

A relying party's checklist for AI agents

Version 0.1 · Published · 12 October 2026

ABSTRACT

This paper is for the people who run the front door at a bank, a hospital or a tax office. It sets out six checks to apply, in order, to every request that comes from an AI agent. Is the identity genuine, is the person or organisation behind it answerable, does its scope cover this exact action, does the delegation chain narrow at each link, what is the data-protection position, and is evidence written before anyone acts. It says what to do when a check can't be completed, explains why a model's confidence score doesn't belong in the decision, and suggests adopting the checks in three phases. It closes with notes for four kinds of front door and a list of what we haven't worked out.

READ THIS FIRST

This is a position, not a specification, and not legal advice. The tax office below is a case we made up, and the checks are our proposal. No bank, hospital, regulator or standards body has adopted them.

Picture a taxpayer services desk in a district office on a Tuesday afternoon. There are three clerks and a queue. A message arrives through the office's online form. It asks to change the contact email on a small business's registration, and it says it was sent by an AI agent working for the business's accountant.

The clerk who picks it up has about two minutes. The business number is right. The wording is sensible. Changing a contact email sounds harmless, until you notice that every notice the office sends afterwards will go to the new address.

What should the clerk check before doing anything?

Most desks today would check that the business exists. That's needed, and it isn't enough, because the thing talking to the office isn't the business. It's software acting for the accountant, who is acting for the business. The office is the relying party: the one that's about to rely on somebody else's claim and act on it. This paper is for the people in that chair.

We're setting up a non-profit called Saakshya to work on agent identity. It isn't incorporated, it has no members yet, and no regulator or company has endorsed anything below. What follows is our view of what a careful front desk should do, written so that a desk can use it without needing anything from us. If you want the wider argument for why this matters, it's in [our first position paper](https://saakshya.org/papers/agent-identity-for-india) (saakshya.org/papers/agent-identity-for-india).

What this paper assumes

Some assumptions, so you can see where they might not fit.

We assume the agents you're dealing with have an identity that can be checked. That's a big assumption today, and it's the one we're working to make true. Saakshya is being set up to register the individual, company or institution behind each agent, check their documents and the

agent's details, and issue the agent an identity. Checking that identity works offline, through an SDK. Nothing else about how it works belongs in this paper, and we won't describe it.

We also assume the office wants to remain in charge of its own decisions. Everything below is a check the office makes for itself. None of it hands a decision to us or to anyone else.

And we assume a mix of requests. Some will come from agents that pass every check. Many, for a while, won't. The checklist tells the two apart.

The checklist

Six checks, in a fixed order. The order isn't decoration. Each check makes the next one meaningful, and an early failure means you can stop. Apply the same six to every agent request, including from an agent you've seen before, because the question is about this request and not about the sender's reputation.

First check: is the identity genuine?

Start with the agent's identity. The office checks it on its own side, offline, through the SDK, and the result is yes or no. Either the agent presented an identity and it checks out, or it didn't.

The failures here are plain. The agent presented nothing, or something the SDK can't make sense of, or something that doesn't check out. In all three, you stop. However good the rest of the message looks, you haven't been told anything you can rely on.

It's tempting to fill a missing identity with judgement: the wording sounds official, the business number is right. Don't. Those are things the sender wrote about itself. The first check asks whether anyone other than the agent has vouched for it.

Second check: is the person or organisation behind it identified and answerable?

An identity that checks out should lead to someone. We mean a named individual, company or institution, the owner, that registered the agent and can be reached. Not "the AI", and not "the platform" unless the platform is the owner.

The test is simple. If something goes wrong tomorrow, whom does the office write to? If the answer is a name and a way to reach them, the check passes. If the answer is "a software company in another country, through a support form", you've learned something worth knowing, and you should record it.

In our approach the owner is registered, and their documents are checked, before an agent gets an identity. That's a principle we hold, not something to assume of every system you meet.

Third check: does the scope cover this exact action?

