THE SOCIAL OPERATING SYSTEM

How shared references decide how fast an organization can move

The social operating system: the sum of sufficiently shared references that every action in an organization runs on, and that every action in turn updates.

A working paper by Christoph Svoboda, author of The Hidden Cost of Speed. Version 1.0, Vienna, October 2026.

Abstract

Accelerated organizations have learned to scale output: analyses, drafts, decisions and, increasingly, the actions of AI agents. What does not scale with them is the shared reading on which coherent distributed execution depends. This paper names the layer that pressure lands on: the social operating system, the sum of sufficiently shared references through which people interpret priorities, reality, meaning, decision logic and next sensible steps, the layer underneath processes and org charts on which every instruction actually runs. We argue that this layer erodes in a characteristic sequence that stays invisible until its late symptoms arrive as rework, escalation and strategic surprise; that AI widens the gap because speed is bought per person while shared readings are built together or not at all; and that the capacity to form new shared references is limited, which makes the binding question not how much is new but how much can become shared at once. Three observations make the condition visible to any leader without instrumentation, and beyond hand range it becomes measurable.

1. What the dashboard does not show

Leaders judge from what is visible, and in an accelerated organization almost everything visible is output: shipped, sent, decided, drafted, produced. The numbers that describe this output are better than they have ever been, and they update daily. What no dashboard shows is whether the people producing that output still read the situation the same way: whether priority three means the same thing on the second floor as it does in the board room, whether a decision that was made is also the decision that was heard, whether two teams executing the same strategy are still executing the same strategy.

A judgment formed from what is visible quietly treats the invisible as absent (Kahneman, 2011). This is not a personal failing; it is how judgment under time pressure works. And it explains a pattern that repeats across fast-adopting systems of every kind: the benefits of acceleration are legible early, in exactly the metrics everyone watches, while its costs accumulate in a layer nobody is instrumenting.

This paper is about that layer. Its claim is simple to state. The operating constraint of accelerated organizations is no longer the volume of output they can produce. It is the stability of their shared references: the sufficiently common reading of priorities, reality, meaning, decision logic and next sensible steps that allows distributed people to act independently without quietly working against

each other. Sufficiently is the operative word. A shared reference does not require identical opinions. It requires enough common ground that local decisions do not silently depart from one another's assumptions. The intuition has deep empirical roots: converging task and team models have long been treated as a condition of coordinated performance in team research (Mathieu, Heffner, Goodwin, Salas and Cannon-Bowers, 2000; DeChurch and Mesmer-Magnus, 2010), and coordination across roles has been shown to rest on shared goals and shared knowledge as much as on communication frequency (Gittell, 2000).

2. The compression layer

Two colleagues agree on a wall color: zircon. It is a new shade, not yet in the color fan they usually work from. One knows zircon as a lipstick color from this season's campaigns; the other saw it on a car in a weekend magazine ad. Both leave satisfied. The mistake only shows once the wall is painted.

The example is trivial, and that is its value, because organizations run on the same mechanism at every level of consequence. Words like focus, quality, speed, ownership and priority behave exactly like that color name. They feel precise in the room and they are anchored in different internal catalogues.

Shared references are why organizations can be fast at all. They work as a compression layer. When a reference is stable, a single word carries an entire agreement, and nobody has to rebuild meaning from scratch. A nod can safely stand in for a contract. Linguists describe the upkeep behind this as grounding, the continuous and mostly invisible maintenance of common ground between speakers (Clark and Brennan, 1991). The risk begins when the compression runs on references that only appear to be shared. Then the organization moves at the speed of the nod and discovers the difference at the speed of execution.

The sum of these references is what this paper calls the social operating system: the layer underneath the processes, systems and org charts, on which every instruction actually runs. The process chart says who does what and in which order. The social operating system decides what the words in the process mean. Organizations maintain their technical operating systems with discipline and budgets. The social one is mostly maintained by accident.

3. Output now scales. Meaning never did.

What has changed in the last few years is not the mechanism. It is the exchange rate.

