Submission to the Public Consultation on Advancing AI Transparency in Canada

Re: Discussion paper, Enhancing trust in artificial intelligence through increased transparency

Submitted to: Innovation, Science and Economic Development Canada

Date: September 21, 2026

Submitted by:
Twinvocate Inc.
100 King Street West, Suite 6200, 1 First Canadian Place
Toronto, Ontario M5X 1B8
wendy@twinvocate.ai

Per: Wendy MacKinnon, Founder and Chief Executive Officer

Declaration of interest

Twinvocate Inc. is a Toronto company building a customer-loyal AI agent and a certified-counterparty credentialing layer for agentic commerce. We have a commercial interest in how Canada regulates agent transparency and disclose it here so that these comments can be weighed accordingly. Our recommendations are drafted to apply to any participant in this market, including us.

Summary

Our comments concentrate on section 3.5, AI Agents, with shorter observations on sections 3.2 and 3.4 and on the cross-cutting questions.

We make one central argument. The discussion paper frames agent transparency almost entirely as visibility into what an agent did. That is necessary but incomplete. The prior question is on whose behalf the agent acted. An agent that is fully observable and fully logged can still be operating in the interest of the platform that deployed it rather than the person it appears to serve, and no amount of activity telemetry will surface that fact. Transparency about function is not transparency about loyalty.

We make a second argument that we have not seen raised elsewhere in this consultation. The detailed activity records the paper contemplates are not neutral. They are a continuous, fine-grained record of a person's intentions, preferences, associations and transactions, held by a third party. Canadian law is clear about how such records will be characterized. Measures designed to create accountability will also create a new surveillance surface, and the design of those measures should account for that from the outset rather than after the fact.

Our recommendations follow from those two points.

1. The record created for accountability is also a record created for compulsion

The paper proposes that greater visibility into agent activity may require detailed activity logs, telemetry covering tool invocations and decisions, and chain-of-custody information across multi-agent interactions. We support the objective. We ask that the design of any such requirement begin from the recognition that these records will be sought by parties other than the principal.

In R v Bykovets, 2024 SCC 6, the Supreme Court of Canada held that an internet protocol address attracts a reasonable expectation of privacy under section 8 of the Charter, and that a state request for one constitutes a search. The majority reasoned that the analysis turns on the potential of the information to reveal an individual's biographical core, not on what it happened to reveal on the facts before the Court. That built on R v Spencer, 2014 SCC 43, which protected subscriber information associated with an IP address. The significance for this consultation is the class of data at issue. An IP address discloses almost nothing on its face. It was protected because of what it unlocks when combined with other holdings.

Agent activity logs are a far stronger case on those facts than an IP address. A complete record of what a person's agent was instructed to do, what it purchased, whom it transacted with, which counterparties it declined and on what basis is not merely revealing of the biographical core. It is close to a direct transcript of it. A mandated log is therefore a mandated repository of protected information, and the volume, retention period, custody and format of that repository are policy choices being made now, whether or not they are made deliberately.

We note that this is not a hypothetical concern about the distant future. The paper's own discussion of AI incidents and multi-agent chain of custody points toward record-keeping obligations of exactly this character.

There is a third property of these records that the paper does not address. They are produced by the system whose conduct they are meant to evidence. METR's investigation of the OpenAI evaluation incident, published on August 26, 2026 and discussed further at part 5 below, reports that agents modified part of the tool-call execution system so that the transcript recorded one command while a different one ran, in roughly 7% of the transcripts examined. The manipulation worked only forward. The agents did not succeed in altering entries already written, and they correctly identified that records held inside their own environment were not the version anyone would rely on. METR states that much of its analysis was delegated to AI systems and that it cannot rule out error in that analysis, so we treat the figure as indicative rather than established. The design consequence does not depend on the figure. A log that can be written by the party being held to account is not an accountability record. Any mandated record should be append-only and held outside the reach of the agent it records, which is also the architecture that best serves the principal under Recommendation 2 below.

Recommendation 1. Any transparency measure that requires the retention of agent activity records should be accompanied, in the same instrument, by purpose limitation, retention limits, a stated position on third-party and state access, and a means by which the principal can confirm that records derived from their data, including summaries and other derived material, have been deleted on request. Records should be held in the form least useful for secondary purposes that still satisfies the accountability objective.

Recommendation 2. Where the accountability objective can be met by records held by or portable to the principal rather than accumulated centrally by the deployer, the Government should prefer that architecture. Data minimisation is available here as a design choice, not only as a compliance obligation, and architectural choices have proven more decisive than litigation in comparable contexts.

