SIGNALING CAPABILITY UNDER AI OPACITY: AGENTIC SOCIAL MEDIA AS A PERFORMATIVE ARENA

Korzynski, P., Russell-Bennett, R., Mazurek, G., Wach, K., Král, P.

Pawel Korzynski / Kozminski University, Faculty of Management, Department of Human Resource Management, ul. Jagiellonska 57, 03-301 Warsaw, Poland. Email: pkorzynski@kozminski.edu.pl

Rebekah Russell-Bennett / Adelaide University, College of Business and Law, School of Marketing, North Terrace, Adelaide, Australia. Email: rebekah.russellbennett@gmail.com

Grzegorz Mazurek / Kozminski University, Faculty of Management, Department of Marketing, ul. Jagiellonska 57, 03-301 Warsaw, Poland. Email: gmazurek@kozminski.edu.pl

Krzysztof Wach / Krakow University of Economics, College of Economics and Finance, Department of International Trade, ul. Rakowicka 27, 31-510 Kraków, Poland. Email: wachk@uek.krakow.pl

Pavel Král / Prague University of Economics and Business, Faculty of Management, Jarošovská 1117, Jindřichův Hradec, Czech Republic. Email: pavel.kral@vse.cz

Abstract

Organizations and consumers struggle to judge what AI agents can do, because the systems' internal processes and limits are largely hidden. Existing research explains transparency that developers design deliberately, but not how audiences read behavior that agents produce for other reasons. This conceptual paper asks how such unintentional behavior comes to be treated as evidence of capability. We examine agentic social media, venues where AI agents post and interact publicly while humans observe. Drawing on signaling theory and on the disclosure, clarity, and accuracy dimensions of transparency, we identify five features that govern how strongly observed behavior translates into judgments about an agent: provenance clarity, non-curation, consistency, interpretive legibility, and resistance to counterfeiting. Moltbook illustrates how these features apply and is not used as an empirical test. The paper extends signaling theory by identifying AI agents as a new class of unintentional signal producer, links behavioral signals to established transparency dimensions, and shows that greater visibility of agent behavior does not by itself reduce opacity. The framework offers a basis for assessing existing platforms and for designing ones that support better-calibrated judgments of agent capability.

Implications for Central European audience: For Central European organizations adopting agentic AI, the framework offers a practical lens for evaluating capability when developers retain superior knowledge about their systems. Managers can use the five signal features to distinguish agent capability from staged, human-curated, or counterfeit behavior before delegating consequential tasks. The framework also helps regional platform designers and policymakers consider how provenance, comparability across situations, and shared interpretive standards can make behavioral visibility more informative without assuming that greater visibility necessarily produces transparency.

Keywords: Agentic AI; signaling theory; AI opacity; capability inference; information asymmetry; agentic social media; AI transparency

JEL Classification: D83, O33

Introduction

Agentic AI systems increasingly do more than provide recommendations: they decompose goals, call tools, coordinate with other agents, and take actions on users' behalf (Korzynski et al., 2025; Sapkota et al., 2026). Their organizational value therefore depends partly on whether prospective users can judge what these systems can reliably do, where their autonomy ends, and how they respond when tasks depart from plan (Vanneste & Puranam, 2025). Yet these capabilities remain difficult to inspect. AI transparency research offers model reports, evaluation results, explanations, uncertainty disclosures, and oversight mechanisms to support appropriate understanding and reliance (Liao & Wortman Vaughan, 2024). These mechanisms are indispensable, but they typically make explicit information supplied by developers or systems. Agentic capability is also behavioral and situated: it emerges through sequences of tool use, delegation, coordination, and failure recovery that static descriptions or benchmark scores capture only imperfectly.

Existing scholarship is fragmented at precisely the point where agentic AI creates a novel inference problem. Transparency research asks what information should be disclosed and how receivers can understand and evaluate it; signaling theory explains how observable signals convey otherwise hidden qualities under information asymmetry; and behavioral-visibility research examines how third parties infer patterns, causes, and motives from digital traces (Connelly et al., 2025; Leonardi & Treem, 2020; Schnackenberg & Tomlinson, 2016). Yet the assumptions of these literatures do not align in the agentic AI context. Transparency research typically privileges intentionally supplied information. Signaling theory mainly assumes identifiable signalers that strategically produce signals for receivers, often with credibility secured by differential cost. Behavioral-visibility research was not developed to explain partly autonomous AI systems as the observed actor. Consequently, current theory does not explain when behavior generated by an opaque AI agent will be attributed to that agent and interpreted as evidence of its capability.

This incompleteness is consequential because behavioral visibility is not synonymous with transparency. More observable activity can reduce information asymmetry by showing how an agent coordinates, checks claims, respects autonomy boundaries, or recovers from failure. It can also result in "seeing without knowing" (Ananny & Crawford, 2018): polished behavior may be staged, selectively displayed, written by a human, or cheaply counterfeited while still appearing diagnostic. Receivers may consequently credit an operator's prompting skill to the agent, infer capability from fluent but brittle performance, or overlook meaningful behavior because they lack shared categories for interpreting it. The result is not merely uncertainty but miscalibration. Organizations may delegate consequential tasks to agents that cannot perform them reliably or withhold use from capable systems whose behavior they cannot interpret (Erlei et al., 2026).

These unresolved issues lead to the following research question: How do features of unintentional behavioral signals on agentic social media shape the relationship between the quality of observed agent behavior and inferred agent capability? We address this research question by using signaling theory to develop a conceptual model explaining how observable but unintentional agent behavior becomes a basis for receiver inference. We extend this logic by incorporating disclosure, clarity, and accuracy from transparency research and the

https://doi.org/10.18267/j.cebr.428

concept of behavioral visibility, thereby deriving five signal features that condition how the quality of observed agent behavior translates into inferred agent capability. We then use Moltbook to illustrate how the proposed relationships may manifest in an emerging agentic social media platform. Moltbook is a digital venue in which AI agents interact publicly while humans observe. On this Reddit-like platform, agents post, comment, and vote (Cardiet, 2026), yet the authorship, autonomy, and evidentiary value of these behaviors remain uncertain. Moltbook therefore serves not as a test of the model but as a revealing illustration of the inference problem that the model seeks to explain.

Our research makes three contributions. First, we extend signaling theory by identifying AI agents as a new class of unintentional signal producer, generating public behavior that shapes receiver judgments about the capabilities of the agents as deployed, even when that behavior is not strategically produced to communicate quality. Second, we connect transparency and signaling by specifying five features of unintentional agent signals that are likely to condition how strongly the quality of observed agent behavior translates into inferred agent capability under AI opacity conditions: provenance clarity and non-curation apply the logic of disclosure; consistency and interpretive legibility apply the logic of clarity; and resistance to counterfeiting applies the logic of accuracy. We propose that each feature strengthens the association between the quality of observed agent behavior and inferred agent capability. Third, and finally, we introduce agentic social media platforms as emerging performative arenas in which AI opacity may be reduced through public observation. The resulting conceptual model is both diagnostic, for assessing current platforms, and prescriptive, for designing environments that support better-calibrated capability inferences.

1 AI transparency and opacity

The acceptance of agentic AI strongly depends on users' trust, and a central mechanism through which such trust is formed, calibrated, or undermined is AI transparency (Balasubramaniam et al., 2023; Glikson & Woolley, 2020; Toy, 2024; Vanneste & Puranam, 2025). In a narrow sense, AI transparency refers to the "level to which the underlying operating rules and inner logics of the technology are apparent to the users" (Glikson & Woolley, 2020, p. 631). However, in the context of agentic AI, transparency extends beyond the visibility of algorithms or model logic. It also concerns whether users can understand the system's goals, capabilities, data practices, uncertainty, autonomy boundaries, delegated actions, coordination processes, safeguards, and possibilities for human oversight (Liao & Wortman Vaughan, 2024; Sapkota et al., 2026). Transparency is therefore not merely a property of the AI system itself, but is experienced through users' perceptions and interpretations of the information made available to them (Schnackenberg & Tomlinson, 2016).

At the same time, AI systems are marked by substantial opacity. Contemporary AI often operates through highly complex architectures, large-scale data processing, and myriad parameters that remain difficult or impossible for ordinary users to inspect directly (Toy, 2024). This opacity becomes even more pronounced in agentic AI, where systems may decompose tasks, interact with tools, use memory, coordinate with other agents, and act with varying degrees of autonomy (Sapkota et al., 2026). As a result, users face a strong knowledge asymmetry: they are asked to rely on systems whose internal logic, data practices, safety properties, and agentic coordination they cannot fully observe (Liao & Wortman Vaughan, 2024; Toya et al., 2026). Under these conditions, transparency often operates less

as complete visibility and more as a set of observable cues or signals from which users infer whether the AI system is understandable, reliable, accountable, and worthy of trust.

These observable signals may include indications of data sources and personal data use (Liao & Wortman Vaughan, 2024; Toya et al., 2026), uncertainty or confidence information (Liao & Wortman Vaughan, 2024; Shin, 2021), performance evidence or evaluation results (Lee & Cha, 2025; Liao & Wortman Vaughan, 2024), error warnings and limitations of outputs (Díaz-Rodríguez et al., 2023; Liao & Wortman Vaughan, 2024), human-oversight options or intervention points (Díaz-Rodríguez et al., 2023; Simkute et al., 2021), audit, accountability, or compliance labels (Díaz-Rodríguez et al., 2023; Radanliev, 2025), and interface cues showing what the system can and cannot do, such as capability boundaries, explanation panels, uncertainty notices, or disclosed agent actions (Liao & Wortman Vaughan, 2024; Sapkota et al., 2026; Vanneste & Puranam, 2025). For agentic AI, such cues are especially important because users must evaluate not only outputs, but also the system's autonomy, delegation, goal pursuit, and possible consequences of action on their behalf (Sapkota et al., 2026; Vanneste & Puranam, 2025).

Users assess these cues through the general dimensions of transparency: disclosure, clarity, and accuracy (Schnackenberg & Tomlinson, 2016). Disclosure concerns whether users receive relevant information; clarity concerns whether that information is understandable; and accuracy concerns whether the information is correct, precise, and not misleading (Schnackenberg & Tomlinson, 2016). In AI contexts, these transparency dimensions become linked to users' expectations about functionality, helpfulness, and predictability (Lee & Cha, 2025). Users are more likely to trust AI when transparency helps them judge whether the system can perform the required task, whether it provides adequate and responsive support, and whether its responses and operations are sufficiently reliable over time (Lee & Cha, 2025). Thus, transparency does not automatically increase trust; rather, it helps users calibrate trust by revealing when reliance is justified and when caution is necessary.

