What you can actually require

Version 0.2 · revised 8 September 2026 · this version as a PDF

Twelve things an institution or a government can require before an AI agent acts on its behalf, what is already mandatory in New Zealand, and the identity control that was written into a standard and then left out of every baseline.

The pursuit of 'Goodness' in AI — Sovereign agents, and what it takes to trust one with real work

---

The question a policy reader arrives with

FIG-30 THE TWELVE, AND WHERE EACH IS ALREADY MANDATORY

	WHAT AN INSTITUTION CAN REQUIRE BEFORE AN AGENT ACTS	REQUIRED ANYWHERE TODAY?
1	A locally governed identity authority for people, workloads and privileged services	NO
2	A unique agent workload identity with short-lived credentials, not static keys	NO · IA-9 is in no baseline
3	Authorisation externalised to a gateway; the model is never the authorisation engine	NO
4	Default-deny tools and plugins, with a signed inventory and approval per version	NO
5	Capability-based, purpose-bound delegation rather than inherited user or admin access	NO
6	Read, draft, commit, send, publish, delete and administer as distinct permissions	NO
7	No unrestricted browser, shell, SQL, filesystem, outbound HTTP or cross-tenant retrieval	NO
8	No autonomous installation or update of agents, models, plugins, policy or credentials	NO · advised, not required
9	Local, exportable, tamper-evident audit records	PARTLY · Art. 26(6), high-risk
10	Human step-up approval for communications, publication, sensitive disclosure, access changes	PARTLY · Art. 14(4), 26(2)
11	Immediate revocation of identity, tool grants, sessions, network paths and secrets	PARTLY · a stop, not revocation
12	A tested offline or degraded-mode plan	PARTLY · suspension on risk

Eight of the twelve are required nowhere. Row 11 is marked partly because the law requires a stop for the SYSTEM, which is not the same as revoking a credential.

FIG-30 The schedule with its status column shaded rather than tabulated, because the shading is the finding: eight of the twelve are required nowhere, and the four that do carry qualifications — high-risk only, from 2027, cloud services in Europe. None of the twelve waits on new legislation.

Somebody responsible for a portfolio, a council, or a department does not need to be persuaded that AI governance matters. They need to know what to write down — what can be required, of whom, under what existing power, and what happens if it is not met. Almost everything published on this subject answers a different question, which is what an organisation might voluntarily consider.

This piece states twelve requirements in the form they would take in a procurement schedule or a chief executive's instruction, and it is specific about which of them are already law somewhere, which are guidance, and which are required nowhere at all. Eight of the twelve are in the last category. That is the finding, and it is a stronger argument for acting than any of the twelve individually.

What is mandatory in New Zealand

Very little, and the instruments say so themselves.

The **Public Service AI Framework**, issued by the Government Chief Digital Officer in January 2025, states its own force plainly: **"Agencies are encouraged to align with the direction set by this Framework, however it's not binding."** Its fifth principle is the right one — **"AI use within the Public Service should be subject to oversight by accountable humans with appropriate authority"** — and *should* is the operative word.

The **Algorithm Charter for Aotearoa New Zealand**, from July 2020, is **"a commitment by government agencies"** that signatories apply against their own risk rating: the Charter *could* be applied at low risk, *should* at moderate, *must* at high — with the agency doing the rating. It contains a human-oversight commitment worth having, including **"providing a channel for challenging or appealing of decisions informed by algorithms"**. It also records its own boundary: **"it cannot fully address important considerations, such as Māori Data Sovereignty."** ⚠️ Whether it remains operative in 2026 is not confirmed here, and should be checked before anyone relies on it.

The **Office of the Privacy Commissioner's** 2023 expectations are guidance, but they rest on binding principles and one of them is unusually direct about the failure mode this series has been describing: **"Simply having a 'human in the loop' may not be enough to uphold the accuracy principle, given the well-known problem of automation blindness."** The OPC also expects agencies to **"have senior leadership approval based on full consideration of risks and mitigations"** before deployment.

