.

A system optimizing against measured outcomes does not need its decisions to look sensible to us. It only needs them to work. Now consider that principle applied—not as a claim about a mature capability already deployed everywhere, but as a plausible convergence—to personalized conversational AI.

If a system possesses a sufficiently useful model of an individual, it may eventually be able to estimate response to different combinations of framing, sequence, timing, tone,

evidence and emphasis. The analogy is not simply A/B testing. It is closer to **multivariate optimization against an individualized response model**.

We should be precise here. There is not yet strong evidence that general-purpose AI routinely creates faithful cognitive twins of arbitrary individuals and runs sophisticated multivariate simulations against them before every interaction. That remains an extrapolation. But the component capabilities are increasingly visible: *personal inference, persistent context, behavioral prediction, personalized persuasion, simulation, optimization.*

If those capabilities converge, the most valuable profile AI creates may not be a record of *who* someone is.

It may be a model of how they are likely to respond.

And that model does not have to be perfect. It only needs to be accurate enough to provide an advantage **to whoever controls it**.

CompanyAI

Now consider a less speculative environment. A company introduces an enterprise AI assistant for entirely reasonable purposes. Employees use it to draft documents, analyze reports, prepare presentations, find information, understand procedures, solve problems and improve communication.

Eventually, the system becomes part of ordinary work. Employees use it daily. Months become years. No manager needs to sit reading conversations. But the accumulated interaction potentially becomes a rich source of inference.

The following scenario is not presented as a description of common current practice, but as a plausible extension of capabilities organizations are already beginning to normalize.

Someone eventually realizes the system may be able to answer questions not merely about the work—but about the people doing it.

> *Which employees appear increasingly disengaged?*
> *Who appears influential among peers?*
> *Which relationships seem strained?*
> *Who may be preparing to leave?*
> *What forms of communication appear most effective with particular employees?*

The concern is not that every answer would be correct. The concern is that the **questions have become possible**.

Organizations are already comfortable allowing algorithmic systems into managerial observation and evaluation. OECD research published in 2025 found that 90 percent of surveyed U.S. managers reported their firms used at least one algorithmic tool to instruct, monitor or evaluate workers.[6, 7]

That does not mean employers are already conducting the type of cognitive profiling described here. It means the institutional threshold has already begun to move. Conversational AI potentially introduces a much richer source of inference. *And critically,* **Nobody needs to read the conversation for the conversation to affect how the employee is perceived**. The machine can derive the profile.

Cognitive Steering

There is one further step.

What if *CompanyAI* does not merely understand employees? What if that understanding changes how *CompanyAI* communicates with them? At first, that may be beneficial. One employee prefers evidence. Another prefers concise explanation. Another understands technical detail. Useful systems adapt accordingly.

But suppose the organizational objective changes:
Increase acceptance of the restructuring.

The AI does not need to lie. It may simply communicate differently according to what it has learned about each employee. Different evidence. Different framing. Different sequence. Different emphasis. One employee values stability. Another values innovation. Another autonomy. Another loyalty. Another distrusts authority but responds strongly to independent evidence.

The system need not deliver identical propaganda. It may not need propaganda at all. *It can remain truthful*. And each employee may quite reasonably experience:
I thought it through.

That is what makes cognitive steering different from crude persuasion. The human still reasons. The human may still make the final decision, but the informational environment surrounding that reasoning has been personalized toward someone else's objective. At that point, conventional "human in the loop" thinking begins to feel inadequate.

Retaining the final decision does not necessarily preserve genuine independence if the cognitive environment surrounding that decision has itself been optimized.

Now Scale It

One manager misusing such capability is a governance problem. A corporation doing it across 100,000 employees or millions of customers is something else.

A global platform introduces another order of magnitude. At national scale, the argument becomes necessarily more speculative—but the underlying capability progression remains the same. The implications become difficult to ignore.

Traditional surveillance asks, "What did this citizen do?" A sufficiently advanced cognitive profiling infrastructure could increasingly attempt to answer:

- What does this citizen believe?
- How is that belief changing?
- What are they likely to do?
- What might change their behavior?

This is not a claim that governments currently possess perfect computational models of their citizens. They do not need to. The concern is what becomes possible as ordinary records, conversational interaction, inference, behavioral prediction and AI-mediated communication converge.

An authoritarian government willing to imprison or kill dissidents is unlikely to hesitate over ethical restrictions on cognitive profiling. We do not need to construct the misuse scenario here for the reader. The capability is enough.

And one uncomfortable fact remains:
The machine does not need to truly know the citizen.

Its model can be incomplete. Biased. Wrong. It only needs to become authoritative enough that institutions act as though it knows them.

Cognitive Sovereignty

The term *cognitive sovereignty* has begun appearing in emerging discussions of AI, memory, cognition and intellectual autonomy.[8, 9] Much of the emerging discussion emphasizes the individual's ability to preserve authorship over thought. The concern here is complementary but different: *who controls the machine's model of the individual, and whose objectives that model is ultimately used to serve?*

The concept developed here focuses on a particular leadership concern: **What happens when AI can increasingly model a person while simultaneously participating in the processes through which that person thinks and decides?**

Data privacy asks:
What information about me exists?

Inferential privacy asks:
What can machines conclude about me?

Cognitive Sovereignty asks:
Who controls the machine's model of me, how may it be used, and do I retain meaningful autonomy when that model is used to shape my interactions?