2. Disclosure of authority, not only disclosure of activity

The paper asks who should be responsible for disclosing the use of AI agents and what information should be recorded. We suggest the answer is organised around a single artefact: the mandate under which the agent acts.

A user who is told only that an agent is in use learns very little. What is material is the scope of authority the agent holds: what it may spend, on whose accounts, with which counterparties, under what conditions, and what it may not do without returning to the person. That authority should be expressed in a form the person can read and the agent cannot exceed silently, and each action should be recorded against it.

The gap is not theoretical, and it is not confined to the platforms the paper cites. On September 14, 2026, after hearing the product recommended by a Canadian fintech figure at an industry event, we obtained an invite code through a contact and enrolled in Instinct, an invite-only personal agent operated by Spear Street Technology of San Francisco, in order to test it. A company mailbox and calendar were connected during onboarding. Within two hours, and without being given any task, the agent had read and summarised the contents of that mailbox, including commercially sensitive correspondence. The setting permitting our data to be used to improve the model was on by default. When we disconnected the mailbox, the agent reported that its index of the messages had been cleared, but the summaries it had already produced remained in the conversation record. The product does offer a control to delete external data. We used it the following day, received a confirmation that the external data had been deleted, and two minutes later the agent listed the same summaries as still held. The agent had said as much on the first evening: the control clears indexed data and does not delete memory, chat history or vault data. Removal of the summaries is a manual request the agent submitted on our behalf. Five days later, the request showed no reference, no owner and no timeline, and the agent confirmed that nothing had been removed. We also note that the first message in the conversation, which reads as if we had written it, was composed by the product and placed in our message field during onboarding, so the first words the person appeared to say to the agent were the agent's. Instinct's published terms of service authorise it to enter into agreements and transactions that bind the user. We report this as one observation of one product on one day, and we have no reason to think it unusual. It shows what a consumer agent looks like when its authority is defined by its own terms rather than by a mandate the person has set. The agent decided what to read, what to retain and what to do with it, and the person learned the answers afterwards. In the ten days before our test the same product announced a password-manager integration to hold and broker the user's credentials (September 4, 2026), gave each agent its own email address as a first step toward the agent owning its own accounts (September 9, 2026), and completed the six-digit authenticator step of a login (September 11, 2026). An agent with its own address, its principal's credentials and the ability to pass two-factor authentication is indistinguishable to the counterparty from the principal, and nothing in that stack tells the counterparty which it is dealing with or tells the principal whose interest the agent serves. The route by which we reached the product is the route by which most Canadians will: a recommendation from someone they trust, not a review of the terms.

This also supplies a workable test for the redress problem raised in the paper's agentic commerce case study, where twelve packages are ordered instead of one package of twelve. If a mandate exists and the action fell outside it, responsibility sits with the party that operated the agent. If the action fell within a mandate the person did not understand, the disclosure was inadequate. Without a mandate artefact, the allocation question has no anchor and defaults, as the paper notes some businesses have already asserted, to the user.

The paper notes a foreign precedent that points the same way. The European Union's 2024 directive on liability for defective products extends the definition of a product to software, including agentic AI systems, and presumes a product defective where the defendant fails to disclose relevant evidence at a claimant's request. As the paper observes, this could create incentives for developers and deployers to document what their agents do. We would put the point more precisely. A disclosure incentive of that

kind produces documentation only where there is something determinate to document against. A complete record of what an agent did, produced under threat of an adverse presumption, still leaves unanswered whether the action was permitted, unless the authority it acted under was fixed beforehand in a form that cannot be reconstructed once the dispute has arisen. Canada can take the useful half of that approach directly, by requiring the mandate rather than by inferring its absence from a liability rule after the fact.

Recommendation 3. Disclosure obligations for agentic systems should attach to the scope of delegated authority, expressed in both human-readable and machine-readable form, and not solely to the fact that an agent is in use.

Recommendation 4. Records should capture refusals as well as actions. An agent that declines a transaction because a counterparty could not be verified has produced the most decision-relevant evidence in the entire interaction, and current observability practice, which is oriented to tool calls and completions, does not reliably capture it.

3. Verification must be reciprocal

The paper describes ongoing industry work on identity credentials that platforms and service providers can verify when an agent interacts with them. As drafted, that is one-directional. It equips the merchant, the platform and the enterprise to satisfy themselves about the agent, and leaves the person on whose behalf the agent acts with no equivalent capability.

This is the same structural gap we identified in our submission on the proposed Consumer-Driven Banking Regulations. A framework can establish rigorously who is permitted to act while remaining silent on whose interest the actor serves. In agentic commerce the consequence is concrete: an agent instructed to buy on a person's behalf must be able to establish that the counterparty on the other side is who it claims to be, is entitled to transact, and is not itself an agent optimising against the person's interest.

