Kashi — Adversarial / Gaming Perspective

Research memo for product, governance, and pitch refinement

Prepared for	Purpose	Date
Kashi project team	Strengthen the concept with a realistic routing-around-the-system analysis	21 April 2026

What this memo is for

This memo does not ask whether Kashi can detect structural meeting asymmetry in principle. It asks the harder question: once people know what is measured, how are they likely to adapt, reroute pressure, launder responsibility, or perform symbolic compliance — and what should Kashi therefore change in its product doctrine, governance language, and claim discipline?

Executive judgment

The research supports a hard conclusion: adversarial adaptation is not an edge case of measurement systems. It is a normal response once actors understand what is visible, what is risky, and what consequences may follow. In other words, Kashi should assume that at least some people will learn the detector surface and route pressure around it.

- Metric improvement is not the same thing as behavioral improvement. Once a metric matters, some actors optimize around the metric rather than around the underlying organizational objective.
- In monitored settings, pressure often moves across channels. Public compliance can coexist with private coercion, side-channel retaliation, or displacement into 1:1s and offline conversations.
- Meeting power is not reducible to interruptions. Sequencing, agenda control, pre-closing, selective response, and topic-credit capture can suppress participation without obvious overlap or overt hostility.
- Harm can be laundered down the hierarchy. Senior actors may keep their own visible metrics clean while delegating exclusion, aggression, or information withholding through intermediaries.
- Kashi is already directionally strong because it is structural, longitudinal, and bounded. But it still needs an explicit “how actors route around the system” section, plus product rules for interpreting suspiciously clean metrics.

Project-use answer in one sentence

Short version

Kashi should not present itself as a system that simply reveals hidden power dynamics; it should present itself as a system that surfaces structural risk patterns while explicitly accounting for adaptation, displacement, and responsibility laundering once the system becomes socially known.

1. Why this point matters for Kashi specifically

Kashi’s current materials are already more mature than a generic workplace-AI pitch. They emphasize structural metadata rather than content, reject emotion inference, refuse HR-decision use, rely on longitudinal patterns rather than single events, calibrate against each speaker’s own baseline, and explicitly reject a company-wide “relationship health” score. Those are real defenses, not cosmetic ones. But they do not yet fully answer the adversarial question: what happens after actors learn the detector logic?

This gap matters because Kashi’s own materials already contain the seed of the argument. The project correctly notes that the moment a visible number becomes a compliance target, the number gets falsified.

The adversarial extension is simple: if a system makes certain behaviors visible, some actors will reduce the visible behavior while preserving the underlying power relation through another mechanism or another channel.

That is not a reason to abandon the idea. It is a reason to harden the concept. The product becomes more credible, not less, when it openly states where people will adapt and how Kashi intends to read around that adaptation.

2. Research synthesis: what the evidence supports

2.1 Measurement systems predictably produce gaming and goal displacement

The strongest general point is also the simplest: when a metric affects rewards, attention, status, or risk, people do not merely "improve" under it; they often optimize against it. Ian Larkin's field study of a large enterprise software vendor is a clean example. Salespeople learned to game the timing of deal closure under a nonlinear commission system, and the mispricing created by this behavior cost the vendor an estimated 6–8% of revenue. The underlying lesson is bigger than sales: once a measurement regime matters, rational actors learn its edges.

Related performance-measurement literature reaches the same conclusion at a more conceptual level. Jenny Lewis argues that performance measurement is never just a neutral technical exercise; it is political, power-laden, and prone to unintended consequences. The literature she reviews explicitly treats gaming as a normal consequence to be anticipated, not a weird bug that appears only when users are unusually malicious.

For Kashi, the implication is immediate: no single behavioral metric should ever be treated as transparent evidence of genuine improvement. A reduction in interruptions may be real improvement, but it may also be substitution, displacement, or strategic signaling.

2.2 Organizational analytics and workplace monitoring create cat-and-mouse adaptation

Recent research on organizational analytics describes the future of measurement at work as a "cat-and-mouse" game: organizations attempt to quantify behavior, while workers alter their signaling to meet situated goals. This is important because it moves the discussion beyond crude cheating. The issue is not only falsification. It is also selective signaling, impression management, and communicating in ways that undermine the usefulness of the metric while remaining legible as compliant.