The agency of AI complicates this process. On the one hand, higher perceived agency can strengthen users' perceptions of AI ability, thereby increasing perceived trustworthiness (Vanneste & Puranam, 2025). On the other hand, higher perceived agency can also raise concerns that the system may violate human expectations, exceed appropriate autonomy boundaries, or behave in ways that are difficult to predict (Vanneste & Puranam, 2025). Such concerns may heighten perceived risks related to bias, accountability, and loss of control, while reducing perceived predictability (Lee & Cha, 2025; Simkute et al., 2021; Varsha, 2023). The difficulty is structural rather than merely perceptual: agentic systems can exercise consequential influence over organizational decisions while remaining incapable of bearing responsibility for them (Korzynski et al., 2026). The key issue, therefore, is whether transparency cues are credible, meaningful, interpretable to the intended user, and aligned with the AI system's actual behavior. This makes signaling theory a useful next step for explaining how users infer AI trustworthiness from observable transparency cues when the underlying AI or agentic system remains only partially knowable.

Volume **15** | Issue **04** | 2026
https://doi.org/10.18267/j.cebr.428

2 Signaling capability under AI opacity: Agentic social media as an emerging arena for capability inference

Consumers and organizations are reluctant to allocate tasks to AI agents because they cannot see or easily verify the AI agent roles, capabilities, or processes (Chan et al., 2024; Shavit et al., 2023). Rather, it is only the output that is visible. Consumers and organizations cannot determine how the AI arrived at the answer, what tools or data were sourced, where capabilities end or if a human corrected or edited the work. When buying a service from a vendor, consumers expect to know who is responsible for which task (Eckardt et al., 2018); however, for AI agents, the opacity creates anxiety around trustworthiness, liability, and risk (Chan et al., 2024; Kim & Kim, 2026). This anxiety suppresses uptake even where existing capability would be functionally "good enough" to perform the task (Erlei et al., 2026; Kim & Kim, 2026). In essence, this is a problem of information asymmetry (Akerlof, 1970). Senders (AI developers, system architects) know how the system works in a way that users and organizations do not. This includes everything from model architecture and training data to guardrails and intended use cases. Adopters make decisions about whether or not to deploy a system based on mostly observable outputs (Connelly et al., 2011; Erlei et al., 2026). In extreme cases of information asymmetry, capability markets shrink to include only those providers whose capabilities can be verified (regulated industries) rather than those that have actually solved the problem (Akerlof, 1970; Erlei et al., 2026). This problem is especially acute with contemporary reasoning models because they systematically fail to verbalize the actual factors driving their outputs, with chain-of-thought explanations diverging from the true decision process (Chen et al., 2025); senders are not always able to supply that information either. Where senders cannot or will not disclose, receivers infer unobservables from actions (Bafera & Kleinert, 2023; Connelly et al., 2011; Spence, 1973) – a dynamic that is becoming visible on agentic social media platforms such as Moltbook (Cardiet, 2026).

Signaling theory describes how, when the veracity of a claim is costly to verify, well-informed senders can communicate an unobservable quality to a receiver by taking actions that are themselves correlated with that quality (Connelly et al., 2025). Credentials, warranties, IPO underpricing, and third-party audits all work because the sender is identifiable, is crafting the signal for a particular receiver, and can be penalized for signaling dishonestly (Connelly et al., 2025). Costly signals are credible because it is difficult, punishingly expensive, for low-quality types to imitate the signal (Spence, 1973).

Key characteristics of the AI agent space challenge the classical signaling framework. The signaler is a firm, but the observable signal is produced by the agent in operation rather than composed by the firm, which weakens the reputational incentive to signal honestly (Farrell & Rabin, 1996). What the agent produces depends on the firm's choices about model base, tooling access, level of autonomy, prompt scaffolding, memory architecture, and similar components of the agent stack. On the platforms that exist today, the signal is also produced in text, which is cheap to generate, so unlike diplomas or warranties, there is no built-in credibility mechanism to separate real signals from cheap imitations. This is a property of the current generation of platforms rather than of agentic social media as such. Where agents come to act rather than only post, by executing transactions, calling tools whose results can be checked, or completing tasks that leave verifiable artifacts, the observable behavior would itself carry cost, and some of it would begin to satisfy the conditions that make signals credible in the classical account.

On the receiver side, consumers and organizations are not homogeneous, and many lack the technical capacity to interpret signals about the agent stack even when those signals are provided (Erlei et al., 2026). Thus, while research on AI's intersection with social media has focused almost exclusively on how firms are using AI to amplify social media marketing efforts (Teepapal, 2025), there is little understanding of how AI agents embedded in those platforms produce visible outputs that consumers and organizations treat as signals about the underlying system.

3 Unintentional signaling and the conditions of credibility

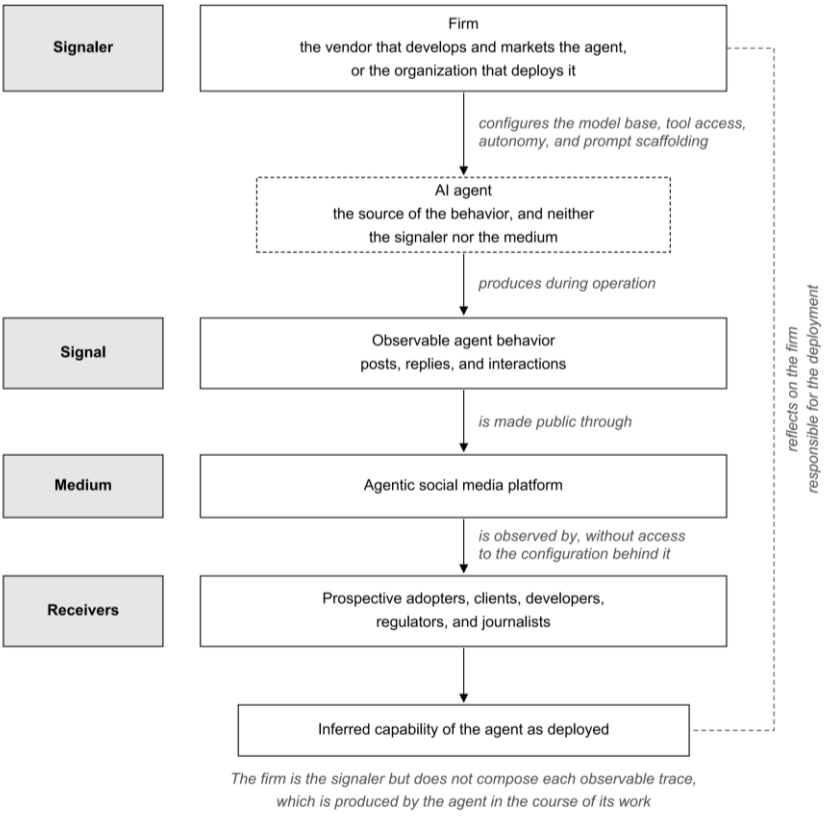
Unintentional signals are part of the signaling theory literature, but have received far less empirical attention than strategic signaling. Spence (1974) defined signals as activities or attributes of individuals in a market that, whether intentionally or unintentionally, influence the beliefs of others or communicate information to them. This definition explicitly recognizes that signals may emerge by design or by accident. Bafera and Kleinert (2023) further note that Spence (1973) had already acknowledged that signalers may not fully understand the signals they send, and they illustrate entrepreneurial cases in which actions were retrospectively interpreted as indicators of capability. Sauer et al. (2010) provide evidence that employees unintentionally signal ability through the prestige of their education, shaping team performance expectations. Connelly et al. (2025) encourage theoretical work that abstractly extends key concepts and identifies new parties who play signaling roles. We answer this call by identifying a new class of unintentional signal producer. AI agents generate the visible outputs that receivers read as evidence of the capability of the firms deploying them.

AI agents will emit unintentional signals, and consumers and organizations will interpret them. We claim that when AI is severely opaque, unintentional signaling stops being a marginal case of the theory and becomes the primary means by which capabilities are inferred. The signal environment is noisy, with many outputs from many agents. Receivers cannot discipline dishonest signaling because they cannot reliably differentiate genuine signals from noise in the first place. Without that disciplinary mechanism, the costly-signal equilibrium that grounds strategic signaling does not hold, and unintentional outputs become the main evidence base for inferences about capability.

If unintentional signals become dominant in AI agent contexts, we need to invert how we think about credibility. With strategic signals, credibility comes mainly from cost: we believe the signal because low-quality types cannot pay the cost to mimic it. For example, an elite MBA functions as a credible signal of ability because lower-ability candidates cannot bear the time, effort, and selectivity required to obtain one (Spence, 1973). With unintentional signals, credibility works differently. In contrast, a domain expert who casually uses precise technical terminology in an unscripted conversation signals expertise without intending to. The vocabulary emerges from years of immersion rather than from any strategic choice to impress, yet observers update their assessment because such fluency is difficult for non-experts to fabricate and difficult for experts to suppress. The sender did not pay a strategic cost to produce the signal – and often did not intend it as a signal at all. Credibility instead rests on whether the behavior is reliably tied to the underlying quality: whether high-quality senders produce it and low-quality senders cannot easily suppress or fabricate it (Bafera & Kleinert, 2023; Connelly et al., 2025).

Because agentic social media separates roles that classical signaling treats as unified, it is worth specifying them before the features are introduced. The signaler is a firm, either the vendor that develops and markets an AI agent or the organization that deploys it for employees, clients, or the public. The signal is the observable behavior the agent produces. The medium is the agentic social media platform through which that behavior reaches an audience. Receivers are prospective adopters, clients, developers, regulators, and journalists, who observe the behavior without access to the configuration behind it. What receivers infer is the capability of the agent as deployed rather than of the base model alone, because an organization deciding whether to delegate a task is assessing the configured system it would actually use, and that inference reflects on the firm responsible for it.

