How much it may do unsupervised

© John Stroh

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

Five declared levels of permitted autonomy, the criteria that set which one applies, and the published objection that a ladder of this kind is the wrong instrument entirely.

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

---

The question, and the phrase that stands in for it

FIG-28 WHAT "HUMAN IN THE LOOP" CONCEALS

human in the loop → FOUR ARRANGEMENTS IT COVERS EQUALLY

Describes a diagram, not an arrangement.

- ✗ approves forty actions an hour without reading them
- ✗ receives a notification after the fact
- ✗ could intervene, has never been told which actions warrant it
- ✓ reads this action, and decides it

THE QUESTION THAT HAS A WRITTEN ANSWER

Which classes of action may this agent perform without a named person approving that particular action — and who decided?

Answerable in writing. Auditable afterwards. Five declared levels, from advisory through to delegated autonomy.
An undeclared level cannot be audited, which is the reason for naming one at all.

The gate fails at the middle levels. In New Zealand's Privacy Act 2020, "automated" appears zero times and "algorithm" once.

FIG-28 One phrase covering four arrangements, three of which provide no check at all: approving forty actions an hour unread, being notified afterwards, and being able to intervene without knowing when to. Replaced by a question that has a written an… action may proceed without a named person approving that one, and …

📖 Ask the Glossary

Almost every published assurance about AI oversight reduces to the phrase *human in the loop*, and the phrase is close to useless because it describes a diagram rather than an arrangement. A person who approves forty actions an hour without reading them is in the loop.

So is a person who receives a notification after the fact, and a person who could intervene in principle but has never been told which actions warrant it.

The useful question is narrower and answerable: **which classes of action may this agent perform without a named person approving that particular action, and who decided?** That is a question with a written answer, and an institution either has one or does not.

This piece sets out five levels of permitted autonomy, the criteria that decide which level applies to which class of action, and — at length, because it is the strongest objection in the literature — the argument that ordering autonomy on a scale of this kind is itself a mistake.

The scales that already exist

Three published scales cover this ground, and the tiers below are a fourth. Saying so first matters, because a reader who meets the fourth without the other three cannot judge what it adds.

Feng, McDonald and Zhang define five levels by the role the person retains: operator, collaborator, consultant, approver, observer. Published in the Knight First Amendment Institute's essay series with a single anonymous reviewer — an essay rather than a peer-reviewed paper, which is worth stating because it is frequently cited as though it were the latter. Their observation about the fourth level is the one that matters here: **"While an agent may be instructed to seek approval prior to taking consequential actions, reliably determining which actions are consequential may be challenging."** And, on the failure mode of approval itself: **"User disengagement can lower the care with which they approve actions and can result in unintended approvals."**

Morris and colleagues at Google DeepMind give six, from "No AI" through "AI as a Tool", "Consultant", "Collaborator", "Expert" and finally "AI as an Agent — fully autonomous AI". A preprint, widely cited.

Mitchell, Ghosh, Luccioni and Pistilli at Hugging Face give five, ordered by how much of the program flow the model controls, from a model with no influence over flow to one that "creates and executes new code". Their argument is not that the top level needs governing: **"we argue fully autonomous AI agents… should not be developed"**, on the ground that **"risks to people increase with the autonomy of a system."** They are the only one of the three willing to say a level should not exist.

The canonical precedent for all of them is **SAE J3016**, the driving-automation levels, and it contains a feature the AI scales have not carried over. In the standard's own framing a vehicle moves between levels according to the driving task and the circumstances at that moment. The level is a property of the task in context, not a fixed attribute of the machine. That

distinction is the whole of what the tiers below add. ⚠️ J3016 is behind a paywall; the level definitions here are as reproduced in secondary references citing it, not read from the standard.

The objection to scales of this kind

The strongest published argument against everything in this piece is **Bradshaw, Hoffman, Johnson and Woods, "The Seven Deadly Myths of 'Autonomous Systems'", IEEE Intelligent Systems, May/June 2013**. Two of their seven myths are aimed directly at the instrument proposed here.

Myth one: **"'Autonomy' is unidimensional."** Myth two: **"The conceptualization of 'levels of autonomy' is a useful scientific grounding for the development of autonomous system roadmaps."** Their case, quoting the US Defense Science Board, is that a levels taxonomy is routinely misread as implying **"that autonomy is simply a delegation of a complete task to a computer, that a vehicle operates at a single level of autonomy and that these levels are discrete and represent scaffolds of increasing difficulty."** Their own verdict is blunter: **"levels of autonomy encourage reductive thinking"**, and autonomy **"isn't a discrete property of a work system… it's an idealized characterization."**

They are right about the failure mode and I have seen no rebuttal that meets them. A single ordered ladder invites exactly the reasoning they describe: that an organisation can be *at* level three, that level four is more advanced, and that progress means climbing. Every one of those inferences is wrong, and the ladder makes each of them easy.