Binding, and narrow: the **Biometric Processing Privacy Code 2025**, a code of practice under the Privacy Act, commenced November 2025.

And the Privacy Act itself contains no rule about automated decisions at all. Searched in its consolidated form, **"automated" appears zero times** and **"algorithm" once**, inside the information-matching provisions. There is no right to human intervention, no threshold at which oversight becomes mandatory, and no equivalent to Article 22 of the GDPR. The Privacy Commissioner said as much in December 2025: **"We also need stronger protections for the significant privacy risks that arise from automated decision-making."**

An institution in Aotearoa is therefore operating without a floor, and whatever it requires it requires on its own authority. It is the reason a schedule like the one below can be adopted immediately, by a board or a chief executive, without waiting for anybody.

NIST IA-9, and the baselines

NIST Special Publication 800-53 — the control catalogue underpinning United States federal security compliance and widely borrowed elsewhere — contains control **IA-9, Service Identification and Authentication**: uniquely identify and authenticate organisation-defined system services and applications before establishing communications with devices, users, or other services.

That is, almost exactly, requirement 2 of the twelve below. It was written. And **it appears in no baseline**: not Low, not Moderate, not High, not the privacy overlay, not the operational-technology overlay. An organisation bound by those baselines is not required to select it and will not select it by default.

A control that exists and is optional is a more interesting state of affairs than an absent one, because it means the requirement was understood, drafted, and then placed where nothing would pull it in — and it disposes of the objection that requiring identity for non-human actors is novel or unreasonable.

The rest of the standards landscape is consistent with that. SPIFFE specifies a workload identifier containing no notion of ownership or accountability. NIST SP 800-63-4, finalised in July 2025, states its scope as the identity of users. ⚠️ NIST's control-overlay work for AI, announced in 2025, names single-agent and multi-agent use cases and says the overlays **"will assume that certain controls are already in place"** — including identification and authentication. The overlays for agents had not been published at the time of this search.

The same twelve at every scale

These requirements are not written for a particular size of organisation, and that is a consequence of the argument rather than a convenience. The questions underneath them are constitutional — who granted this authority, what are its limits, who may withdraw it, to whom is the holder answerable, what record establishes what was done. A cabinet and a golf club committee both grant authority, both can be asked afterwards for their reasons, both can be overruled by the body that constituted them, and both carry it if the thing goes wrong.

What each requirement *looks like* differs enormously. Requirement 2, that an agent hold an identity of its own, is a national identity-federation programme in one setting and a separate login on the club's booking system in another. Requirement 10, human approval before consequential action, is a statutory instrument in one and a rule that the treasurer sees anything touching money in the other. **The shape is the same. The implementation is whatever an organisation of that size can actually operate**, and a requirement an organisation cannot operate is a requirement it will document and ignore.

What does not scale is purchasing power. A club secretary can ask a supplier the five questions in Who actually holds the controls and will generally receive no useful answer, because the club is not a customer whose departure would be noticed. A government asking the same questions in a procurement rule receives answers, because the answers are a condition of a market worth having.

That asymmetry is the argument for national action, and it is not paternalism. A state should require these things not to protect institutions from their own choices, but because **only a buyer of that size can cause products to exist that are capable of answering** — and once such products exist, every organisation below inherits the answer without needing any purchasing power of its own. The club benefits from a procurement rule it was never party to, in the same way it benefits from electrical standards it did not negotiate.

The converse is the position most small organisations are in today. The questions are askable, the answers are unavailable, and the reason is not that anybody refuses. It is that nobody with sufficient purchasing power has yet asked, so no supplier has had to build the capability to answer.

How long a requirement has to last

The purchasing argument above has a condition attached, and without it the argument fails quietly.

Procurement changes what suppliers build. But building the capability to answer these questions — issuing an identity to a workload, externalising authorisation, producing an exportable tamper-evident record — is a multi-year engineering commitment made against an expectation of demand. A supplier weighing that investment against a requirement that may lapse at the next election does not refuse. **It waits.** Waiting is free, costs nothing in reputation, and is the correct commercial judgement if the requirement's expected life is shorter than the build.