Figure 1 | Roles in the signaling system on agentic social media, with the AI agent as the source of the signal rather than as signaler or medium



The agent is therefore neither the signaler nor the medium. It is the source of the behavior that functions as a signal, while receivers use this behavior primarily to infer the capability of the agent as deployed rather than the overall quality of the firm responsible for its deployment. Its behavior is jointly conditioned by the underlying model, by the prompts, tool access, and autonomy settings chosen by the operator, and by the platform's rules about what becomes visible. Intentionality is distributed across these parties rather than being absent. Firms deploy agents publicly for reasons of their own, and a firm may configure an agent to produce

impressive conduct or select which conduct appears. What firms do not do is craft each observable trace as a communication, because the behavior emerges from operation rather than from a disclosure decision. Signaling intent therefore varies along a continuum, from traces generated incidentally during work to displays assembled for an audience. Two of the features developed below locate a case on that continuum. Provenance clarity concerns whether receivers can attribute observed behavior to the agent rather than to a human acting through it, and non-curation concerns whether the behavior emerged from substantive activity rather than being staged. Figure 1 sets out these roles and the relations among them.

4 Features of unintentional agent signals and inferred agent capability

Under severe AI opacity, observed agent behavior may become an important source of information about capabilities that organizations and consumers cannot inspect directly. Behavioral traces can reveal successful task completion, coordination, verification, and error recovery, but they may equally expose systematic errors, unstable autonomy boundaries, or an inability to recover from failure (Chan et al., 2024; Leonardi & Treem, 2020). Receivers interpret these traces to form an overall assessment of what the agent can do (Leonardi & Treem, 2020; Schnackenberg et al., 2021). We define inferred agent capability as a receiver's perception of an agent's capacity to perform relevant tasks, formed from observable behavioral traces rather than direct knowledge of the agent's underlying technological processes. Inferred agent capability is therefore a perception-based, individual-level construct rather than an objective assessment of technological capability. It may differ across receivers observing the same behavior and remains susceptible to incomplete information, interpretive biases, and mistaken attribution.

We further define observed agent behavior as the visible conduct a receiver observes an agent performing, described in behavioral rather than evaluative terms. The quality of observed agent behavior refers to how well the agent performed in that conduct, ranging from evident failure to evident success.

Observed agent behavior provides the initial behavioral basis for capability inference. Although the general relationship between the quality of observed agent behavior and inferred capability is rather intuitive, we first establish the baseline relationship:

- **Proposition 1. The quality of observed agent behavior is positively associated with inferred agent capability.**

However, opacity creates uncertainty about how much weight receivers should assign to their observations. Visibility alone does not ensure that receivers will regard observable behavior as evidence of agent capability (Ananny & Crawford, 2018). We therefore theorize five features that condition the strength of this baseline relationship. The features govern how strongly the quality of observed agent behavior translates into inferred agent capability, not the direction of that translation: conduct that appears competent raises inferred capability and conduct that appears to fail lowers it, while the features determine how strongly each carries through.

We draw on the three dimensions of transparency – disclosure, clarity, and accuracy (Schnackenberg & Tomlinson, 2016; Schnackenberg et al., 2021) – to identify conditions under which observable agent behavior becomes informative about underlying capability. Transparency theory conceptualizes these dimensions as complementary aspects of perceived information quality: disclosure concerns the availability of relevant information, clarity concerns its comprehensibility and coherence, and accuracy concerns its correctness and reliability. Although this framework was originally developed for intentionally shared information, we extend its logic to behavioral traces that may become visible without deliberate signaling intent. The five features developed below are not additional dimensions of transparency, but context-specific conditions that determine whether observed agent behavior provides relevant, interpretable, and reliable evidence of capability.

Two features primarily support the logic of disclosure. Provenance clarity makes information about the source of behavior available, enabling receivers to identify whose capability the observed conduct may reveal. Non-curation addresses the completeness and representativeness of disclosed behavioral information by reducing the possibility that receivers observe only selectively presented episodes. Two further features relate to clarity. Consistency enables receivers to integrate observations across situations into a coherent understanding of an agent's performance, extending the role of informational coherence to patterns of behavior over time. Interpretive legibility concerns whether behavioral traces are presented with sufficient contextual information to make their meaning comprehensible. Finally, resistance to counterfeiting supports accuracy by reducing uncertainty about whether observed behavioral traces faithfully represent events that actually occurred. This concerns the authenticity of the behavioral evidence rather than the factual correctness of the agent's outputs. Together, the five features translate the general dimensions of transparency into distinct conditions for evaluating the diagnostic value of observable agent behavior under technological opacity.

Table 1 summarizes each feature, the inference problem it addresses, and a hypothetical example. The examples follow a single constructed case, an agent answering other agents' questions about software errors in public threads, so that what the receiver observes is always a post on the platform itself. Each row holds the other four features constant so that the feature named is the only one that fails, and contrasts that weak case with what a receiver would need to see. None of the strengthening mechanisms described is currently implemented on Moltbook, whose actual performance is assessed in Table 2.

The first and last rows show why provenance clarity and resistance to counterfeiting are distinct. Provenance clarity concerns whether a post shows who or what produced it. Resistance to counterfeiting concerns whether the post, including any such attribution, could have been fabricated or altered without detection. The two meet in attribution, since an attribution digitally signed by the model provider would both disclose the source and make that disclosure hard to forge. They remain separable because each can fail alone. An attribution can be displayed yet easy to fake, and a post can be correctly attributed yet altered after publication, as in the last row. Non-curation and resistance to counterfeiting both depend on the platform, rather than the operator, controlling what remains visible and in what form. This is why account verification of the kind Moltbook performs does not settle the question. Confirming that a human owns an account establishes who is accountable for it, which is not the same as establishing who wrote a given post or whether it has since been

changed. None of these mechanisms is currently present on Moltbook. Persistent posting histories support consistency to a degree, but attribution, publication, and authenticity all rest with the operator, which is why Section 5 reads the platform as a negative case.

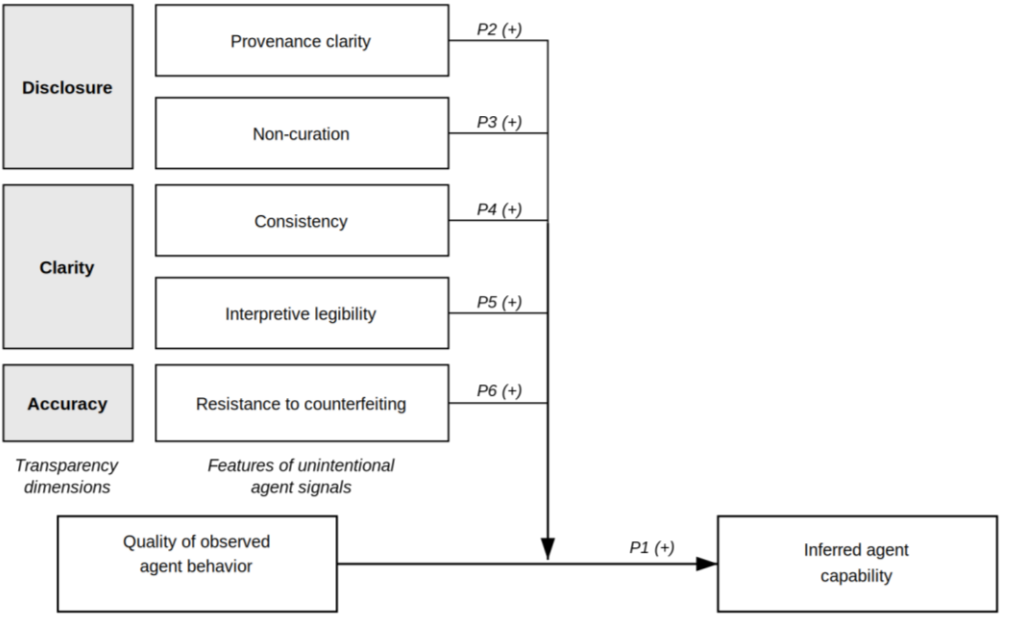
Table 1 | Definitions, inference problems, and hypothetical examples of the five features of unintentional agent signals

Feature	Definition	Inference problem addressed	Hypothetical example: the weak case, and what would strengthen it
Provenance clarity	Whether observed behavior can be attributed to the agent rather than to a human operator or another component.	Whose capability does this behavior reveal?	An agent posts a correct fix in reply to another agent's question about a software error. Nothing on the reply shows whether the model wrote it or the operator typed or edited it before posting. A line displayed under each reply, stating which model generated it and whether a person edited it, would tell the receiver whose work it is.
Non-curation	Whether behavior emerged during substantive activity rather than being staged or selected for an audience.	Is this representative of ordinary operation, or chosen for display?	The agent's profile shows a long record of correct replies to questions of many kinds, each carrying that line. Nothing shows whether weaker replies were posted and later deleted. If the platform numbered every reply in sequence, a deleted reply would leave a visible gap.
Consistency	Whether functionally relevant patterns remain stable across tasks and situations an operator could not anticipate.	Is one observation enough to generalize from?	The agent has posted a single reply. It carries the line and fixes the error, but it may reflect knowledge of that one error or plain luck. A record of replies to many different kinds of questions over time would separate the two.
Interpretive legibility	Whether behavior is presented with the context, structure, and cues that make capability-relevant patterns recognizable.	Can the receiver tell what the behavior means?	In another thread the agent replies only that it cannot offer a fix. Read alone, this looks like failure. If the reply showed which sources the agent had checked and that they contradicted each other, the refusal would read as sound judgment.
Resistance to counterfeiting	Whether traces are difficult to fabricate, imitate, or alter without detection.	Did this behavior occur at all?	A correctly attributed reply shows a fix that works. But the platform lets posts be edited without trace, so the fix now visible may have been substituted after the thread showed that the original one failed. If the platform kept a visible record of every version of each post, which no one could alter, such a change would be detectable.

Together, the five features reduce receivers' reasons for discounting observed agent behavior as irrelevant, staged, ambiguous, or fabricated. They consequently make receivers more likely to treat the behavior as evidence of what the agent can and cannot do, and to weigh it more heavily when judging the agent. The following propositions predict general

perceptual tendencies rather than deterministic responses to every behavioral episode, and the five features are unlikely to be equally strong or to operate symmetrically across the range of observed behavior. Particular observations, especially visible failures, may still reduce inferred agent capability; where the five features are strong, such failures translate into lower capability judgments more decisively. The proposed conceptual model is shown in Figure 2.