Speed, in organizations, is bought individually: a faster tool, a faster analysis, a faster draft, per person, per desk. A shared reading cannot be bought that way, because it is built together or not at all. When tools accelerate individual output and nothing deliberately maintains the collective reading, the gap between the two compounds.

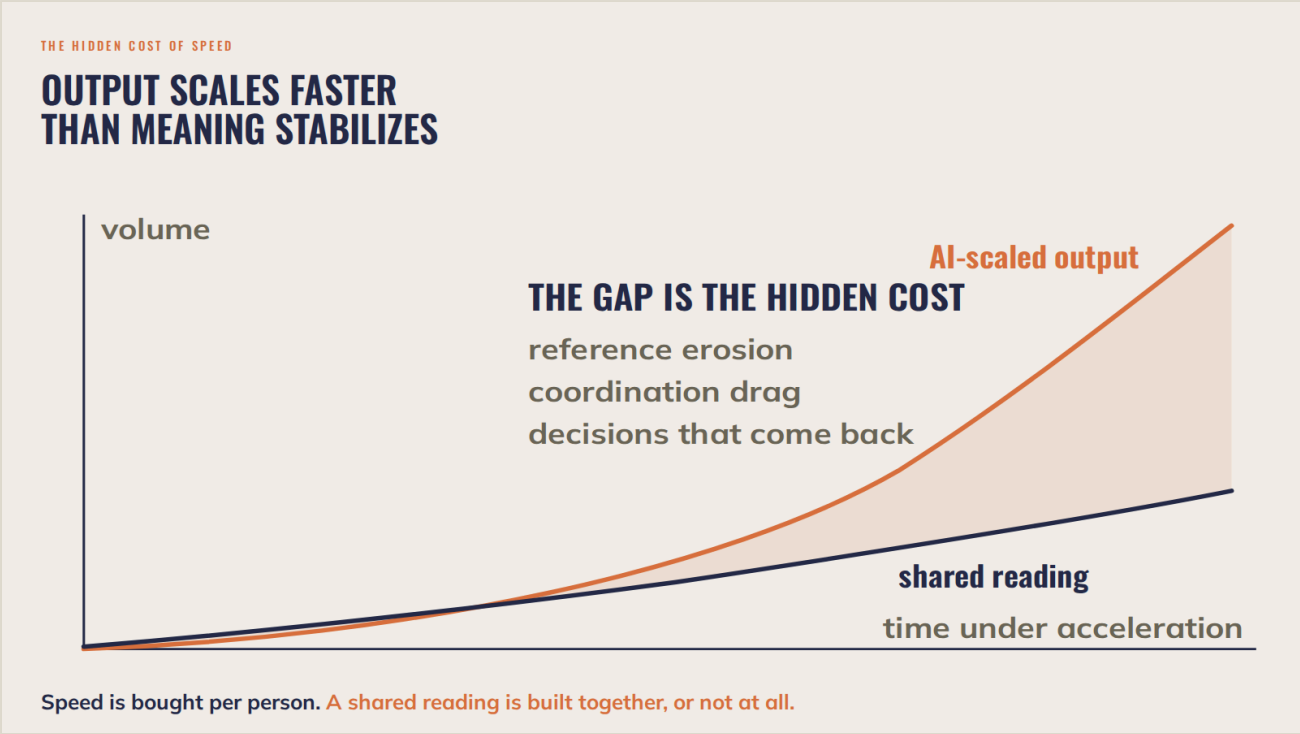


Figure 1. Output scales with tools; a shared reading is built together. The widening gap is where the hidden costs accumulate. Conceptual illustration, not measured data.

AI has made this exchange rate brutal, and not in the way most commentary suggests. The problem is rarely a single wrong answer. The problem is large volumes of plausible, decision-ready material entering the system faster than the system can interpret it in common. Each document is locally useful. Each team acts on its own reasonable reading. Collectively, the readings thin out, while the appearance of productivity has never been better.