So a procurement rule adopted by one government and reversible by the next does not produce products capable of answering. It produces compliance theatre for the duration, and the capability is never built — which is indistinguishable, from the outside, from the requirement having been unreasonable.

The condition on the whole argument is therefore durability across a change of government, and that is a structural requirement rather than a political preference. It is the same condition this series places on any institutional value: that it can be changed deliberately, by whoever is entitled, rather than eroding through the ordinary turnover of the people who happen to hold office. At the level of a firm that means a minuted decision with an expiry. At the level of a country it means a commitment that survives an election, because from the perspective of any institution or supplier relying on it, a policy that changes with each government changes by drift.

The question has been worked through on this site from four directions, and each addresses a different way durability fails. They are worth reading in place of this summary.

The horizon. Our own machines states the difficulty plainly:

> AI is the clearest test yet of whether New Zealand's politics can lift its attention from the electoral cycle to the country's foundations.

It also records that the work has begun. In June 2026 an independent cross-party proposal, *In Our Own Hands*, was offered to every party in Parliament — seven commitments and a three-phase pathway for AI under New Zealand authority. The first of those commitments is the one this series has been arguing toward from a different direction: **"Authority stays here — New Zealanders keep rightful authority over the AI that runs public services and holds their data."**

And on the mechanism, that framework reaches the same conclusion as this document by an independent route, which is worth noting because agreement arrived at separately is worth more than either argument alone:

> MBIE amends the Government Procurement Rules — an executive-level act, done before without legislation — to insert the cross-party test pair: jurisdiction of inference (where does the AI actually run, and whose court order does it answer?) and data residency…

> This is deliverable inside a parliamentary term, and it converts procurement from a purchasing function into the enforcement lever for everything else in this document.

The money. A requirement survives if the funding behind it does, and the instrument matters more than the amount. Nine changes government could make sets out why:

> Committed revenue is more bankable than a grant, survives a change of government better than a line item, requires no novel structure, and — critically — does not put the Crown into the governance of the entities it buys from.

That last clause is the one this series cares about. A state that funds by grant acquires influence over the governance of what it funds, which is the failure mode the whole argument here is built to avoid: authority migrating to whoever holds the purse. Committed revenue for a service actually wanted buys the capability without buying the governing body.

The custodian. A commitment held by whoever is currently in office lasts as long as they do. Democratic AI locates it correctly, and the observation applies to everything in this document:

> Which is the part that should concern a permanent secretary more than a minister: it survives the change of government.

A requirement written into procurement rules, administered by officials, reviewed on a schedule and reported to a committee is held by the permanent part of the state. One announced as policy is held by the temporary part. The difference is not ceremonial: it determines whether the requirement is still there when the supplier's investment decision matures.

The time horizon, and it is longer than anybody's term. Why records signed today may not survive is about cryptography rather than politics, and it establishes the interval the argument actually has to span. What that piece finds New Zealand lacks is:

> A timestamp authority that is independent, durable across decades, and recognised in evidence here.

Decades. A school, a council or a firm holding records under a retention obligation is making a commitment measured in decades, discharged by systems procured under rules made in a three-year cycle. The mismatch is arithmetic, and it is why the ninth requirement below — a local, exportable, tamper-evident record — is the one whose absence is hardest to repair afterwards. A missing control can be added. A decade of records that cannot be shown to be what they claim cannot be reconstructed.

Three things follow for a reader of the present piece. The first is that the twelve requirements are not a proposal awaiting a legislative programme: the instrument that would carry them can be amended by an executive act, there is precedent, and it is achievable within one term. The second is that achieving it within one term is not sufficient. **A test pair inserted in 2027 and removed in 2030 will have changed nothing, because no supplier will have built for it.** The third is that the durability has to be constructed rather than hoped for — a funding instrument that survives a change of government, a custodian in the permanent part of the state, and a horizon set by the retention obligations of the institutions relying on it rather than by the parliamentary calendar.