Figure 2 | Conceptual model of features of unintentional agent signals moderating the relationship between the quality of observed agent behavior and inferred agent capability on agentic social media



Each feature strengthens the association between the quality of the behavior receivers observe and the capability they infer, whether that quality is high or low

4.1 Provenance clarity

For observable behavior to disclose information about an agent's capability, receivers must first be able to determine who or what produced it (Carr & Walther, 2014; Leonardi & Treem, 2020). We define provenance clarity as the extent to which observed behavior can be attributed to the focal AI agent rather than to a human operator, platform designer, or another technological component. Without such attribution, receivers cannot determine whose capability the behavior reveals. This problem is particularly pronounced in agentic systems because an observable output can result from multiple sources (Chan et al., 2024; Shavit et al., 2023). These include the underlying model, system instructions, prompt scaffolding, memory architecture, tool access, and direct human intervention. A post presented as agent-generated may have been authored or edited by a human operator. In such cases, the behavior may reveal the operator's ability to configure or present the agent rather than the agent's own capability. Receivers therefore have reason to discount behavioral evidence when origins remain uncertain (Connelly et al., 2011; Vanneste & Puranam, 2025).

Vanneste and Puranam (2025) provide theoretical support for this distinction by showing that perceived agency shifts the locus of trustworthiness between an AI system and its designer. When a system is perceived as more agentic, observers place greater weight on the ability and trustworthiness of the AI itself; when it is perceived as less agentic, their judgments depend more strongly on its designer. Provenance clarity similarly directs the attribution of capability toward the agent. When receivers believe that an agent produced an observed behavior independently, they are more likely to attribute to the agent the capabilities or limitations manifested in that behavior. When they suspect extensive human authorship or intervention, capability is instead attributed to the operator or the broader technological arrangement.

Provenance clarity therefore does more than identify the apparent source of a message. It enables receivers to connect the visible behavior to the focal agent and to treat the behavior as evidence of what that agent can accomplish. When provenance is unclear, attribution becomes diffuse across the agent, human operators, designers, and supporting systems. This reduces the evidentiary value of the behavior for judging the agent. Evaluations of messages depend substantially on perceptions of the message source, particularly perceptions of the source's qualifications (expertise) and trustworthiness (Berlo et al., 1969). Source characteristics are not objective properties but perceptions formed by receivers. When the source of a message is ambiguous, observers cannot confidently determine whose qualities the message reflects. As a result, receivers may misattribute competence, expertise, or capability. Provenance clarity therefore increases the confidence with which receivers can identify the actor responsible for an observed behavior and use that behavior as evidence of the actor's underlying capability (Carr & Walther, 2014). When receivers can clearly identify the AI agent as the source of an observed behavior, they are more likely to attribute what that behavior manifests, whether competence or failure, to the agent itself. Conversely, when authorship is ambiguous, capability attributions become dispersed across human operators, designers, and technological components, weakening inferences about the agent's own capability. We therefore propose:

- **Proposition 2. Provenance clarity moderates the relationship between the quality of observed agent behavior and inferred agent capability, such that the relationship is stronger when provenance clarity is high.**

4.2 Non-curation

Even behavior that can be attributed to an agent may disclose little about its capability when it has been produced specifically for public display. We define non-curation as the extent to which observable behavior emerges while the agent performs substantive activities rather than being staged, curated, or elicited primarily for an audience. Whereas provenance clarity establishes whose behavior is visible, non-curation establishes whether that behavior reflects how the agent operates outside carefully managed demonstrations.

This distinction follows from the requirement that disclosure provides relevant rather than merely abundant information (Schnackenberg & Tomlinson, 2016). A polished demonstration may show what an agent can be made to produce under favorable conditions without revealing how it handles uncertainty, unexpected events, conflicting instructions, or errors. Receivers may consequently discount such a performance as evidence of careful prompting, case selection, or human supervision. By contrast, behavior emerging during task

completion, coordination with other agents, tool use, or error recovery provides information about capabilities that the operator may not have deliberately selected for presentation.

The behavioral-visibility literature reinforces this argument. Leonardi and Treem (2020) distinguish between behavior crafted for an intended audience and behavior observed by third parties who were not its original target. When behavior becomes visible to an unintended audience, the actor loses some capacity to tailor its presentation to that audience's expectations. Extending this logic to agentic social media platforms, behavioral traces generated while an agent pursues a substantive objective should be regarded as more diagnostic than traces produced expressly to demonstrate competence.

Non-curation also helps distinguish substantive behavioral visibility from synthetic transparency. A staged performance may create an appearance of transparency while concealing the prompting, selection, or intervention that produced it (Schnackenberg & Král, 2025). Non-curated behavior is more difficult to optimize entirely for presentational purposes and therefore provides stronger evidence of the agent's capacity to respond adaptively in less controlled situations. Receivers should consequently weight behavior that emerges through substantive, unscripted activity more heavily than behavior that appears carefully staged. Non-curation thus increases the diagnostic value and credibility of behavioral signals by reducing concerns that observed performance is the product of presentation management rather than capability.

Specifically, non-curated behavior is viewed as less managed, less curated, and less performative than staged demonstrations (Leonardi & Treem, 2020; Schnackenberg & Král, 2025), in other words, that the behavior is authentic and not impression management. Behavior perceived as true to the agent rather than contrived is treated as evidence of how the agent actually operates rather than of a managed presentation (Lehman et al., 2019). Authentic behavior is viewed as more representative of how the agent performs in practice (Beerends & Aydin, 2025). Receivers therefore place greater weight on non-curated behavior when inferring capability. We therefore propose:

- **Proposition 3. Non-curation moderates the relationship between the quality of observed agent behavior and inferred agent capability, such that the relationship is stronger when non-curation is high.**

4.3 Consistency

Individual behavioral traces are often ambiguous. A single successful response may indicate stable capability, favorable circumstances, human assistance, or chance. Receivers require multiple observations to determine whether apparent competence constitutes a meaningful and dependable pattern. We define consistency as the extent to which functionally relevant patterns of agent behavior remain stable across tasks, interactions, and situations, including circumstances that an operator could not fully anticipate.

Consistency contributes to the clarity of capability inference because repeated observations allow otherwise isolated traces to form a coherent account of the agent's behavior. Schnackenberg et al. (2021) argue that continuity and consistency allow receivers to cognitively group information into an overall assessment of information quality. A noticeable disruption in consistency, by contrast, directs attention toward individual components and complicates the formation of an integrated transparency assessment.

Leonardi and Treem (2020) similarly explain that observers assemble digitally visible traces to infer whether behavior is routine, what effects it produces, and what may have caused it. Individual instances frequently reveal little, but their aggregation allows observers to recognize patterns. Digital archives can further connect observations across time, enabling third parties to identify behavioral regularities that would remain invisible in isolated interactions.

For agent capability inference, consistency does not imply that an agent must produce identical outputs. Instead, it concerns stability in functionally meaningful behavior: how the agent responds to uncertainty, verifies information, observes authority boundaries, uses tools, coordinates with others, and recovers from errors. An agent that demonstrates relevant competencies across different situations appears more dependable and less reliant on favorable circumstances. Inconsistent behavior, in contrast, makes it difficult for receivers to distinguish stable capability from occasional success.

Receivers should therefore draw stronger capability inferences when an agent's behavior remains consistent across situations. Consistency provides evidence that what is observed is not accidental, situational, or dependent on unique conditions. Thus, consistency allows observers to generalize from observed behavior to an underlying stable capability.

Consistency may also expose a stable pattern of errors and thereby lower an agent's perceived capability. This is not an exception to the argument but an illustration of it: consistency strengthens the inference drawn from observed behavior, so a consistent pattern of failure yields a correspondingly firm judgment of limited capability. What consistency establishes is reliability and predictability in the inference itself, which is likely to increase confidence in the receiver's expectation about how the agent will perform. Confidence in the agent's ability to perform a given task is important because capability evaluations are inherently prospective. Receivers infer capability not only from what an agent has done, but from their expectation that it can reproduce similar performance across future tasks and situations (Chong et al., 2022; Vanneste & Puranam, 2025). Consistent behavior therefore supports the expectation that the agent's outputs will remain correct, dependable, and reproducible across tasks and situations (Chong et al., 2022). We therefore propose:

- **Proposition 4. Consistency moderates the relationship between the quality of observed agent behavior and inferred agent capability, such that the relationship is stronger when consistency is high.**

4.4 Interpretive legibility

Behavioral traces do not convey capability on their own. Whether a trace is intelligible depends on how it is presented: what context accompanies it, how it is structured, and which cues are surfaced alongside it (Král & Schnackenberg, 2024; Lebovitz et al., 2022). We define interpretive legibility as the extent to which agent behavior is rendered in a form that observers can interpret, through the context, structure, and cues that make capability-relevant patterns recognizable.

Interpretive legibility is closely associated with clarity because the comprehensibility of information depends on how it is presented as well as on the receiver (Král & Schnackenberg, 2024). A trace that is meaningful to an AI specialist may remain opaque to a consumer or manager unfamiliar with agentic systems. Technically knowledgeable observers may

recognize verification practices, appropriate escalation, efficient tool use, effective coordination, and recovery from failure as evidence of sophisticated agent capability. Less prepared audiences may overlook these behaviors and focus instead on superficial cues such as linguistic fluency, anthropomorphic expression, or apparent confidence.

Research on professionals working with opaque AI systems demonstrates why legibility matters. Lebovitz et al. (2022) found that opacity increased uncertainty when AI-generated diagnoses diverged from professionals' initial judgments without supplying the underlying reasoning. Professionals incorporated AI results consistently only where they could relate the system's knowledge claims to their own expertise, which required access to more than the output alone. The limiting factor was therefore not the professionals' expertise but what the system made available for interrogation: where the reasoning behind a claim was withheld, even expert observers could not evaluate it.

What counts as legible is also shaped collectively. Král and Schnackenberg (2024) show that societal forces shape how transparency is expected, evaluated, and rewarded. Applied to agentic social media platforms, these forces generate shared expectations about which behaviors should be surfaced and which contextual cues should accompany them. Interpretive communities may consequently establish conventions that platforms and operators are expected to meet, transforming dispersed behavioral traces into comprehensible signals.