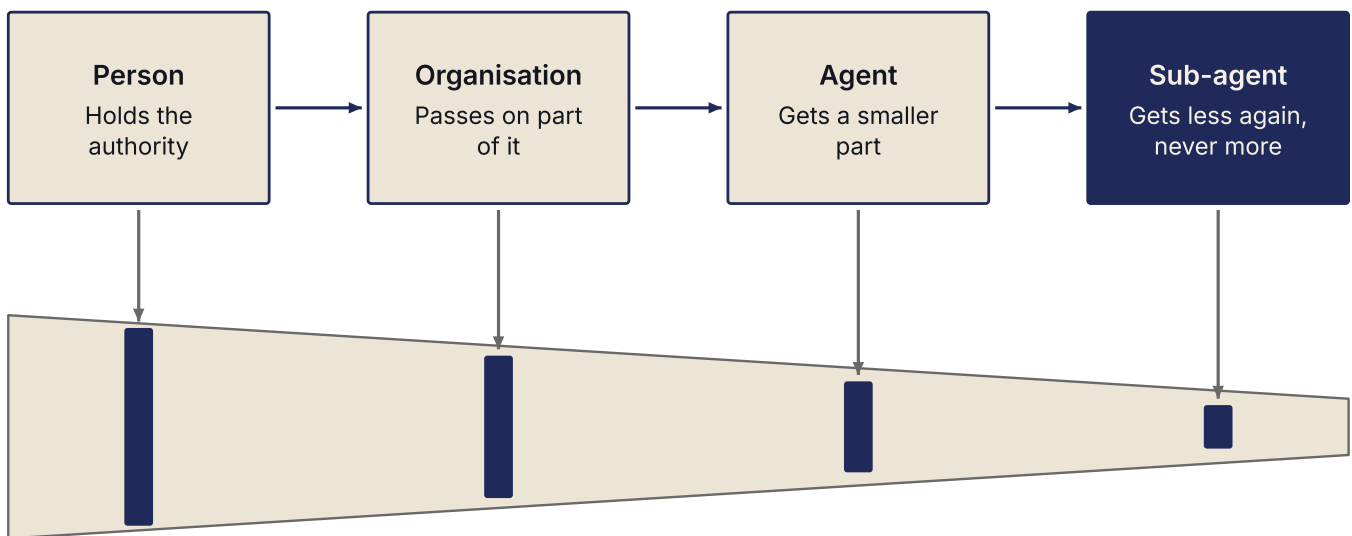
Now compare what the agent is allowed to do with what it's asking for right now. This is the check offices most often skip, because the earlier ones feel like enough.

Take our tax office. Permission to look at a registration is not permission to change it. Permission to prepare a return is not permission to file it. Permission to change a contact email is a different thing again from permission to change a bank account for refunds. The check is against this request, for this business, for this task. A general trust in the agent doesn't count.

Be literal. If the scope says "view", and the request is an edit, the answer is no, even if the edit looks small. If you can't tell what the scope covers, you haven't completed the check. We come back to that below.

Fourth check: does the delegation chain get narrower at each link?

Agents rarely act alone. A business owner gives some authority to an accountant. The accountant gives some of it to a firm's software. The software hands a task to a helper agent. Each of those is a link in a chain.



What each may do

Figure 1. Authority passes down from a person to an agent, and gets smaller at each step.

► Read this diagram as text

1. It starts with a person, who holds the authority to act.
2. The person passes part of that authority to an organisation.
3. The organisation passes a smaller part to an agent.
4. The agent can pass a smaller part again to a sub-agent.
5. At every step the band gets narrower, because nobody can hand on more than they were given.

The rule is that nobody can hand on more than they were given. Each link passes on the same or less. If you read down the chain and any link claims more than the one before it, the chain is broken, and you stop. A helper agent can't have more authority than the agent that engaged it.

There are two practical points. First, look at the whole chain and not only the last link. An agent that looks well-scoped may be acting on behalf of one that was never allowed to delegate this at all. Second, if a request plainly comes from a helper but nothing shows how it got there, you haven't seen the chain, and the check isn't done.

We've written about this at more length in [Local processing is not local authority](https://saakshya.org/articles/local-processing-is-not-local-authority) (saakshya.org/articles/local-processing-is-not-local-authority), which looks at the same shape from the phone's end.

Fifth check: what is the data-protection position?

The request will almost always involve personal data, either what the agent is asking for or what it's about to hand you. India's Digital Personal Data Protection Act, 2023 matters here. Take what follows as our reading, and ask your counsel.

We read the Act from the Gazette of 11 August 2023 (<https://egazette.gov.in/WriteReadData/2023/248045.pdf>) and the Rules from the MeitY copy of the notification of 13 November 2025 (<https://www.meit.gov.in/static/uploads/2025/11/53450e6e5dc0bfa85ebd78686cadad39.pdf>), both on 12 October 2026. A commencement notification (<https://egazette.gov.in/WriteReadData/2025/267647.pdf>) puts most of the Act, and most of the Rules, into force eighteen months after publication. On the Gazette's dateline that is 13 May 2027. So these duties are scheduled, not yet running, and an office that builds the habit now is ahead of them.

Three points matter at the desk.

First, the office stays responsible. Section 8(1) says a data fiduciary is responsible for complying "irrespective of any agreement to the contrary", for processing done by it or on its behalf by a processor. We read that to mean an office can't pass its duties to an agent by saying the agent asked for it.