The same logic now extends to AI agents, and this may be the most consequential version of it. Agents do not join the org chart; they join the coordination layer. Every agent acts on some reading of the strategy, whether anyone made that reading explicit or not. An organization that cannot keep its people on one map will not fare better once agents are added to it. Agent readiness, which Fagerland (2026) defines as the organizational capacity to integrate, govern, monitor, override and learn from increasingly agentic systems, may therefore turn out to be in large part reference readiness (on shared mental models in teams of humans and AI agents, see Andrews et al., 2023).

One reported case makes the cost concrete. In the engineering unit of a large industrial group, retold here with details altered to protect the source, AI-assisted planning compressed a standard task from roughly a week to a fraction of it, and throughput per planner multiplied. Around the same time, one planner developed a substantially simpler design for a common subsystem, cutting its cost by a large multiple. In the old rhythm, such a practice spreads across desks within days, through the working exchange that ran alongside the projects themselves. That rhythm was never slack; the same planners carried full loads before, and the exchange still happened, at a rate healthy enough to move innovations and better practices through the unit quickly. What changed is the arithmetic. Three projects a week instead of one mean three ramp-ups instead of one, each with its own context to absorb, and all of it comes out of the same week. Nobody decided to cut the exchange. It was simply the first thing the new arithmetic reallocated. The better design took weeks to travel,

several projects shipped without it, and the difference ran into millions. Speed in one dimension had been bought with slowness in another. Innovation kept appearing. It stopped traveling. The mechanism is well described in research on transactive memory, the system by which groups know who knows what: it is built and refreshed through interaction, and face-to-face exchange plays an outsized role in it (Lewis, 2004).

4. How the erosion runs

Reference erosion follows a sequence, and the sequence explains why it is so rarely caught early.

It begins invisibly: the common reading thins. Nothing happens yet. Meetings conclude, documents circulate, work continues. Then coordination begins to drag. More effort is needed to keep parallel work compatible, decisions take longer to operationalize, and leaders spend growing energy repairing meaning between layers. This stage is felt but rarely attributed; it registers as friction, personalities, workload. Only in the third stage do the visible symptoms arrive, the ones organizations actually act on: rework, repeated clarification rounds, escalation meetings, confusion about priorities, a slow decline in the general quality of what ships, and eventually the expensive discovery that strategy and execution have been describing different companies for some time. Research on cross-functional teams draws the same line: the expensive boundaries between specialists are rarely about vocabulary; they are semantic and pragmatic, differences of meaning and of what is at stake (Kotlarsky, van den Hooff and Houtman, 2015).

The middle of this sequence deserves a precise name, because it is where organizations spend the longest and see the least. Benja Stig Fagerland, Associate Professor at the University of South-Eastern Norway, in a recent exchange about this work, called it decoupling: an organization that remains operationally coordinated while its interpretations quietly part ways. The schedule holds, the rituals run, the reports arrive, and the common reading underneath them is already gone. Fagerland adds a second observation this paper fully subscribes to: what is eroding here is "not a stock of alignment; a living capability". A shared reading is not achieved once and then owned. It reproduces itself through use, whether anyone manages it or not: every decision draws on the frame and writes something back into it. The updating never stops. What breaks under speed is the coordination of the updates. Decisions keep writing back, but locally, per person, per team, and the references they renew stop being shared ones. References remain; shared references thin out. Erosion is not a system that has stopped updating. It is a system whose updates no longer converge.

By the time the symptoms are visible, the cause is months old. That is what makes the hidden cost hidden: the feedback arrives too late to point at its origin.

5. The renewal constraint

If updates converge only where people build the reading together, then the capacity to form new shared references becomes a strategic quantity, and it has two properties leaders tend to discover late.

First, forming a shared reference takes common time. A new term, a new priority, a new tool, a new structure: each starts as a private reading in many heads and becomes a shared one only through use, friction and correction. This cannot be delegated to a document. The announcement is the beginning of the process, not its end.

Second, the capacity is limited, and every piece of novelty draws on the same budget. A few changes, introduced in sequence, each reach the point of being sufficiently shared. Many changes, introduced in parallel, crowd each other out, and none of them anchors: everyone is partially briefed on everything and jointly fluent in nothing. The question for a leadership team is therefore not how much is new. It is how much can become shared at once.