OECD evidence on algorithmic management points in the same direction from a different angle. In European countries and Japan, staff resistance is one of the main reasons firms do not adopt algorithmic management tools, and the report repeatedly emphasizes worker consultation as a condition for acceptance. The same report notes a mitigating pattern: where employees have substantial influence over company decisions and are consulted during the implementation of new systems, algorithmic management does not show the same negative effects on autonomy, trust, job satisfaction, and motivation. This matters because opaque, top-down, high-stakes systems generate more resistance pressure — which in turn increases the incentive to route around the system.

For Kashi, this means adversarial behavior should not be modeled only as deliberate bad-faith sabotage by toxic managers. Some adaptation will be broader and more structural: defensive compliance, channel shifting, or strategic minimalism in response to being monitored.

2.3 Monitored behavior often moves offstage into private channels

Research on platform labor provides a useful analogue. Yu and colleagues show that delivery workers create private WeChat groups that function as "hidden transcripts" of resistance: public behavior remains compliant under the platform's gaze, while tactical discussion, dissent, and mutual aid move into spaces

invisible to power. The specific setting is not meetings, but the mechanism generalizes well: once a monitored surface is known, behavior that is costly onstage can migrate offstage.

Applied to Kashi, the risk is straightforward. Overt pressure in a group meeting may decline, while coercion, intimidation, exclusion, or retaliation move into 1:1 calls, corridor conversations, off-calendar chats, or delegated managerial conversations that the system does not observe. This does not make Kashi useless. It means Kashi must state clearly that absence of in-meeting signal is not proof of absence of pressure overall.

2.4 Meeting power can shift from interruption to sequencing, agenda control, and pre-closing

A major research-backed weakness of interruption-heavy narratives is that institutional meeting power often operates through sequencing rather than overlap. Månsson's conversation-analytic work on meeting talk shows that the chair controls topic progression by introducing items on the agenda and closing topics through summarizing formulations. Those formulations do not merely repeat what was said; they can transform it, narrow it, and move the meeting to closure. In plain terms, a person can stop interrupting and still choke participation by deciding what counts as relevant, what gets summarized, and when a topic is considered closed.

Voice research reinforces this from another side. Burris shows that managers view employees who use more challenging forms of voice as worse performers and endorse their ideas less than those who use more supportive voice. This means a superficially polite or orderly meeting can still structurally punish dissent. No overt overlap is required. Participation can be sanctioned through endorsement patterns, floor allocation, and the reputational cost attached to speaking plainly.

This is directly relevant to Kashi's own "slightly dangerous" scenario: the system under-calls the content-hostile, dismissive dynamic because the structural indicators remain borderline. That is not a failure of engineering so much as a reminder that actors can preserve surface order while still narrowing who gets heard.

2.5 Harm can be laundered down the hierarchy

Abusive-supervision research suggests that harmful behavior often propagates rather than staying attached to the original source. Mawritz and colleagues find support for a trickle-down model of abusive supervision across three hierarchical levels. In practice, this means the person with the most formal power does not need to remain the visibly aggressive actor for the organization to remain abusive.

For Kashi, the adversarial implication is important: once a senior actor understands what is visible, they can keep their own dashboard relatively clean while pressure is executed by a direct manager, team lead, or senior individual contributor. The underlying climate remains unhealthy, but the visible signature shifts position in the hierarchy.

2.6 Lower-stakes, consultative accountability systems distort less than punitive opaque systems

The fact that people adapt to metrics does not mean all accountability regimes fail equally. Elston's work on low-stakes accountability argues that feedback and standard-setting can improve performance even without sanction-heavy regimes, because they redirect attention and control efforts toward deficits that were previously ignored. The OECD evidence similarly suggests that consultation and worker involvement reduce the trust and autonomy damage associated with algorithmic systems.

This is good news for Kashi. The project is stronger when it stays in a bounded, assistive, conversation-starting lane rather than sliding into ranking, punishment, or formal HR decision support. The more Kashi becomes a high-stakes evaluative instrument, the stronger the incentive to game it. The more it remains a

constrained governance and reflection system, the more likely it is to reveal useful structure without contaminating the signal too badly.

3. What adversarial adaptation is likely to look like in Kashi

The table below separates direct empirical support from stronger design inference. Not every row has been empirically tested in a Kashi-like product. That is normal. The goal here is not false precision; it is to anticipate the most plausible routing-around behaviors before they show up in deployment.