Second, ask whether the data being requested or handed over is what the stated purpose needs. Section 6(1) says consent is to be limited to "such personal data as is necessary for such specified purpose". An agent that asks for a whole file when it needs one field is a warning, and the office can say no to the extra.

Third, ask where the data goes next. Section 11(1)(b) gives a person the right to learn the identities of other fiduciaries and processors their data was shared with. If the agent is going to pass what you give it to a third party, the office should be able to say who, because the person may ask.

The check, in practice, is a short question put to yourself and written down. Is the personal data in this request needed for the purpose? Do I know who it goes to? Have I told the person what I do with it, where I need to? If you can't answer, you haven't completed the check.

We've gone through related ground in Read-only is a permission (saakshya.org/articles/read-only-is-a-permission), including what a person is entitled to ask afterwards, and in Who is the relying party, and what should it check? (saakshya.org/articles/who-is-the-relying-party), an earlier and shorter

piece on the same idea. This paper's list differs a little from it, and where they disagree, this one is the later view.

Sixth check: is evidence written before acting?

Before the office acts, it writes down what it saw and what it decided. The order matters. A record written afterwards can quietly turn into a story about a decision, told by someone who already knows how it turned out.

At a minimum, we'd write down the time, the identity presented, the owner it names, the scope it showed, the chain it showed, which of the six checks passed, and the decision with a one-line reason. Write that down whether the answer is yes or no. Refusals are the records you'll be asked about.

Keep the record light. A log full of copied personal data is a liability. Note that the data was needed and what for, and leave the data where it already lives.

On retention, the Rules say something worth knowing. Rule 6(1)(c) lists, as a minimum security safeguard, "visibility on the accessing of such personal data, through appropriate logs, monitoring and review". Rule 8(3) says a data fiduciary must keep personal data, associated traffic data and other logs of the processing for a minimum of one year from the date of processing, for the purposes listed in the Seventh Schedule, unless another law requires more. We read those as applying to the office's own processing logs. Whether your agent-check records count is a question for counsel, and we don't know the answer.

When a check can't be completed

If any check can't be completed, the answer is no. A lookup timed out. A document is missing. The scope is written in a way nobody can interpret. The chain stops halfway. In every case, the office refuses. We call that failing closed.

It will annoy people. A request that would have been fine gets turned away because something was missing. We think that's the correct trade for anything involving money, health or tax. The alternative is a default of yes, and a default of yes means the checks work only against agents that were already behaving.

A few details make refusing bearable.

Say which check failed, in words the sender can act on. "Rejected" teaches nobody anything. "We could not confirm the person behind this agent" tells an accountant exactly what to fix.

Have a path to a human. A refusal isn't the end of the matter. A person at the office can review a request, ask the sender to try again, or take the work some other way. What we mean by failing closed is that the agent's request doesn't go through on the strength of a guess. A person making a decision is a different thing.

Keep an emergency route that doesn't involve the agent at all. The Act recognises that processing can be necessary to respond to a medical emergency. Its section 7(f) lists "responding to a medical emergency involving a threat to the life or immediate threat to the health of the Data Principal or any other individual" among the legitimate uses. That is about what the office may do with personal data. It doesn't tell a hospital to take an unchecked agent's word, and it doesn't change who the agent is. If someone's life is at risk, the hospital should be talking to a person on the phone.

And record every refusal. They tell you which senders to help, and they tell a regulator, later, that the checks did something.

Why a model's confidence score doesn't belong in the decision

It's tempting to add a seventh check: a number that says how trustworthy the agent seems. Some systems produce one from the agent's writing style, its past behaviour or the model's own sense of how sure it is.

We'd keep any such number out of the decision path altogether, and that includes using it to settle a close call.

A score is a guess. The guess is made by a system the office doesn't control, usually the agent's own side. When a regulator asks why the office acted, "a model said 0.93" is not an answer the office can defend, because nobody on the office's side can say what the number means, how it was produced or whether it would be the same tomorrow.

A score can also be gamed. A sender who knows the office watches one has every reason to shape what it sees. And a score doesn't tell you whom to ring. When something goes wrong, you want a person who answers, and a probability can't be written to.

None of this says scores are useless. An agent's own operator might find one helpful for monitoring its own agents, and that is the operator's business. Our point is only about where it goes. The office's yes or no rests on facts someone else has vouched for, and on the office's own rules.

Adopting the checks in phases

No office can turn all this on in a day. Trying to would make a pile of refused requests and angry customers, and the checks would get blamed. We suggest three phases.