⚠️ *In Our Own Hands* is described here from Our own machines, which records it and states where it agrees and departs. This piece has not examined the proposal independently, and a reader relying on its contents should go to that framework and to the proposal itself rather than to this summary.

The twelve

Written as they would appear in a schedule. The status column states what is required somewhere today, not what ought to be.

⚠️ **What "required" counts.** The column below scores each requirement against states, standards bodies, and Te Kāhui Raraunga's *Māori AI Governance Framework* — a requirement set by a body with standing in Aotearoa is not commentary, and scoring only states and standards bodies would treat it as though it were while treating one set by a European regulator as law. The framework's requirements are noted in the column where they bear, and its standing is a matter for the bodies that hold it rather than for this table to rank.

	REQUIREMENT	REQUIRED ANYWHERE?
1	A locally governed identity authority for people, workloads and privileged services	**No**
2	A unique agent workload identity with short-lived credentials, not static keys	**No** — NIST IA-9 exists and is in no baseline
3	Authorisation externalised to a policy-enforcing gateway; the model is never the authorisation engine	**No**
4	Default-deny tools and plugins, with a signed inventory and approval per version	**No** — the UK's transparency standard requires a public record of algorithmic tools, which is an inventory rather than a deny-list
5	Capability-based, purpose-bound delegation rather than inherited user or administrator access	**No**
6	Read, draft, commit, send, publish, delete and administer as distinct permissions	**No**
7	No unrestricted browser, shell, SQL, filesystem, outbound HTTP or cross-tenant retrieval	**No**
8	No autonomous installation or update of agents, models, plugins, prompts, policy or credentials	**No** — the UK Playbook advises caution; nothing requires it
9	Local, exportable, tamper-evident audit records	**Partly** — EU AI Act Art. 26(6) requires logs for at least six months, high-risk only, applying from 2027
10	Human step-up approval for communications, publication, sensitive disclosure, access changes, deletion, payments, agreements and production changes	**Partly** — EU Art. 14(4) and 26(2), high-risk only; US OMB M-25-21 for high-impact federal AI; Canada's Directive at its higher impact levels
11	Immediate revocation of identity, tool grants, sessions, network paths and secrets	**Partly** — Art. 14(4)(e) requires a stop for the *system*, not revocation of a credential
12	A tested offline or degraded-mode plan	**Partly** — Art. 26(5) requires suspension on risk; Canada requires contingency at higher impact levels

⚠️ Canada's **Directive on Automated Decision-Making** is the closest existing thing to a tiered mandatory regime, with four impact levels and requirements escalating across them. Its Appendix C could not be retrieved for this piece — the department's site refused automated access — so nothing from it is quoted here and its requirements are described only in outline from the department's own summary pages. Anyone building on this should open it directly.

Forbidden, and unavailable

A requirement that can only be broken by someone ignoring a document is a different kind of requirement from one that cannot be broken at all. That is the reason the twelve are worth writing down rather than leaving them as principles: **ten of them describe an arrangement in which the prohibited thing is not available to be done, rather than one in which it is available and forbidden, and an eleventh does so in part.**

That is not a claim that they are easy, or that they are common. It is a claim about their form. Each names a point at which a request is refused by whatever is standing there, whether or not anyone is reading a policy that day, and whether or not the policy still exists.