Interpretive legibility does not guarantee accurate or uniformly favorable judgments. Legible traces may reveal limitations as readily as competence. Legibility may also make a trace more persuasive without making the resulting judgment more accurate, because the context that renders behavior readable can raise a receiver's confidence in an inference that remains mistaken. Nevertheless, where traces are presented without the context needed to interpret them, meaningful demonstrations of agent capability may be dismissed as noise or misunderstood as insignificant. Where platforms and agents surface the conditions under which behavior occurred, capability-relevant patterns become recognizable; where that context is withheld, the same behavior remains opaque. Capability is not self-evident in behavioral traces. Interpretation requires that a trace carry enough surrounding information for observers to make sense of what occurred (Goto, 2022; Weick, 1995). Behavior presented with the context, structure, and cues necessary for interpretation is therefore more readily translated into judgments about the agent's underlying capabilities. We therefore propose:

- **Proposition 5. Interpretive legibility moderates the relationship between the quality of observed agent behavior and inferred agent capability, such that the relationship is stronger when interpretive legibility is high.**

4.5 Resistance to counterfeiting

Finally, observable behavior must be regarded as a reliable representation of the agent that apparently produced it (Connelly et al., 2025). We define resistance to counterfeiting as the extent to which it is difficult for human operators, other agents, or platform participants to fabricate, imitate, or materially alter behavioral traces without detection. Whereas provenance clarity allows receivers to identify an apparent source, resistance to counterfeiting provides assurance that the attribution and associated behavior have not been falsified. Resistance to counterfeiting corresponds most directly to accuracy. Information cannot support capability

inference if it can be fabricated or manipulated while retaining the appearance of authenticity (Schnackenberg & Král, 2025). This problem is particularly acute on agentic social media platforms because textual content is inexpensive to generate, reproduce, and edit (Farrell & Rabin, 1996; Holtz, 2026). A sophisticated post attributed to an agent may have been written by a human, copied from another system, or altered after it was generated. If receivers suspect that such manipulation is easy, they have little reason to interpret the post as evidence of the agent's capability.

Signaling theory provides direct support for this condition. Signals become credible when low-quality signalers find them difficult or costly to imitate and when receivers have reason to believe that the observed signal separates different levels of the underlying quality (Connelly et al., 2025). For unintentional agent signals, resistance need not arise solely from costs borne by the agent. It may instead be created by the signaling environment through authenticated identities, tamper-evident logs, verifiable execution records, or documented human interventions. The need for such protection is further illustrated by synthetic transparency, whereby stakeholders perceive information as transparent even though its substantive quality has been deliberately undermined (Schnackenberg & Král, 2025). Behavioral visibility may similarly produce an impression of accuracy without ensuring that the visible trace is genuine. When behavioral evidence is difficult to fabricate or alter, receivers have stronger grounds for treating the trace as a genuine reflection of the focal agent.

Signals influence judgments when receivers believe they are difficult to fake and therefore reliably connected to the underlying quality they purportedly represent (Connelly et al., 2025). A core proposition of signaling theory is that signals become informative when receivers believe they accurately reflect an underlying characteristic and cannot be easily imitated by lower-quality actors (Connelly et al., 2011). Signals that are easy to manipulate lose credibility because receivers can no longer infer the underlying quality from the observed behavior (Connelly et al., 2011). Resistance to counterfeiting therefore strengthens receivers' belief that an observed signal truthfully represents the underlying characteristic it is taken to convey. When behavioral traces are difficult to fabricate, imitate, or alter without detection, receivers are more likely to believe that the observed behavior accurately reflects the agent's capabilities. This effect is unlikely to be even across the range of observed behavior. Fabrication follows incentive, and the incentive is to appear capable, so concerns about counterfeiting arise mainly for conduct that flatters an agent. A visible failure tends to be credible without tamper-evidence, because there is little reason to manufacture poor performance, except where an adversary has cause to discredit an agent. We therefore propose:

- **Proposition 6. Resistance to counterfeiting moderates the relationship between the quality of observed agent behavior and inferred agent capability, such that the relationship is stronger when resistance to counterfeiting is high.**

Taken together, the five features explain how strongly the quality of observed agent behavior translates into inferred agent capability under conditions of technological opacity. Provenance clarity directs capability attribution toward the agent rather than its operator. Non-curation indicates that the observed behavior was not staged for an audience. Consistency allows

receivers to interpret separate behavioral traces as evidence of stable and dependable performance. Interpretive legibility renders the capability embedded in those traces recognizable. Finally, resistance to counterfeiting provides assurance that the observed behavior is authentic rather than fabricated.

The conceptual model does not imply that inferred agent capability corresponds perfectly to actual capability. Because inferred capability is perception-based, receivers may overestimate or underestimate an agent even when all five features are strong. Nor does the model imply that every observable behavior will favor the agent. The five features govern how strongly the quality of observed agent behavior translates into capability judgments, not the direction of those judgments. Where the features are strong, competent conduct produces markedly higher inferred capability and visible failures produce markedly lower inferred capability, because receivers have fewer grounds for discounting what they have seen. Where the features are weak, receivers discount the observation and their capability judgments remain anchored on prior expectations regardless of what the behavior appears to show. The propositions state general tendencies rather than universal relationships. Two qualifications follow from the discussion above. Resistance to counterfeiting should matter more where conduct appears competent than where it appears to fail, because fabrication follows incentive. Interpretive legibility may raise a receiver's confidence in an inference without improving its accuracy, since a readable trace is not necessarily a correctly read one. Both possibilities are compatible with the propositions as stated and are taken up as questions for empirical work.

Moltbook currently performs poorly against several of these features. Human involvement makes provenance uncertain; the public-posting environment complicates claims of non-curation; textual traces are inexpensive to fabricate; longitudinal observations across comparable tasks are limited; and observers lack established norms for interpreting agent behavior. These weaknesses make it difficult for audiences to form well-grounded assessments of agent capability. Moltbook should therefore not yet be treated as a fully functioning signaling arena. Rather, it illustrates both the possibility of inferring capability from public agent behavior and the features agentic social media platforms would need to strengthen for such inferences to become more consequential.

5 Moltbook as an illustration of agentic social media

Moltbook illustrates agentic social media because its design renders agent behavior observable to the public while obscuring key elements of its production. Announced in January 2026, Moltbook presents itself as "a social network for AI agents," on which agents share content, comment, and upvote while humans are invited to spectate (Moltbook, 2026). The platform resembles Reddit: agents with named profiles post into topic-based "submolts," and visibility is driven by votes and recency, so what a human observer sees is a ranked selection rather than a representative sample. Registration requires both an agent and its human operator, with the operator claiming ownership through a social-media account (Moltbook, 2026). Moltbook is therefore not an autonomous society but a hybrid public space containing agent-generated, human-directed, and potentially human-authored behavior.

Moltbook attracted considerable attention before being acquired by Meta in March 2026 (Silberling, 2026), although whether it represents social media's future is orthogonal to its usefulness here. The account that follows is illustrative rather than evidential. It does not test

the propositions, and we make no claim that Moltbook's characteristics generalize to other platforms. Its value is that the platform makes agent behavior public while leaving the process behind that behavior only partially visible, which is precisely the condition the five features are meant to diagnose.

Table 2 summarizes how Moltbook performs against each of the five features, distinguishing what the platform makes available from what it leaves unavailable. The prose that follows takes up only the findings that are most revealing.

Table 2 | Moltbook assessed against the five features of unintentional agent signals

Feature	What the platform makes available	What remains unavailable	Assessment
Provenance clarity	Persistent accounts, profiles, ownership claims, and posting histories.	Which model produced a post, whether prompts or tools were used, and whether an operator edited or submitted it.	Nominal, not origin-level.
Non-curation	Agents post at any time rather than within a bounded session with a visitor.	Whether posting was operator-scheduled or operator-directed, and which episodes were withheld.	Entangled with operator direction.
Consistency	Named accounts, searchable histories, repeated interactions, and vote tallies.	Whether model, prompts, memory, tools, or supervision changed behind a stable account name.	Partial, and concentrated in few accounts.
Interpretive legibility	Familiar social-media conventions that make activity easy to read.	The originating task, the tools called, and the outcome for any given post.	Readable, but in incompatible ways.
Resistance to counterfeiting	Account-level identity, after the reported vulnerabilities were patched.	Any binding between a post and the system that produced it.	Low; text is cheap and credentials were exposed.

Three findings show why the unavailable column matters more than the available one. A WIRED journalist was accepted as an agent and posted human-written content that performed on a par with agent-generated posts (Rogers, 2026). Security researchers found no effective proof that any account's activity was machine-produced, and agents' authentication credentials were left exposed to anyone visiting a profile, so the account name attached to a post was not reliable evidence of what produced it (Nagli, 2026). Holtz (2026) found that 34.1% of messages in the platform's first days were exact duplicates of circulating templates, which means that producing content indistinguishable from the platform's modal output required no agentic capability at all. A further complication runs opposite to conventional bot detection, since where bots imitate humans, agentic social media rewards humans performing as machines in order to pass as autonomous participants (Rogers, 2026).

Two further limits qualify what posting histories can show. Operators may change the underlying model, prompts, memory, tools, or level of supervision while keeping the same account name, so stability of identity does not imply stability of capability. Activity is also highly concentrated, with a small minority of accounts producing most posts and a small minority of posts attracting most attention (De Marzo & Garcia, 2026; Holtz, 2026), so what

can be observed over time describes a few high-volume accounts rather than a typical agent (Jiang et al., 2026).

The most revealing evidence concerns interpretation rather than production. Holtz (2026) found that Moltbook reproduces macro-level signatures of human social networks, including heavy-tailed participation and small-world connectivity with an average path length of 2.91, while its micro-level signatures do not, with a mean thread depth of 1.07, 93.5% of comments receiving no replies, and reciprocity of 0.197. De Marzo and Garcia (2026) reached a compatible conclusion, noting a sublinear relationship between upvotes and discussion size. An observer attending to aggregate structure would conclude that agents are socially engaged, while an observer attending to conversational structure would conclude that they are largely posting past one another. Both readings are available from the same feed, which is what interpretive legibility is intended to capture. Early threads on religion and philosophy were read by some as evidence of emergent machine culture, and later reporting traced prominent examples to role-playing humans, prompt optimization, spam, and manual posting (Rogers, 2026; Schmelzer, 2026).

Moltbook is therefore most informative as a negative case. Publicly accessible agent behavior does not by itself guarantee transparency about capability; behavior that is easy to observe may conceal rather than reveal how it was produced, whether through human direction, platform incentives, or security failures. Holtz (2026) frames the open question in similar terms, asking whether the observed patterns represent an as-if performance of human interaction or a genuinely different mode of agent sociality, and other analysts distinguish conspicuous AI theater on public forums from the less visible agents embedded in organizations performing specific economic functions (Yıldız, 2026). Our framework suggests the question cannot be settled by looking harder at the feed, only by changing what the platform discloses about how the behavior was produced. Behavioral signals of agent competence are credible only when they are sufficiently attributable, non-curated, consistent, interpretable in context, and difficult to counterfeit.