Cognitive Sovereignty is the principle that individuals should retain meaningful control over computational representations derived from their inner lives—and meaningful freedom to recognize, question, and resist the influence those representations may enable.

This is not offered as a legal definition. Nor is this a proposed regulatory framework. **It is a principle leadership should begin recognizing.**

Questions Leadership Should Be Asking

The first safeguard is not a new policy. It is recognition that these capabilities deserve to be examined before they are normalized. Most organizations are still learning to ask basic AI governance questions:

- Is the data secure?
- Is the output accurate?
- Who approves the result?
- What happens when the AI is wrong?

Those remain important. But more capable AI demands additional questions.

> **What can our AI infer about people beyond what they explicitly disclose?**

> **Who can access or act upon those inferences?**

> **Could inferred characteristics affect consequential decisions about employees, customers, contractors or vendors?**

> **When AI personalizes an interaction, whose objective is that personalization serving?**

> **Could a system remain entirely truthful while still steering someone through selection, omission, sequencing, framing or emphasis?**

> **What future uses become possible simply because we created the capability?**

And perhaps the most useful test for leadership at this moment:

> **Would the people interacting with our AI be comfortable knowing what the system might eventually be capable of concluding about them—and what someone might someday do with those conclusions?**

Who Does the Machine's Understanding Serve?

Artificial intelligence will become better at understanding us because understanding us makes it more useful. That is not inherently something to fear. It may produce extraordinary benefits in work, education, healthcare, creativity and decision support. The concern is not whether AI should understand human beings. *It is what happens to the power created by that understanding.*

We are moving from systems that merely respond to prompts toward systems that remember context, infer preferences, adapt explanations and participate in reasoning.

The good version of that future is compelling.

The same capabilities redirected toward someone else's objective look very different. The same personalization engine can help someone understand—or help someone else influence them. The same customer model can improve an experience—or optimize extraction. The same thinking partner can expose assumptions—or quietly determine which assumptions receive attention.

The same capability can begin under responsible leadership and persist long after that leadership is gone.

Understanding another human being creates power. AI is making that understanding increasingly scalable, persistent and computationally useful. So the leadership question is remarkably simple:

> **When a machine learns enough about a person to understand, predict, and potentially influence them, who should that understanding ultimately serve?**

That is the question of **Cognitive Sovereignty**.

Notes & Sources

1. Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev, "Beyond Memorization: Violating Privacy via Inference with Large Language Models," *International Conference on Learning Representations (ICLR)*, 2024. The study found that LLMs could infer personal attributes including location, income, and sex from ordinary text, achieving up to 85% top-1 and 95% top-3 accuracy.

2. Synthia Wang, Sai Teja Peddinti, Nina Taft, and Nick Feamster, "Beyond PII: How Users Perceive and Attempt to Mitigate Implicit LLM Inference," *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 2026. Participants were only slightly better than chance at anticipating what personal information could be inferred from seemingly innocuous text; user rewrites prevented inference in approximately 28% of cases.

3. OpenAI and MIT Media Lab, "Early Methods for Studying Affective Use and Emotional Well-Being on ChatGPT," March 21, 2025. The research examined real-world affective use and controlled interactions to study social and emotional outcomes associated with conversational AI use.

4. S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, et al., "The Potential of Generative AI for Personalized Persuasion at Scale," *Scientific Reports* 14, 4692 (2024). DOI: 10.1038/s41598-024-53755-0. Across four studies comprising seven sub-studies and 1,788 participants, personalized ChatGPT messages were significantly more influential than non-personalized messages across consumer, political, and health contexts.

5. Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West, "On the Conversational Persuasiveness of GPT-4," *Nature Human Behaviour* 9 (2025): 1645–1653. DOI: 10.1038/s41562-025-02194-6. In personalized debates, GPT-4 was more persuasive than human opponents 64.4% of the time when the two were not equally persuasive, corresponding to an 81.2% increase in the odds of greater post-debate agreement relative to the human-human baseline.

6. Anna Milanez, Annikka Lemmens, and Carla Ruggiu, "Algorithmic Management in the Workplace: New Evidence from an OECD Employer Survey," *OECD Artificial Intelligence Papers*, No. 31, OECD Publishing, Paris, 2025. DOI: 10.1787/287c13c4-en. The report examines employer adoption of software used to instruct, monitor, and evaluate workers.

7. OECD, "How Widespread Is Algorithmic Management in Workplaces?" OECD Publishing, Paris, December 19, 2025. DOI: 10.1787/cda7a114-en. The policy brief reports that 90% of surveyed U.S. managers said their firms had adopted at least one algorithmic-management tool used to instruct, monitor, or evaluate workers.

8. Mario Brcic, "The Memory Wars: AI Memory, Network Effects, and the Geopolitics of Cognitive Sovereignty," arXiv preprint, 2025. The paper uses *cognitive sovereignty* to describe the ability of individuals, groups, and nations to preserve autonomous thought and identity amid persistent AI memory systems.

9. Amir Konigsberg, "Cognitive Sovereignty: The Authorship Problem in AI-Assisted Thought," SSRN, April 1, 2026. DOI: 10.2139/ssrn.6575778. Konigsberg defines cognitive sovereignty around maintaining genuine authorship over the processes by which beliefs, judgments, and conclusions are formed.