The concession is not rhetorical, so it should be stated at its full cost. **The five tiers below are a single ordered scale, which is the thing Bradshaw and colleagues say should be discarded.** NIST's own ALFUS framework, written for unmanned systems, splits autonomy across three axes — human independence, mission complexity and environmental complexity — precisely to avoid collapsing it into one. A three-axis treatment is more faithful to what autonomy actually is.

The reply is that the tiers are not measuring autonomy. They are **allocating permission**, and permission is one-dimensional in the only sense that matters: an action either requires a named person's approval before it happens, or it does not. That is a binary at each action class, and five tiers are five useful groupings of those binaries. Bradshaw and colleagues are describing a scale that purports to characterise a system; this one records a decision an institution has taken. An institution's permission is one-dimensional even where the system's autonomy is not, and it is the permission that has to be written down, because it is the permission somebody will later be asked to account for.

Whether that reply is sufficient is a fair question and it is left open below.

The five levels

Each level is declared for a **class of action**, not for an agent. One agent will ordinarily hold different levels for different classes on the same afternoon: drafting correspondence at one level, sending it at another, and touching a payment at a third.

Advisory. The agent produces analysis and invokes no external tool. Nothing it does changes anything outside the conversation. Suitable for research, drafting and decision support; the person owns every action that follows.

Draft and approve. The agent prepares an action and a named person approves it before it takes effect. The approval is recorded against the person. This is the level at which approval most often becomes nominal, the failure examined under “Where the human gate fails” below.

Bounded execution. The agent performs low-impact, pre-approved actions within narrow written limits — tagging, filing, internal tickets, read-only retrieval — and escalates anything outside them. Nobody approves each action; somebody approved the boundary.

Supervised autonomy. The agent runs consequential workflows inside a defined perimeter, with continuous monitoring, a policy gateway that can refuse, and the ability to suspend it in seconds rather than after a review.

Delegated autonomy. Long-horizon work, sub-agents, broad effects. The position taken here is the one Mitchell and colleagues argue for and it should be stated plainly rather than hedged: outside a sandbox this level should generally not be used, and where it is, it needs independent assurance rather than internal governance.

What decides the level

Six factors move the permitted level down, and the useful discovery is that five of them are already written into European law — not as an oversight dial, but as the criteria by which the Commission may classify a system as high-risk.

Article 7(2) of the EU AI Act directs attention to, among others: **“(c) the nature and amount of the data processed… in particular whether special categories of personal data are processed”**; **“(d) the extent to which the AI system acts autonomously and the possibility for a human to override a decision”**; **“(g) the extent to which persons who are potentially harmed are dependent on the outcome”**; **“(h) the extent to which there is an imbalance of power, or the persons who are potentially harmed are in a**

vulnerable position in relation to the deployer"; and **"(i) the extent to which the outcome produced involving an AI system is easily corrigible or reversible, taking into account the technical solutions available."**

Irreversibility, override, dependence, power imbalance and data sensitivity — five, and all five are quoted above. The sixth factor used here — whether the action crosses a trust boundary into another party's systems — is not in Article 7 and is added on the evidence of the incidents below.

The distinction worth drawing is that Article 7 uses these to classify *systems* for regulatory purposes, whereas an institution needs them to set the level for *classes of action*. The criteria transfer; the unit of application does not. An organisation adopting them is not implementing the AI Act, and should not tell a regulator that it is.

The precedent from financial markets

Financial markets settled this question in law well before AI, and the settlement is instructive because it did not choose per-action approval.

Commission Delegated Regulation (EU) 2017/589 (RTS 6) requires an investment firm using algorithmic trading to be able to **"cancel immediately, as an emergency measure, any or all of its unexecuted orders"**, and to identify **"which trading algorithm and which trader… is responsible for each order"**. Article 15 requires pre-trade controls: price collars that **"automatically block or cancel orders that do not meet set price parameters"**, maximum order values, maximum volumes, message limits.

SEC Rule 15c3-5, in force since 2010, requires controls **"reasonably designed to prevent the entry of orders that exceed appropriate pre-set credit or capital thresholds, or that appear to be erroneous"**, and — the phrase that matters — those controls must be **"under the direct and exclusive control of the broker or dealer with market access."**

Neither rule asks a human to approve each trade, which would be absurd at the speeds involved. Both bound the system in advance, require an identity per order, and put an emergency stop under the exclusive control of the accountable party. That is bounded execution and supervised autonomy, written into securities law fifteen years ago, and it is the answer to anyone who says the tiers below are impractical because approval does not scale.