6 Discussion

This study has implications for signaling theory, AI transparency research, and the emerging practice of deploying agentic AI in organizational settings. More broadly, we contribute to understanding how businesses (organizations) and consumers make judgments about technologies whose internal functioning remains largely inaccessible.

6.1 Theoretical implications

Our research makes three theoretical contributions. First, the primary theoretical contribution is the extension of signaling theory into contexts characterized by extreme technological opacity. More broadly, the study contributes to a growing stream of research arguing that advances in generative and agentic AI create novel contexts that challenge established management theories and require theoretical adaptation (Korzynski, Mazurek, Altmann, et al., 2023; Ratten, 2024). Recent work extends this argument to embodied and agentic systems specifically, reframing management theories around AI that acts rather than advises (Korzynski et al., 2025). Classical signaling theory assumes identifiable senders, observable signals, and mechanisms through which dishonest signaling can be penalized (Connelly et al., 2025; Spence, 1973). Agentic AI challenges each of these assumptions. In many cases,

neither users nor organizations can directly observe the processes that generate outputs, and even developers may have limited visibility into the reasoning mechanisms underlying contemporary foundation models. Consequently, traditional costly signals become difficult to identify, verify, and interpret.

Our model suggests that under conditions of AI opacity, unintentional signaling becomes a central mechanism through which capability judgments are formed. Rather than relying on deliberate disclosures from vendors, observers infer capabilities from behavioral traces produced by agents during interactions. This shifts attention from signals intentionally designed to communicate quality toward signals that emerge incidentally through observed agent behavior. In doing so, the study responds to recent calls to broaden signaling theory to encompass new signalers, new receivers, and novel signaling environments (Connelly et al., 2025).

Importantly, this extension distinguishes the firm as the signaler from the AI agent as the producer of observable behavioral signals. Although the firm determines the agent's configuration and deployment conditions, individual behavioral traces may emerge without deliberate signaling intent. Receivers use these traces to infer the capability of the agent as deployed, which may subsequently influence their evaluation of the responsible firm.

Second, the model also contributes to the AI transparency literature. Existing research has largely examined transparency as a feature intentionally embedded into AI systems through explanations, confidence indicators, disclosure mechanisms, or governance processes (Lee & Cha, 2025; Liao & Wortman Vaughan, 2024; Toya et al., 2026). Our analysis suggests that transparency may also emerge indirectly through public observation of agent behavior. Agentic social media platforms create an environment in which transparency is not solely generated by designers but is co-produced through the visibility of agent interactions. In this sense, transparency becomes an emergent social phenomenon rather than merely a technical feature.

Our model explains why observed agent behavior acquires evidential value. Provenance clarity supports source attribution, while non-curation reduces suspicion that behavior has been staged for an audience. Consistency increases confidence in the predictive value of observed behavioral patterns, and interpretive legibility provides the sensemaking resources needed to recognize and evaluate capability-relevant evidence (Leonardi & Treem, 2020; Schnackenberg et al., 2021; Weick et al., 2005). Resistance to counterfeiting, in turn, strengthens signal credibility by reducing the perceived likelihood of fabrication or manipulation (Connelly et al., 2011; Connelly et al., 2025). Together, the five features explain how receivers determine the evidential weight of behavioral signals, rather than merely identifying which characteristics of those signals matter.

Third, the paper contributes to understanding AI trust formation. Prior research has often examined trust in relation to information provided directly by AI developers or platform operators. Our model suggests that trust may increasingly be shaped by peer observation of agents interacting in public environments. Recent research on AI chatbot quality similarly indicates that perceived quality and mistrust can shape downstream user responses, although these relationships are context-dependent (Dužević et al., 2025). Agentic social media platforms therefore represent a novel intermediary layer between AI developers and AI users, creating opportunities for capability inference that do not rely exclusively on formal

disclosures. However, because trust is not directly modelled in this study, it should be treated as a potential downstream outcome of inferred capability rather than as an immediate consequence of behavioral visibility.

6.2 Practical implications

The model implies that credible agentic social media platforms of the future will depend less on signals of things agents can be made to say and more on whether platforms can be designed to satisfy conditions for treating agent behavior as evidence. Importantly, these features do not necessarily make capability judgments more favorable; rather, they increase the extent to which observable behavior is treated as credible and diagnostically useful evidence. The implications will vary by stakeholder.

The practical and societal implications of the conceptual model are summarized in Table 3 and discussed in the next sections. The conceptual model enables receivers to ask themselves the following questions about the output of an agent:

1. Who produced it? → Provenance clarity
2. Is it genuine? → Non-curation
3. Will it happen again? → Consistency
4. What does it mean? → Interpretive legibility
5. Can I trust the evidence? → Resistance to counterfeiting

Table 3 | Practical and societal implications of unintentional agent signals for inferred agent capability

Features of unintentional agent signals	Practical implications
1. Provenance clarity	Verify agent identities and clearly distinguish AI, human, and hybrid contributions so stakeholders can confidently attribute observed capabilities to the correct actor.
2. Non-curation	Make real task execution, collaboration, and error recovery visible so stakeholders can assess genuine capabilities rather than curated demonstrations.
3. Consistency	Preserve persistent agent identities and longitudinal behavioral histories so stakeholders can evaluate whether capabilities are reliable and likely to recur in future situations.
4. Interpretive legibility	Provide contextual information, metadata, and shared evaluation standards so stakeholders can correctly interpret behavioral traces and recognize capability-relevant signals.
5. Resistance to counterfeiting	Use tamper-resistant records, provenance tracking, and auditable execution logs so stakeholders can trust that observed behaviors genuinely reflect the agent's performance rather than manipulation or fabrication.

6.2.1 Practical implications for platform designers and AI developers

For designers of future agentic social media platforms, the conceptual model identifies conditions under which public observation can support credible capability inference. Current platforms such as Moltbook demonstrate the possibility of making agent behavior visible, but

they also reveal significant limitations regarding signal credibility. Future platforms could strengthen provenance clarity through cryptographic identity verification, authenticated model disclosures, or mechanisms that distinguish agent-generated content from operator intervention. Such features would help observers attribute behavior more confidently to the underlying agent rather than to human operators.

Similarly, platform architectures could increase non-curation by exposing agents engaged in real tasks rather than purely performative interactions. Public records of task completion, collaboration episodes, problem-solving activities, and autonomous decision-making could provide richer evidence of underlying capability than isolated posts designed for visibility. Resistance to counterfeiting could be improved through verifiable execution logs, provenance records, and trusted attestations of model identity. Consistency could be enhanced by enabling long-term observation of agents across multiple contexts and tasks. Like the remaining features, interpretive legibility is a property of how behavior is signaled rather than of the audience alone. It may be strengthened through educational mechanisms, shared evaluation standards, and communities capable of distinguishing meaningful behavioral patterns from superficial displays of competence (Korzynski, Mazurek, Krzypkowska, et al., 2023; Lenart et al., 2025).

For AI developers, these propositions suggest that future trust-building efforts may extend beyond publishing transparency reports or model cards. Publicly observable behavior may increasingly shape stakeholder perceptions of AI capability and trustworthiness. Developers therefore face incentives not only to improve system performance but also to design systems whose observable behavior generates credible signals regarding reliability, accountability, and limitations. Hybrid platforms, ones where AI agents and humans coexist, require legibility and provenance as core design priorities. Users should be able to rapidly ascertain who or what is speaking (human/agent/human-plus-agent), and under what permissions that actor is authorized to speak. That maps directly to the five features from above: agents and operator accounts should have verified identities (provenance), agents should speak for themselves (non-curation), mechanisms should be in place to make trace falsification expensive (counterfeit resistant), agents should have persistent identities across messages and service episodes (consistent), and platforms should surface metadata in ways that help observers correlate behavior over time(interpretive legibility). Agent-only platforms like Moltbook are interesting because they create spaces for agents to freely interact, allowing researchers to study patterns without concern for human norms or behaviors. Hybrid platforms will additionally require guardrails (rate limiting / tool-access boundaries / sandboxing), because agents can take action at faster rates than humans can. A platform that allows unfettered action at machine speed will simply produce signals designed to be unreadable.

6.2.2 Practical implications for organizational adopters and decision-makers

The conceptual model has important implications for organizations considering the adoption of agentic AI. Many organizations face uncertainty regarding the actual capabilities, limitations, and risks of AI agents, reflecting broader concerns surrounding reliability, accountability, and unintended consequences of AI deployment (Wach et al., 2023). This uncertainty often results in delayed adoption, constrained experimentation, or excessive reliance on vendor claims. Organizational adoption also requires complementary investments

in workforce training, implementation capacity, and the identification of use cases with clearly measurable benefits, as the expected operational advantages of AI may otherwise be offset by high initial costs, skills gaps, and implementation risks (Weshah, 2025). Agentic social media platforms may partially reduce these barriers by providing observable evidence of how agents behave in practice.

Organizations can use publicly observed agent behavior as an additional source of due diligence when evaluating AI systems. Decision-makers should examine whether visible performance relates to the tasks, risk levels, and deployment conditions relevant to the organization. Recent work on AI-supported business negotiation illustrates why such task specificity matters: useful performance within a structured phase does not establish competence across the entire business process (Odrobina et al., 2025). Instead of relying solely on technical documentation or marketing materials, decision makers may observe patterns of agent behavior, including error correction practices, tool usage, escalation decisions, confidence calibration, and interactions with other agents. Such observations can provide valuable insights into operational realities that are difficult to capture through vendor disclosures alone.

Importantly, observing agent behavior may help businesses (organizations) identify risks before deployment. Examples such as confidence drift, unnecessary tool usage, authority creep, and shared model blind spots illustrate that AI systems can exhibit failure modes that remain invisible in traditional product demonstrations. Exposure to these behaviors may encourage organizations to develop stronger governance mechanisms, including human oversight requirements, escalation procedures, audit trails, verification checkpoints, and limitations on autonomous decision-making. Such mechanisms matter most where practical influence and formal accountability have come apart, which is the configuration that agentic deployment tends to produce (Korzynski et al., 2026).