	WHAT STOPS IT
1	The authority issues the credential or it does not. An identity it never issued cannot be presented, and one it will not renew stops working without anybody intervening.
2	A short-lived credential expires on a clock. A leaked one is worthless within its window, and the window is a number somebody chose rather than a practice somebody follows.
3	The gateway sees the call before the resource does. A model that is not the authorisation engine cannot authorise anything, however it is prompted, because nothing downstream is listening to it for that.
4	An unsigned tool does not load. Approval per version means the artefact that runs is the artefact that was examined, and a substituted one fails its signature rather than passing unnoticed.
5	A capability that was never delegated is not held. The request has nothing to present, so the refusal happens at the resource and needs no rule to be consulted.
6	Separate permissions are separate grants. An agent holding draft and not send cannot send; there is no configuration in which the two are the same grant.
7	An egress path that is not open is not open. A shell that was never attached cannot be reached by persuading the model to ask for one.
8	Write access to the artefact store is withheld. An agent that cannot write to the place its own updates come from cannot update itself, whatever it concludes about the merits.
9	Append-only storage makes an alteration visible rather than preventing it. **This one is partly documentary** — the record is structural, and somebody choosing to read it is not.
10	The action sits behind a second credential the agent does not hold. Step-up approval that the agent can satisfy on its own is not step-up approval.
11	Revocation at the authority ends every session that depends on it. The question to ask a supplier is how long, measured, not whether it is supported.
12	**Nothing. This one is documentary and cannot be made otherwise.** A degraded-mode plan is a practice: it can be written and never tested, and the day it is needed is the day that is discovered. It is in the list

	WHAT STOPS IT
	because leaving it out would imply the other eleven are sufficient.

One of the twelve is documentary, and one is half. The argument against governance by document is not that documents are useless — the twelve are a document — but that a document is the *weakest* available form for ten of these, and it is the form nearly all of them are currently left in. Where a requirement can be made structural and is left documentary instead, somebody has chosen the weaker option, and the choice is usually invisible because both versions read identically in a policy.

⚠️ **None of this says an institution holding all twelve has a good AI system.** It says the institution's judgement about what is good can reach what its software does. Whether that judgement is any good is the question this series says nobody else can answer for it.

The lever that already exists

A government wanting these twelve in force does not need an AI Act. It needs a procurement schedule, and the reason is structural: every one of the twelve is a property of an arrangement between an institution and a supplier, and procurement is where that arrangement is written.

Two existing instruments show the shape. The European Commission's model contractual clauses for AI procurement exist in a full version aligned to the AI Act and a lighter version for systems outside it — voluntary templates, but templates that make the requirements concrete and quotable. And the **EU Data Act**, at Article 23, is binding: it requires providers to remove **"pre-commercial, commercial, technical, contractual and organisational obstacles"** to a customer terminating, switching, and porting exportable data and digital assets **including to on-premises infrastructure**, with functional equivalence. That is the eleventh and twelfth requirements above, in force, for cloud services in Europe.

New Zealand's Government Procurement Rules reached their fifth edition in December 2025. ⚠️ Whether they contain any AI provision is not established here; the rule text could not be retrieved. If they do not, that is the shortest path from this piece to something enforceable — a schedule appended to the rules binds every agency that buys, takes no legislation, and reaches suppliers directly rather than through the deploying agency's good intentions.

What twelve requirements will not do

A schedule is worth no more than its account of what it does not reach, and this one does not reach three things.

It does not make an agent safe. Every requirement here concerns authority — who permitted what, and what can be shown afterwards. An agent operating entirely within its permissions can still be wrong, and nothing in this list catches that. It bounds the damage; it does not prevent the error.

It does not bind the model provider. Requirements 1 to 8 are enforceable against a deploying institution and largely enforceable against a supplier through a contract. None of them stops a provider changing model behaviour, altering what a system will refuse, or withdrawing a capability. That is the gap identified in Who actually holds the controls and a procurement schedule narrows it without closing it.

And it does not address Māori data sovereignty, which is not a subsection of this schedule and should not be written as though it were. Te Kāhui Raraunga's *Māori AI Governance Framework* sets out requirements in that domain from a body with the standing to set them, including that decommissioning is a partnership decision and that **"AI systems must not be implemented in Aotearoa without fully realising Māori authority over Māori data."** The Algorithm Charter records the same limitation about itself. An institution adopting these twelve has not discharged that obligation and should not tell itself otherwise.