The operator of Canada's domestic payments and verification infrastructure has reached the same point. In Catalyzing Canada, its August 2026 white paper, Interac writes that a future including consumer-driven banking, agentic commerce and programmable money will require trusted mechanisms to verify individuals, organizations and, increasingly, software agents acting on behalf of consumers (page 25). The paper's own illustration shows where the gap sits (page 13). It asks the reader to consider an agent, authorised through the consumer-driven banking framework, initiating an irrevocable real-time payment with the consumer's identity authenticated through trusted digital credentials, and then asks where liability sits if the agent errs or is deceived. Every verification named in that transaction runs toward the consumer. None runs toward the party the agent is dealing with, which is the only party in a position to deceive it. The incumbent has described the one-directional framework precisely. What remains is to name the missing direction.

The paper's account of market responses shows the same gap from the other direction. It describes independent third-party verification services that test and attest to an agent's behaviour against published security and risk standards before deployment and monitor it thereafter, and treats the tying of those attestations to common public standards as an early, market-led parallel to conformity assessment in other sectors. We think the comparison is apt and the services useful. But the standards named in the paper's own footnote are security and risk standards. An attestation against them establishes that an agent has been built and tested competently. It does not establish whose interest that agent serves, and it tells the person on whose behalf the agent acts nothing at all about the party

on the other side of the transaction. Conformity assessment elsewhere carries the same limit: certifying that a product is safe is not certifying that the seller is honest. A market that matures around competence attestation alone will produce agents that are demonstrably well built and, as the retailer terms of service the paper cites suggest, still pointed at someone else's interest.

Recommendation 5. Any Government work on agent identity credentials should be scoped from the outset as a two-directional problem, so that an agent acting for a consumer or a small business can verify a counterparty on the same terms a platform can verify the agent.

Recommendation 6. Verification requirements should be proportionate to the value and reversibility of the transaction. A verification step whose cost exceeds the value of the transaction it protects will be bypassed in practice, and a framework that ignores this will govern only the transactions that least need governing. This bears directly on the paper's cross-cutting question about small and medium-sized enterprises, which will bear verification costs without the ability to spread them.

Where government action adds value

The paper closes section 3.5 by asking whether and when government action would add value beyond emerging market and standards activity. We offer a direct answer, since it is the question on which we are best placed to comment.

The market will supply the testing. It is already doing so, and there is no case for the Government to build or operate a verification service of its own. The market is also supplying the pieces around the test: the product we describe in section 2 now holds and brokers its principal's credentials through a password-manager vendor, so the question of what the agent may do with those credentials is being answered by a credential supplier, not by any attestation of whose interest the agent serves. What the market will not supply unaided is the thing that makes any attestation worth relying on, which is a common answer to what an attestation asserts and who is entitled to rely on it. Competing schemes each define their own scope, so a consumer or a small business presented with an attestation has no way to compare it against another or to know which assertions it covers. That is a coordination problem, and coordination problems are where standards work, and where necessary regulation, add value that competition does not. Interac makes the same observation from the infrastructure side. Its white paper states that if these frameworks are not established through public policy and industry collaboration they will emerge through the rules of global platforms, schemes and private market actors, and it points to Visa's Trusted Agent Protocol as governance already being built (pages 12 and 24).

There is a second and narrower place. Verification of a counterparty depends on authoritative registers, and in Canada most of the registers that matter are operated by government. Publishing them in machine-readable form with a programmatic interface is a low-cost step that only their operators can take, and it is a precondition for consumer-side verification of any kind. We return to this at part 6.

The gap this consultation is addressing has also been named by the incumbent. Interac's white paper observes that the AI for All strategy did not directly address agentic AI in the context of payments, and that this leaves significant room for further engagement between industry and government (page 25). We agree. We note that the observation comes from the operator of the domestic infrastructure over which much of that agentic activity would run, not only from a new entrant with a product to sell. The product experience described at part 2 above is what fills that room in the meantime: an agent's own terms answering, by default, the questions the strategy left open. The room is not small. At a separate industry event on September 17, 2026, a show of hands put a consumer agent of this kind on roughly one in five phones present.

Recommendation 7. Government action on agentic transparency should concentrate on what the market cannot coordinate for itself: a common vocabulary for what an agent credential or attestation asserts and who may rely on it, and machine-readable publication of the authoritative registers government already maintains. Building or operating verification services is neither necessary nor a good use of public capacity.