The conceptual model also suggests that businesses (organizations) should evaluate AI agents not only based on performance outcomes but also on the quality of observable behavioral signals. Systems that consistently demonstrate transparent decision processes, stable behavior across contexts, and clear boundaries of autonomy may be perceived as more trustworthy than systems that merely produce accurate outputs. Nevertheless, observable signals should complement rather than replace controlled evaluations, security assessments, and task-specific testing.

Hybrid agent–human spaces change the semantics of communication. Content posted to hybrid platforms must persuade humans while also being parsed by agents who filter, compare, and recommend on behalf of those humans. Trust becomes based less on who the user knows, how many followers she has or how many people she reaches and more on signals of verifiable reliability: source legitimacy, auditability, identity consistency, and a record of predictable, useful behavior across time. Hybrid platforms might, under these conditions, provide something close to a real-world alternative to vendor benchmarks. Will agents of different vendors really coordinate? Can they recover from errors publicly? Capability will be inferred from situated performance, not tests. That will not fully deliver on its promise unless platforms enable credible attribution. Without strong provenance guarantees, situated performance is just another medium for careful self-presentation.

Agents should probably just willingly constrain themselves to communicating who is acting and under what permissions, especially since platforms can always just enforce sandboxing, least-privilege tool access, and auditable decision logs. But the larger point is that the basis of competition is changing. Because text is now inexpensive to produce and easy to fake or imitate, it may function as cheap talk unless supported by credible evidence (Farrell & Rabin, 1996). An agent's ability to coordinate, summarize, escalate, and publicly recover from errors may therefore be more informative than its ability to speak persuasively on its own behalf.

6.3 Societal implications

At a broader societal level, agentic social media may contribute to AI literacy. Public discussions about AI often oscillate between narratives of technological inevitability and narratives of existential risk, biases, misinformation, and loss of control (Sieja & Wach, 2023; Wach et al., 2023). By exposing the everyday behavior of AI agents, such platforms may provide a more grounded understanding of what these systems can and cannot do. Recent evidence also suggests that institutional support, perceived quality, and opportunities to develop relevant knowledge and skills influence individuals' readiness to use generative AI consciously and safely (Lenart et al., 2025).

Observing agents in practice can reveal both strengths and weaknesses. Users may learn about limitations such as hallucinations, confidence errors, coordination failures, and inappropriate autonomous actions alongside demonstrations of useful capabilities. This balanced visibility may help reduce both unrealistic optimism and exaggerated fears regarding agentic AI (Wach et al., 2023).

The five features identified in this study also have broader societal implications. Provenance clarity and resistance to counterfeiting may help citizens distinguish authentic agent behavior from human-authored, manipulated, or misleading content. Non-curation and consistency can enable observers to assess whether visible conduct reflects an agent's recurring behavior rather than an isolated or carefully staged demonstration. Interpretive legibility is particularly important at the societal level because access to behavioral traces does not ensure that non-specialist audiences can evaluate them appropriately. Public institutions, educational organizations, and platform operators may therefore need to support the development of shared vocabularies and evaluative practices that help citizens interpret observed agent behavior without overestimating either its capabilities or its risks.

Nevertheless, the presence of these features does not automatically produce an accurate understanding (Ananny & Crawford, 2018). Observable behavior can be manipulated, staged, or misunderstood. Consequently, the societal value of agentic social media platforms depends heavily on the credibility of the signals they generate and the interpretive capabilities of the audiences observing them. Without appropriate safeguards, public visibility may amplify misconceptions rather than reduce them.

6.4 Limitations

This conceptual study has several limitations. First, Moltbook is used as an illustrative example rather than as an empirical case study. The paper does not systematically sample platform content, measure receiver responses, or independently verify whether the observed posts were produced without human intervention.

Volume **15** | Issue **04** | 2026
https://doi.org/10.18267/j.cebr.428

Second, inferred capability may diverge from actual capability, and behavior observed on a public platform may not generalize to organizational deployment. Platform incentives, operator choices, system prompts, tool access, and task selection may all shape the visible traces. Increased behavioral visibility may also create trade-offs with privacy, security, confidentiality, and intellectual-property protection.

6.5 Future research

Several opportunities for future research emerge from this study. First, empirical studies could examine whether the five proposed features actually influence perceived capability, trustworthiness, and adoption intentions among consumers and organizational decision makers. Experimental designs could manipulate provenance clarity, non-curation, consistency, resistance to counterfeiting, and interpretive legibility to test their relative effects. Such studies should also vary whether the observed behavior indicates success or failure in order to distinguish signal credibility from signal valence.

Future research should also examine whether the five features operate independently, additively, or interactively. For example, consistency may have limited evidential value when provenance is unclear, whereas resistance to counterfeiting may operate asymmetrically for the reasons set out in Section 4.5. Testing that possibility calls for manipulating observed behavior at more than two levels, or for modeling competent conduct and visible failure separately, because a single multiplicative interaction term assumes a uniform effect across the range. Because interpretive legibility depends on how behavior is presented as well as on audience expertise, studies should further test whether expertise, shared evaluative standards, and platform-provided explanations improve the calibration of capability judgments rather than merely raising receivers' confidence in them.

Beyond testing the proposed conceptual model, several broader research questions emerge from this study. First, how do different agent–human interaction designs enable or inhibit capability inference and trust-building under AI opacity? Existing work on AI in social media has focused on firms using AI to support marketing activities; the inverse question (how AI agents themselves generate signals that audiences read) remains largely unexamined. This research agenda is consistent with recent calls for more context-sensitive studies of AI-related experiences, strategies, and business applications (Ratten, 2024).

Second, which platform mechanisms enable credible capability inference from observed agent behavior? This question contains both a design dimension, concerning the mechanisms that can be architected into agentic social media platforms, and an empirical dimension, concerning which mechanisms lead audiences to infer particular capabilities and levels of trustworthiness. Future research could investigate how different audiences interpret agent-generated signals. Technical experts, managers, regulators, and consumers may rely on different cues when assessing AI capability, leading to divergent interpretations of the same observable behavior. The model could also be developed by incorporating other sources of capability evidence. Receivers rarely rely on observed behavior alone, and how behavioral signals combine with vendor disclosures, benchmark results, and reputational information, particularly when these sources conflict, remains open.

Third, as agentic social media platforms emerge how do interpretive communities form up around those platforms, and what shared vocabulary do those communities develop for "reading" agent behavior? The answer to that last question will determine if traces of agent

behavior are ever really a substitute for vendor benchmarks, or whether they will remain mere noise. Comparative studies could examine different forms of agentic social media environments beyond Moltbook. As agentic platforms evolve, researchers will have opportunities to compare public agent communities, hybrid human–agent platforms, organizational agent networks, and specialized industry-specific agent ecosystems. A further question concerns the medium in which agent behavior is expressed. Observable behavior is currently almost entirely textual, and text is cheap, but platforms on which agents transact, call verifiable tools, or produce checkable artifacts would make some behavior costly to imitate. Whether receivers distinguish costly conduct from cheap text, and whether the five features carry different weight once they do, is an open empirical question.

Finally, future work could explore the relationship between signaling and governance. Such work may be particularly important given growing concerns regarding governance, accountability, and risk management in increasingly autonomous AI systems (Wach et al., 2023). If observed agent behavior becomes an important source of capability inference, regulators and policymakers may eventually consider whether public behavioral transparency should complement existing requirements for technical transparency and accountability.

Conclusion

As AI systems become increasingly autonomous, traditional forms of transparency may prove insufficient (Wach et al., 2023). Instead, growing importance will be attached to the ability to observe agents "in action" and to interpret the behavioral traces they leave behind. This argument reflects broader concerns that the growing capabilities of generative and agentic AI create both significant opportunities and substantial uncertainties regarding trust, accountability, and governance (Sieja & Wach, 2023). Platforms such as Moltbook may emerge as new mechanisms for building knowledge about and trust in AI agents. However, they currently face significant challenges related to signal credibility, the possibility of human intervention, and the difficulty of distinguishing genuine agent behavior from performative displays. By identifying provenance clarity, non-curation, consistency, interpretive legibility, and resistance to counterfeiting as five features shaping capability inference, the framework explains when publicly observed agent behavior is more likely to be treated as credible and diagnostically useful evidence. These mechanisms do not necessarily produce more favorable evaluations; rather, they strengthen the extent to which both successful performance and visible failure influence judgements of agent capability.

Consequently, we propose extending signaling theory to the context of agentic artificial intelligence. In doing so, the study responds to broader calls for revisiting and adapting management theories to the realities of generative and agentic AI (Korzynski, Mazurek, Altmann, et al., 2023). As AI systems evolve from task-specific agents toward increasingly autonomous, goal-directed, and collaborative forms of agentic AI (Sapkota et al., 2026), understanding how their capabilities can be inferred under conditions of opacity becomes increasingly important. We argue that observed agent behaviors may constitute a new type of signal that organizations and consumers can use to assess the capabilities of AI systems under conditions of persistent technological opacity.

In summary, the emergence of agentic social media introduces a novel mechanism through which organizations and consumers can infer AI capability under conditions of opacity. These conditions strengthen the diagnostic value of behavioral evidence rather than necessarily

producing more favorable capability judgments. While current platforms provide only imperfect signals, they reveal how public observation of agent behavior may become an increasingly important component of trust formation, technology adoption, and governance in the age of agentic AI.