How you would know this is wrong

First, on the central negative claim. Two searches were run for a mandatory requirement that an AI agent hold its own identity, both keyword-based over standards and regulatory corpora, which means they shared an assumption and are not two independent methods. **The finding is: no such requirement was found. It is not: none exists.** Singapore's agentic governance framework is reported to require a verifiable identity per agent; that could not be confirmed from the framework text, and if it does, it is the first and this piece must say so. Anyone who knows of one is asked to say so through the contact route on the about page.

Second, if institutions adopting requirements of this kind turn out to govern their agents no better than those that do not — because the requirements are met on paper and evaded in practice — then this is a compliance exercise with a good conscience, and its authors would rather know.

Third, if a jurisdiction adopts something equivalent and it produces documents rather than outcomes, the diagnosis in this series is wrong at the level that matters. The prediction being made is specifically that requirements attached to *procurement* behave differently from re-

quirements attached to *conduct*, because the first reach the supplier and the second do not. If that turns out to be false, it is the load-bearing claim of this piece and its failure should be reported here.

Related

The four properties underneath sets out the four properties these twelve requirements are attempting to secure, and why each of them can be checked rather than asserted.

The six, in reading order. Each stands on its own; read together they build one argument.

- What has to be settled first — why the prior questions are prior, and what each of the others answers
- Who actually holds the controls — five questions establishing whether an authority to act actually exists
- How much it may do unsupervised — five levels of permitted autonomy, declared rather than left implicit
- The four properties underneath — four properties a delegation must carry, each checkable by observation
- What nobody has measured yet — four open questions, and the observations that would settle them
- An invitation to become a Distributor — not part of the argument above: a proposal to work with its author, for anyone who wants to take this further

> ACKNOWLEDGEMENT
>
> **Leslie Stroh** — my big brother, a product philosopher, and a magnificent role model. He helped me find my way to a vision of goodness in AI.

▸DISCLAIMER

> **How this was made.** The version number counts drafts of the text. It does not measure the inquiry behind it, which has run over days and across several AI systems, with argument between those systems and within them, directed, refused and repeatedly redirected by the author. The source material was AI-generated, and then adversarially and iteratively refined across a range of tools — systems built by different companies in different jurisdictions, set against each other and against the author. No one of them produced this text, and no one of them reviewed it alone. **The plurality is deliberate rather than incidental.** A single model carries a single set of priors about which sources are authoritative, and this series argues that an evidence base narrowed in exactly that way is how a contested question comes to look settled. Using one model to investigate that claim would have been the claim refuting itself. To name a single model on it would credit that model with work that was neither its own nor done in a single pass. The diagram on each page was drawn afterwards, with the same assistance, from the argument the essay had already made. **The plurality was also necessary, and the record should say why.** In drafting, the assisting model repeatedly led with United States institutional sources — a national laboratory, an industry association, a market study nineteen years old — and presented conclusions drawn from them as the state of knowledge. On one occasion European measured data contradicting those conclusions was present in the same research return and was placed below them. Framings were proposed that would have argued against this series' own position using that evidence base, and offered as rigour. Each was refused by the author and the material rebuilt. That is the mechanism these documents describe, occurring in their own making, and it is recorded because a series arguing that evidence bases narrow without anyone deciding to narrow them cannot credibly claim its own production was exempt. The framing, the corrections and the judgements are the author's, and so are the errors.

Cite this version

```
What you can actually require, version 0.2 (September 2026).
agenticgovernance.digital/downloads/what-to-require-v0-2.pdf
```


PREVIOUS
The four properties underneath

Series contents

NEXT
What nobody has measured yet

Alongside: questions and answers · sources and provenance · slides

Share this Mastodon Email Copy link

Found something wrong?

This piece names what would show it is wrong. If you have that — a figure, a source, an argument — it is worth more to us than agreement. Tell us which claim and why.

Which section, if you know it

for example: 2. Four properties

The claim you are responding to

quote it, or describe it

What is wrong with it

Your email, only if you want an answer

Used to reply to you and nothing else. No list, no newsletter, never passed on.

Send this