Adversarial route	What it may look like	What Kashi should do
Metric substitution	Interruptions fall, but unanswered questions, selective non-response, topic-closure control, or ignored-turn capture rise.	Read improvement only at the multi-metric level. A “cleaner” interruption metric should be treated as provisional unless adjacent metrics improve too.
Channel displacement	Group meetings look safer, but pressure moves to 1:1s, off-calendar calls, corridor talk, private chats, or delegated feedback sessions.	State this explicitly as a limitation. Keep worker-controlled evidence tools and escalation pathways because the observed channel will not contain all harm.
Hierarchical laundering	Senior leaders keep their own visible metrics clean while leads or deputies execute exclusion or aggression.	Do not over-index on clean senior dashboards. Examine dyads, chains of interaction, and where negative patterns cluster after leadership feedback cycles.
Polite structural exclusion	No obvious overlap or harsh language, but input is time-boxed, summarized away, deferred, or never substantively answered.	Continue investing in unanswered-question, topic-credit, and sequence-aware proxies. Do not market interruption decline as the main sign of health.
Symbolic compliance	Users learn the detector language and perform to it: fewer overtly risky moves, same underlying politics.	Add an adaptation-audit layer after deployment. Compare metric improvement to broader outcomes like silence, churn, retaliation reports, and escalation behavior.
Role camouflage	Agenda control is justified as “chair duty,” “time discipline,” or “keeping us on track,” even when participation is selectively narrowed.	Treat role context as a confound, not a free pass. Combine role labels with differential treatment patterns and baseline shifts.

4. What Kashi already gets right

- It treats a single meeting as weak evidence and patterns over time as stronger evidence.
- It calibrates behavior against each speaker’s own baseline rather than against a crude team average.
- It stays structural instead of drifting into emotion, affect, voice-stress, or psychologizing claims.
- It refuses performance, promotion, discipline, and compensation use.
- It already rejects the company-wide “relationship health” bar and acknowledges that visible numbers get falsified once they become targets.
- Its victim-centered roadmap items — especially the explainer page and worker-owned encrypted evidence vault — are directionally aligned with off-channel risk and contested interpretation.

5. Where Kashi is still exposed

The strongest current exposure is interpretive rather than computational. Kashi can already compute several useful structural patterns. The bigger danger is over-reading apparently cleaner signals once the system becomes socially known.

- If interruptions fall but other exclusion patterns remain stable, Kashi needs language and logic that treats this as possible adaptation, not automatic improvement.
- Kashi's current materials are strong on what the system refuses, but weaker on explicitly naming how pressure can migrate into uninstrumented channels.
- The deck and governance materials should speak more directly about responsibility laundering: a clean senior mirror does not guarantee a clean team dynamic.
- The product still needs a sharper "interpret with caution" rule for polite exclusion, sequencing, and agenda control. These are exactly the cases that look professional on the surface while remaining structurally suppressive.
- Kashi should avoid implying that people who know they are being measured will continue behaving as if the detector does not exist. That assumption is too clean to survive contact with real organizations.

6. Required project decisions

Priority	Decision	Why
P0 — must add now	Add an explicit "How actors may route around the system" section to the deck / concept note / governance page.	This closes the realism gap. Without it, the idea sounds cleaner than the evidence allows.
P0 — must add now	Adopt a metric-interpretation rule: improvement in one metric is provisional unless corroborated by adjacent metrics and broader context.	Prevents overclaiming and reduces false reassurance.
P0 — must add now	Add a plain-language limitation: absence of in-meeting signal does not prove absence of pressure outside meetings.	This is essential claim discipline and strengthens trust.
P0 — must add now	Keep Kashi low-stakes: no ranking, no punitive manager leaderboard, no silent HR decision support.	The research supports bounded, consultative accountability much more strongly than sanction-heavy use.
P1 — build into product logic	Create an "adaptation watch" layer that looks for suspiciously clean improvement in one metric while related signals stay flat or worsen.	This is the product analogue of red-team thinking.
P1 — build into governance	Add deployment guidance requiring worker consultation, manager training, and interpretation guardrails before rollout.	OECD evidence suggests consultation materially affects trust and acceptance.
P2 — future exploration	Investigate safe proxies for sequencing / agenda-control risk without drifting into content-heavy overreach.	Useful, but should be approached carefully so Kashi does not lose its structural-defensibility advantage.

7. Claim discipline: what Kashi should and should not say

This is the easiest place to make the concept stronger immediately. The goal is not to sound modest for its own sake. The goal is to make claims that remain true after users adapt.