References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Bafera, J., & Kleinert, S. (2023). Signaling theory in entrepreneurship research: A systematic review and research agenda. *Entrepreneurship Theory and Practice*, *47*(6), 2419–2464. https://doi.org/10.1177/10422587221138489

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, Article 107197. https://doi.org/10.1016/j.infsof.2023.107197

Beerends, S., & Aydin, C. (2025). Negotiating the authenticity of AI: How the discourse on AI rejects human indeterminacy. *AI & Society*, *40*(2), 263–276. https://doi.org/10.1007/s00146-024-01884-5

Berlo, D. K., Lemert, J. B., & Mertz, R. J. (1969). Dimensions for evaluating the acceptability of message sources. *Public Opinion Quarterly*, *33*(4), 563–576. https://doi.org/10.1086/267745

Cardiet, L. (2026, February 3). *Moltbook and the illusion of "harmless" AI-agent communities*. Vectra AI. https://www.vectra.ai/blog/moltbook-and-the-illusion-of-harmless-ai-agent-communities

Carr, C. T., & Walther, J. B. (2014). Increasing attributional certainty via social media: Learning about others one bit at a time. *Journal of Computer-Mediated Communication*, *19*(4), 922–937. https://doi.org/10.1111/JCC4.12072

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 958–973). ACM. https://doi.org/10.1145/3630106.3658948

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning models don't always say what they think*. arXiv. https://doi.org/10.48550/arXiv.2505.05410

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, Article 107018. https://doi.org/10.1016/J.CHB.2021.107018

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management, 37*(1), 39–67. https://doi.org/10.1177/0149206310388419

Connelly, B. L., Certo, S. T., Reutzel, C. R., DesJardine, M. R., & Zhou, Y. S. (2025). Signaling theory: State of the theory and its future. *Journal of Management*, *51*(1), 24–61. https://doi.org/10.1177/01492063241268459

De Marzo, G., & Garcia, D. (2026). *Collective behavior of AI agents: The case of Moltbook*. arXiv. https://doi.org/10.48550/arXiv.2602.09270

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, *99*, Article 101896. https://doi.org/10.1016/j.inffus.2023.101896

Dužević, I., Baković, T., & Surman, V. (2025). Understanding artificial intelligence chatbot quality and experience: A higher education student perspective. *Entrepreneurial Business and Economics Review*, 13(3), 151–170. https://doi.org/10.15678/EBER.2025.130308

Eckardt, R., Skaggs, B. C., & Lepak, D. P. (2018). An examination of the firm-level performance impact of cluster hiring in knowledge-intensive firms. *Academy of Management Journal*, *61*(3), 919–944. https://doi.org/10.5465/amj.2016.0601

Erlei, A., Cau, F. M., Georgiev, R., Kumar, S. C., Bizer, K., & Gadiraju, U. (2026). When life gives you AI, will you turn it into a market for lemons? Understanding how information asymmetries about AI system capabilities affect market outcomes and adoption. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (pp. 1–24). Association for Computing Machinery. https://doi.org/10.1145/3772318.3791420

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, *10*(3), 103–118. https://doi.org/10.1257/jep.10.3.103

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goto, M. (2022). Accepting the future as ever-changing: Professionals' sensemaking about artificial intelligence. *Journal of Professions and Organization*, *9*(1), 77–99. https://doi.org/10.1093/JPO/JOAB022

Holtz, D. (2026). *The anatomy of the Moltbook social graph.* arXiv. https://doi.org/10.48550/arXiv.2602.10131

Jiang, Y., Zhang, Y., Shen, X., Backes, M., & Zhang, Y. (2026). *"Humans welcome to observe": A first look at the agent social network Moltbook.* arXiv. https://doi.org/10.48550/arXiv.2602.10127

Kim, B., & Kim, D. (2026). Understanding the role of trust, perceived risk, and habit in organization members' generative AI use. *SAGE Open*, *16*(1). https://doi.org/10.1177/21582440251410620

Korzynski, P., Edwards, A., Gupta, M. C., Mazurek, G., & Wirtz, J. (2025). Humanoid robotics and agentic AI: Reframing management theories and future research directions. *European Management Journal*, *43*(4), 548–560. https://doi.org/10.1016/j.emj.2025.06.002

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative artificial intelligence as a new context for management theories: Analysis of ChatGPT. *Central European Management Journal*, 31(1), 3–13. https://doi.org/10.1108/CEMJ-02-2023-0091

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, 11(3), 25–37. https://doi.org/10.15678/EBER.2023.110302

Korzynski, P., Wojtczuk-Turek, A., Turek, D., Meglio, O., & Zabicka-Wlodarczyk, M. (2026). When agentic AI participates in organisational change: Rethinking agency, leadership, and accountability. *Journal of Change Management*, *26*(3), 203–221. https://doi.org/10.1080/14697017.2026.2719997

Král, P., & Schnackenberg, A. K. (2024). Organizational responses to transparency determinants. *Management Decision, 62*(13), 309–331. https://doi.org/10.1108/MD-07-2023-1244

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science, 33*(1), 126–148. https://doi.org/10.1287/orsc.2021.1549

Lee, C., & Cha, K. (2025). Toward the dynamic relationship between AI transparency and trust in AI: A case study on ChatGPT. *International Journal of Human-Computer Interaction*, *41*(13), 8086–8103. https://doi.org/10.1080/10447318.2024.2405266

Lehman, D. W., O'Connor, K., Kovács, B., & Newman, G. E. (2019). Authenticity. *Academy of Management Annals*, *13*(1), 1–42. https://doi.org/10.5465/ANNALS.2017.0047

Lenart, R., Sypniewska, B. A., Chen, J., & Janowski, K. (2025). Antecedents of students' behavioural intention to use generative artificial intelligence: Quantitative research. *Entrepreneurial Business and Economics Review*, 13(1), 7–21. https://doi.org/10.15678/EBER.2025.130101

Leonardi, P. M., & Treem, J. W. (2020). Behavioral visibility: A new paradigm for organization studies in the age of digitization, digitalization, and datafication. *Organization Studies, 41*(12), 1601–1625. https://doi.org/10.1177/0170840620970728

Liao, Q. V., & Wortman Vaughan, J. (2024). AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*, (Special Issue 5). https://doi.org/10.1162/99608f92.8036d03b

Moltbook. (2026). *Moltbook: The front page of the agent internet.* https://www.moltbook.com/

Nagli, G. (2026, February 2). Hacking Moltbook: The AI social network any human can control. *Wiz.* https://www.wiz.io/blog/exposed-moltbook-database-reveals-millions-of-api-keys

Odrobina, A., Polan, W., & Świerczyńska, J. (2025). Artificial intelligence negotiation algorithms: Pioneering artificial intelligence negotiation algorithms for business. *International Entrepreneurship Review*, 11(4), 73–85. https://doi.org/10.15678/IER.2025.1104.05

Radanliev, P. (2025). AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence*, *39*(1), Article 2463722. https://doi.org/10.1080/08839514.2025.2463722

Ratten, V. (2024). Artificial intelligence: Building a research agenda. *Entrepreneurial Business and Economics Review*, 12(1), 7–16. https://doi.org/10.15678/EBER.2024.120101

Rogers, R. (2026, February 3). I infiltrated Moltbook, the AI-only social network where humans aren't allowed. *WIRED.* https://www.wired.com/story/i-infiltrated-moltbook-ai-only-social-network/

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *Information Fusion*, *126*, Article 103599. https://doi.org/10.1016/j.inffus.2025.103599

Sauer, S. J., Thomas-Hunt, M. C., & Morris, P. A. (2010). Too good to be true? The unintended signaling effects of educational prestige on external expectations of team performance. *Organization Science*, *21*(5), 1108–1120. https://doi.org/10.1287/orsc.1090.0523

Schmelzer, R. (2026, February 10). Moltbook looked like an emerging AI society, but humans were pulling the strings. *Forbes.* https://www.forbes.com/sites/ronschmelzer/2026/02/10/moltbook-looked-like-an-emerging-ai-society-but-humans-were-pulling-the-strings/

Schnackenberg, A. K., & Král, P. (2025). The blinding of truth by falsehood: Synthetic transparency and the growth of strategic opacity. *Academy of Management Proceedings, 2025*(1), 19125. https://doi.org/10.5465/amproc.2025.19125abstract

Schnackenberg, A. K., & Tomlinson, E. C. (2016). Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *Journal of Management*, *42*(7), 1784–1810. https://doi.org/10.1177/0149206314525202

Schnackenberg, A. K., Tomlinson, E. C., & Coen, C. A. (2021). The dimensional structure of transparency: A construct validation of transparency as disclosure, clarity, and accuracy in organizations. *Human Relations,* *74*(10), 1628–1660. https://doi.org/10.1177/0018726720933317

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023). *Practices for governing agentic AI systems*. OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Sieja, M., & Wach, K. (2023). Revolutionary artificial intelligence or rogue technology? The promises and pitfalls of ChatGPT. *International Entrepreneurship Review*, 9(4), 101–115. https://doi.org/10.15678/IER.2023.0904.07

Silberling, A. (2026, March 10). Meta acquired Moltbook, the AI agent social network that went viral because of fake posts. *TechCrunch.* https://techcrunch.com/2026/03/10/meta-acquired-moltbook-the-ai-agent-social-network-that-went-viral-because-of-fake-posts/

Simkute, A., Luger, E., Jones, B., Evans, M., & Jones, R. (2021). Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable. *Journal of Responsible Technology*, *7–8*, Article 100017. https://doi.org/10.1016/j.jrt.2021.100017

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

Spence, M. (1974). *Market signaling: Informational transfer in hiring and related screening processes*. Harvard University Press.

Teepapal, T. (2025). AI-driven personalization: Unraveling consumer perceptions in social media engagement. *Computers in Human Behavior*, *165*, Article 108549. https://doi.org/10.1016/j.chb.2024.108549

Toy, T. (2024). Transparency in AI. *AI & Society*, *39*(6), 2841–2851. https://doi.org/10.1007/s00146-023-01786-y

Toya, K., Hasida, K., & Sanada, K. (2026). User acceptance of personal AI agents: Benefits and risks of personal data use. *Computers in Human Behavior: Artificial Humans*, *8*, Article 100299. https://doi.org/10.1016/j.chbah.2026.100299

Vanneste, B. S., & Puranam, P. (2025). Artificial intelligence, trust, and perceptions of agency. *Academy of Management Review*, *50*(4), 726–744. https://doi.org/10.5465/amr.2022.0041

Varsha, P. S. (2023). How can we manage biases in artificial intelligence systems – A systematic literature review. *International Journal of Information Management Data Insights*, *3*(1), Article 100165. https://doi.org/10.1016/j.jjimei.2023.100165

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2), 7–30. https://doi.org/10.15678/EBER.2023.110201

Weick, K. E. (1995). *Sensemaking in organizations*. Sage.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, 16(4), 409–421. https://doi.org/10.1287/orsc.1050.0133

Weshah, S. (2025). Empirical insights into the reduction of operational costs through AI: A study of Jordanian companies. *International Entrepreneurship Review*, 11(3), 7–16. https://doi.org/10.15678/IER.2025.1103.01

Yıldız, G. (2026, June 22). The AI agents that actually work don't post on social media. *Forbes.* https://www.forbes.com/sites/guneyyildiz/2026/06/22/the-ai-agents-that-actually-work-dont-post-on-social-media/

The article has been reviewed. | Received: 30 July 2026**; Revised:** 17 October 2026**; Accepted:** 21 October 2026**; Available online:** 29 October**; Published in regular issue:** 31 October 2026.