Observe. The office runs the six checks on incoming agent requests and writes down the results, but carries on exactly as it did before. Nothing is refused. After a while the office knows what share of requests would have failed, at which check, and from which kinds of sender. This is the simplest way to find out whether your front door has a problem at all.

Warn. Requests that fail still go through, but the clerk sees a notice, and the sender gets a message saying what was missing. This is where senders learn. We expect many senders to fix things quickly, because we expect most failures to be mundane: a scope written too broadly, an owner not named.

Enforce. Requests that fail are refused, with a reason the sender can act on. By now the office knows how many of those to expect, and the people at the desk know what to say.

We can't say how long each phase should last. It depends on how many agent requests you get and how fast your senders adapt. We have no figures, and we won't make one up. Skipping the first phase is the most common way to make the last one painful, and we say that from reasoning, not from watching it happen.

Decide, before you start, who owns the result. Somebody should read the observation records, decide when to move on and answer the senders. A checklist with no owner turns into a document.

Notes for four kinds of front door

We don't have a list of sector rules, and we haven't invented any. These notes are what we'd think about at each kind of desk. Your own compliance team, and your regulator's published requirements, come first. We haven't cited a sector regulator here because we haven't read one closely enough to quote fairly.

Banking. Requests that move money, open credit or trigger a credit enquiry carry the heaviest consequences, so the third check, scope, does most of the work. A request to read a balance and a request to apply for a loan are different acts and need different scopes. A particular risk is the same agent approaching several banks for the same person. Each bank sees one request, so each bank's own record, written before acting, is the only trace it will have.

Health. The mistakes here can hurt a person, so we'd lean toward a human on any request that changes care: moving an operation, changing a medicine, sharing a record. Scheduling and administration are different from clinical content, and the scope should say which one the agent has. A patient's family often acts for them informally, but an agent working for a family member should still name its owner. The emergency point in the section above applies most here.

Tax. Filing and drafting are different acts, and a very common failure is an agent allowed to prepare something that then files it. The office has to read scope literally. Tax offices can also get requests from professionals acting for many clients at once, so the fourth check, the chain, matters: the accountant, the firm and the software may all be links.

Government counters. Offices that give out certificates, benefits or permits deal with the widest range of senders, from large firms to individuals helping their parents. Some counters are small, with little technical support. The checks must fit a person at a desk. If the office has to install a whole new system before it can run them, many counters won't. We think the work of the next year is making them light enough for that desk.

Questions we have not answered

What we don't know about the checks

We don't know whether six is the right number. An office may tell us two of these belong together, and a particular sector may want a seventh. We'd like to hear which.

We don't know how long the observe phase should last, or how many failures an office will put up with before it decides the checks are too strict. That depends on the sector, and we have no figures.

We don't know how a small clinic with one receptionist can do any of this. The checks have to be light enough for a desk with no IT department, or only large offices will use them, and we haven't worked out how.

We don't know how much detail to keep in the written record. Too little and it can't be defended. Too much and it becomes a store of personal data nobody wanted. We lean towards less, and we expect regulators to say more.

Open questions about the law and about adoption

We don't know whether the data protection law will treat an agent's request as the person's own request, a request from a representative, or something else. Neither the Act nor the Rules uses the word "agent" in this sense, as far as we can tell, and the answer matters to the fifth check.

And we don't know whether offices will run the checks at all. A desk with a queue is tempted to accept the request and hope. Whether checking becomes routine depends on how easy it is and on what regulators expect. Neither is up to us.

How to respond

Write to contact@saakshya.org. We aim to reply within a week.

If you run a front door at a bank, a hospital, a tax office or a government counter, we'd most like to know which check you'd drop, which you'd add, and which would be hardest to do with the people and tools you have. If you're a lawyer or work for a regulator, tell us where our reading of the Act is wrong. If you build agents, tell us what would make the checks easier for you to pass.

Saakshya is independent of NPCI, UIDAI, MeitY, RBI, W3C and every company named in these papers, and none of them has endorsed this checklist. This is version 0.1, dated 12 October 2026. When we change our mind, the earlier version stays up with a note saying what changed.

How to cite this paper

PLAIN TEXT

Saakshya (2026). A relying party's checklist for AI agents (Version 0.1). Saakshya Position Paper 2. <https://saakshya.org/papers/relying-party-checklist>

BIBTEX

```
@techreport{saakshya_relying_party_checklist_v0_1,  
  title      = {A relying party's checklist for AI agents},  
  author     = {Saakshya},  
  institution = {Saakshya},  
  year       = {2026},  
  number     = {Position Paper 2},  
  type       = {Position paper},  
  version    = {0.1},  
  url        = {https://saakshya.org/papers/relying-party-checklist},  
}
```

Version history

Version 0.1 · 12 October 2026

First version.