What happens without a boundary of that kind is documented. In the first forty-five minutes of trading on 1 August 2012, Knight Capital's router sent **more than four million orders** into the market while attempting to fill 212 customer orders, traded **397 million shares**, and lost

more than $460 million. The SEC's order found the firm **"did not have adequate safeguards in place to limit the risks posed by its access to the markets"**. It was the Commission's first enforcement action under the market access rule.

Where the human gate fails

Draft-and-approve is the level almost every organisation adopts, and it is the level with the best-documented failure.

European data protection law reached the question first. The Article 29 Working Party's guidance on automated decision-making, adopted in 2017 and last revised in February 2018, is unambiguous: **"The controller cannot avoid the Article 22 provisions by fabricating human involvement. For example, if someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing. To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision."**

The Court of Justice took the same view in **SCHUFA** (C-634/21, 7 December 2023), holding that a credit score is itself an automated individual decision where the recipients **"attribute to it a determining role in the granting of credit"** — the human downstream does not launder the decision if the human defers to it.

Two further findings converge on the same point. Feng and colleagues note that disengagement lowers the care with which approvals are given. And one widely cited body of survey research on software delivery reports that formal external approval bodies were associated with worse delivery performance, and finds no evidence that a more formal external review process was associated with fewer failures. ⚠️ It is published by a cloud vendor, it surveys software deployment rather than AI agents, and it is self-reported. The figures are not reproduced here because a single cross-domain survey statistic cannot carry a conclusion about AI approval regimes, and printing it would invite exactly that. What it supports is narrower and still worth having: that an approval step is capable of costing throughput without buying safety, and that whether it does is a measurable question rather than a matter of principle. The conclusion is not that approval is worthless. It is that approval is worth something only where the approver has the authority, the competence and the time to refuse — which is a claim about staffing and about standing rather than about the interface.

The New Zealand position

The consolidated Privacy Act 2020 was searched for the relevant vocabulary. **The word "automated" appears zero times. The word "algorithm" appears once**, inside the information-matching provisions. There is no equivalent to Article 22, no right to human intervention in an automated decision, and no threshold at which oversight becomes mandatory.

That is not an inference from silence. The regulator says so. Writing in December 2025 on the Act's fifth anniversary, Privacy Commissioner Michael Webster stated: **"We also need stronger protections for the significant privacy risks that arise from automated decision-making, which can cause problems such as inaccurate predictions, discrimination, unexplainable decisions, and a lack of accountability."**

An institution in Aotearoa deciding how much its agent may do on its own is therefore deciding without a floor. Whatever it settles on, it settles on by its own authority.

Two questions each organisation answers for itself

Who classifies an action, and when? Feng and colleagues identify the difficulty exactly: an agent may not reliably know which of its actions are consequential. If the agent classifies at run time, the classification inherits every weakness of the system being governed. The position taken here is that classes are declared in advance by a person, in writing — which is slower, cannot cover every case, and is the only version that can be audited afterwards.

What is the organisation prepared to lose by pausing? Every tier below supervised autonomy trades throughput for the ability to answer for what happened. The trade is real, the survey research cited above suggests it is sometimes a bad one, and an organisation that has not priced it will discover its position by accident during an incident rather than by decision beforehand.

How you would know this is wrong

First, if Bradshaw and colleagues are right that any single ordered scale encourages reductive thinking regardless of what it orders, then the distinction drawn here between scaling autonomy and allocating permission is a distinction without a difference, and these five tiers will be misread exactly as they predict. The test is observational: if organisations adopting them start describing themselves as being *at* a tier, rather than declaring tiers per class of action, the instrument has failed in the way its critics said it would.

Second, if institutions operating agents with no declared tiers turn out to gate consequential actions as reliably as those with them, the declaration is ceremony and should be abandoned rather than defended. Nothing found establishes that declared tiers reduce harm; the argument here rests on auditability, which is a weaker claim and is meant to be.

Third, if the financial-markets precedent turns out not to transfer — if pre-trade bounding works for orders because they are homogeneous and typed, and fails for agent actions because they are neither — then the central practical recommendation of this piece is borrowed from a domain that cannot lend it, and the tiers need a different mechanism than a boundary.

Related

Who actually holds the controls establishes that a named body must hold the authority to set these levels, and what would make that answer more than a name on a contract.

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
- The four properties underneath — four properties a delegation must carry, each checkable by observation
- What you can actually require — twelve requirements, and what currently requires each of them
- 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

```
How much it may do unsupervised, version 0.2 (September 2026).
agenticgovernance.digital/downloads/how-much-may-it-do-on-its-own-v0-2.pdf
```

Licensed CC BY 4.0. Reuse is free, including commercially, provided the author is credited and the reuse does not suggest that the author endorses it. This link is frozen at version 0.2; /how-much-may-it-do-on-its-own always shows the current one.

PREVIOUS
Who actually holds the controls

Series contents

NEXT
The four properties underneath

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