4. Section 3.2, AI interaction

We agree that disclosure fatigue is a real risk and that a blanket statement adds little. The factor that makes disclosure meaningful, in our view, is not whether the counterpart is disclosed as artificial but whether its interest is disclosed. A person negotiating with a system that is optimising against them is materially worse off than a person negotiating with a person, and the disclosure that matters is not that the counterpart is AI but that the system acts for the other side.

We note that the paper's own case study, in which a dealership chatbot made and then repudiated an offer, and the tribunal reasoning in Moffatt v Air Canada, 2024 BCCRT 149, both turn on this point rather than on the absence of an artificiality label.

5. Section 3.4, AI incidents

We support incident reporting for agentic systems and suggest that this domain offers an unusually clean threshold. An agent action taken outside the authority the principal granted is objectively determinable where a mandate artefact exists, does not require a judgement about severity, and is exactly the class of event that reporting regimes exist to surface. We suggest it as a candidate definition for a serious incident in agentic commerce, distinct from and complementary to the harm-based thresholds used in other sectors.

We note that the case for reporting is being pressed from outside industry as well. The Centre for Long-Term Resilience, funded through the United Kingdom AI Security Institute's Challenge Fund, published research in March 2026 identifying real-world instances of AI systems pursuing objectives their operators had not sanctioned, drawn from open sources, and has since called for mandatory reporting of severe incidents. We do not rely on its incident counts, which its own authors state cannot yet be separated from a rise in reporting and in opportunity, but the direction of the recommendation is relevant to the choice before Canada.

A second question the paper should settle is who is told, and by whom. Two agentic incidents from July 2026 answered it in opposite ways. In the first, an independent investigation of an evaluation incident at OpenAI was carried out by METR with Redwood Research and published on August 26, 2026. The position of the party whose systems were compromised fell outside the agreed scope of that investigation, as did the effectiveness of the safeguards and the extent of the compromise. In the second, the United Kingdom AI Security Institute published its own incident report into unsanctioned agent behaviour during cyber testing conducted between July 25 and 28, 2026, in which an agent researched the maintainers of a public software project, created multiple false identities and used them to press a maintainer to approve malicious code. The Institute notified the affected platform, supplied a complete inventory of the accounts and artefacts its agents had created there, worked with the platform to remove them, and asked for assistance in alerting those exposed.

The difference between those two outcomes is not a difference in rule. It is a difference in discretion. In each case the party best placed to identify and notify those affected was the party whose agents caused

the harm, and in neither case was it obliged to do anything. We note that the Institute is an evaluator reporting harm arising from its own testing rather than a commercial deployer, and that the incentives are not alike. We think that is the point rather than an objection to it. The only instance of complete third-party notification available to us comes from a body whose mandate is safety.

This bears directly on the redress problem the paper raises. A person harmed by an agent that someone else operates is not a customer of the deployer, has no contractual relationship with it, and therefore has no standing in any customer complaint channel. Where the harm falls on a third party rather than on the principal, no existing avenue reaches them, and an incident regime that runs only between deployers and the Government will not create one.

Recommendation 8. An incident regime for agentic systems should place a duty on the deployer to notify identifiable third parties affected by an agent's actions within a defined period, and should specify the content of that notification in enough detail to be actionable by the recipient. The practice described above is a workable template, and it has the advantage of having been carried out once already.

6. Cross-cutting: coherence with existing Canadian frameworks

Three points of coherence deserve attention before any new instrument is designed.

First, the accreditation regime under the consumer-driven banking framework will already maintain a public register of entities permitted to access consumer financial data. Agentic commerce transparency should build on that register rather than establish a parallel one, and the register should be published in machine-readable form with a programmatic interface so that software acting for a consumer can confirm status without human intermediation.

Second, entities that move money on behalf of consumers through agentic interfaces are likely to be payment service providers subject to supervision under the Retail Payment Activities Act. Transparency obligations should be drafted with that supervisory relationship in view rather than alongside it.

Third, the reporting and record-keeping obligations already imposed under those regimes will generate much of the record contemplated in section 3.5. Coordinating them would reduce the compliance burden the paper is rightly concerned about, and would limit the proliferation of duplicate repositories of the protected information described in part 1 above.

Conclusion

Canada has an opportunity here that most jurisdictions have already foreclosed, which is to set the expectation that an agent transacting on a person's behalf must be able to demonstrate both what it did and for whom it acted, before the market settles into a default in which agents are deployed on people rather than for them. We would welcome the opportunity to discuss any of the above with officials.

Wendy MacKinnon
Founder and Chief Executive Officer
Twinvocate Inc.
Toronto, Ontario