Under execution pressure, the time in which references form is the first thing to be cut, because from the outside it looks like slack: the loose exchange, the overlap between projects, the conversation that is not on any agenda. The engineering case above is exactly this cut, made silently by a throughput gain. Which raises the question every accelerating organization now faces about its productivity dividend: where do the freed hours go? If all of them are reinvested in more throughput, the organization is spending its coherence to buy its speed. Speed that consumes the time in which references form is not acceleration. It is borrowing against coherence, at interest.

6. What a leader can see without instruments

None of this requires believing a theory. Three observations, available to any leader, make the state of the shared reading visible at hand range.

Have your priority explained back to you. Ask three people on three levels to justify the current top priority in their own words. Not repeat it, justify it. Three clean but different justifications mean the agreement is surface: people said yes, but did not hear the same thing.

Count the decisions that come back. Decisions that are formally made and still reappear, reopened in other rooms, renegotiated by email, escalated through side doors, are usually not resistance. They are divergent readings surfacing late. A decision that comes back was not a decision. It was a postponement with minutes.

Watch the translation load. Ask your direct reports how much of their week goes into explaining decisions downward and translating questions upward. When that load grows while communication increases, volume has stopped being a remedy and has become a symptom.

7. Making the reading visible

Within hand range, attention is enough. Beyond hand range, across hundreds or thousands of people, levels, functions and regions, no leader can hear the readings of the whole system, and this is where the condition becomes measurable rather than merely discussable: where an organization reads its decisions the same way, and where the readings quietly part ways, by level, by function, by region. We built the Shared Reference Diagnostic for exactly this purpose. This paper is deliberately

not about the instrument. The argument stands on its own; the measurement exists for the scale at which argument stops being enough.

8. About this paper

This is a working paper, and it is published as one: the argument is stable, the edges are open. Before publication it went to a small number of readers whose own work touches its core, and it will keep changing with its readers' readings of it, which is only fitting for a paper about shared readings. If something here is wrong, half right, or named better in your world than in ours, that is exactly the response this paper is for.

References

Andrews, R. W., Lilly, J. M., Srivastava, D. and Feigh, K. M. (2023). The role of shared mental models in human-AI teams: a theoretical review. Theoretical Issues in Ergonomics Science, 24(2), 129-175. https://doi.org/10.1080/1463922X.2022.2061080

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine and S. D. Teasley (eds.), Perspectives on Socially Shared Cognition (pp. 127-149). Washington, DC: American Psychological Association. https://doi.org/10.1037/10096-006

DeChurch, L. A. and Mesmer-Magnus, J. R. (2010). Measuring shared team mental models: a meta-analysis. Group Dynamics: Theory, Research, and Practice, 14(1), 1-14. https://doi.org/10.1037/a0017455

Fagerland, B. S. (2026). Fra KI-adopsjon til agentberedskap: Tre målbare gap – og hvordan kunstige agenter endrer premissene for ledelse [From AI adoption to agent readiness: three measurable gaps, and how artificial agents reshape the premises of leadership]. Magma.

Gittell, J. H. (2000). Organizing work to support relational co-ordination. The International Journal of Human Resource Management, 11(3), 517-539. https://doi.org/10.1080/095851900339747

Kahneman, D. (2011). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux.

Kotlarsky, J., van den Hooff, B. and Houtman, L. (2015). Are we on the same page? Knowledge boundaries and transactive memory system development in cross-functional teams. Communication Research, 42(3), 319-344. https://doi.org/10.1177/0093650212469402

Lewis, K. (2004). Knowledge and performance in knowledge-worker teams: a longitudinal study of transactive memory systems. Management Science, 50(11), 1519-1533. https://doi.org/10.1287/mnsc.1040.0257

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E. and Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. Journal of Applied Psychology, 85(2), 273-283. https://doi.org/10.1037/0021-9010.85.2.273