Claim type	Recommended wording
Safe claim	Kashi surfaces structural risk patterns in meeting interaction over time.
Safe claim	Kashi helps institutions notice patterns they would otherwise explain away one meeting at a time.
Safe claim	Kashi is designed for bounded visibility and human review, not automated judgment.
Unsafe claim	If the measured pattern improves, the underlying behavior has improved.
Unsafe claim	If Kashi sees no problem in meetings, the team is likely fine.
Unsafe claim	Clean dashboards mean the relevant actors are behaving well.
Unsafe claim	Kashi can eliminate power abuse in communication by making it visible.

8. Paste-ready section for Kashi materials

Suggested section title
How people may route around the system

Because Kashi measures structural meeting behavior, we assume that some actors will adapt once they know what is visible. That adaptation does not need to look like crude cheating. It can take the form of metric substitution, channel displacement, hierarchical laundering, or symbolic compliance.

For example, interruptions may decrease while exclusion persists through agenda control, selective non-response, takeover after proposals, or delayed acknowledgment. Pressure may also move out of monitored meetings into 1:1s, private chats, corridor conversations, or delegated feedback through lower-level actors. A cleaner measured pattern therefore does not automatically mean a cleaner underlying dynamic.

Kashi is designed with this in mind. We treat single events as weak evidence, patterns over time as stronger evidence, and any improvement in one metric as provisional unless it is supported by broader pattern movement. We also treat absence of in-meeting signal as only that: absence in the observed channel, not proof that no pressure exists elsewhere.

This is why Kashi remains bounded by design. It is a governance and reflection system, not a punitive scoring machine. The more a system is tied to sanctions, rankings, or silent HR decisions, the greater the incentive to game it. Kashi is strongest when it supports accountable human review without pretending that observed metrics are identical to the full social reality of a team.

9. Final judgment

The adversarial / gaming perspective does not weaken the Kashi concept. It actually makes the concept more credible. A mature governance product does not assume that visibility produces honesty in a straight line. It assumes that people react to visibility, and then designs around that reaction.

So the right move is not to promise that Kashi reveals the full truth of workplace power. The right move is to say that Kashi surfaces structural risk patterns in one important channel, while explicitly accounting for adaptation, displacement, and contested interpretation. That version is more realistic, more trustworthy, and ultimately more defensible.

Selected references

Internal project materials

Kashi. (2026, April 21). Kashi — Progress & Project Overview. Internal project document.

Transparency That Drives Institutional Accountability. Detailed concept proposal. Internal project concept note.

External sources

Burris, E. R. (2012). The risks and rewards of speaking up: Managerial responses to employee voice. Academy of Management Journal, 55(4), 851–875. https://journals.aom.org/doi/10.5465/amj.2010.0562

Elston, T., Dixon, R., & Yang, J. (2025). Low-stakes accountability and public service turnarounds. Journal of Public Administration Research and Theory, 36(1), 19–35. https://academic.oup.com/jpart/article/36/1/19/8297176

Larkin, I. (2014). Employee gaming in enterprise software sales. Journal of Labor Economics, 32(2), 319–351. Harvard Business School working-paper version used here: https://www.hbs.edu/ris/Publication%20Files/13-073_cbb24c28-9e84-47d9-8a32-f01b73cfda13.pdf

Lewis, J. M. (2015). The politics and consequences of performance measurement. Policy and Society, 34(1), 1–12. https://academic.oup.com/policyandsociety/article/34/1/1/6401372

Mawritz, M. B., Mayer, D. M., Hoobler, J. M., Wayne, S. J., & Marinova, S. V. (2012). A trickle-down model of abusive supervision. Personnel Psychology, 65(2), 325–357. https://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2012.01246.x

Månsson, L. (2015). Pre-closing formulations in meeting talk. https://www.diva-portal.org/smash/get/diva2%3A823684/FULLTEXT01.pdf

OECD. (2025). Algorithmic management in the workplace. OECD Artificial Intelligence Papers, No. 31. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/algorithmic-management-in-the-workplace_3c84ed6d/287c13c4-en.pdf

Treem, J. W., et al. (2023). Signaling and meaning in organizational analytics. Journal of Computer-Mediated Communication, 28(4). https://academic.oup.com/jcmc/article/28/4/zmad023/7210220

Yu, Z., Qiu, J. L., & Chen, W. (2022). The emergence of algorithmic solidarity: Unveiling mutual aid practices and resistance among Chinese delivery workers. https://orca.cardiff.ac.uk/id/eprint/146523/8/1329878x221074793.pdf