

Sample: Oversight readiness review

SAMPLE. This report describes a fictitious organisation and a fictitious system. Every finding below is invented to show what the deliverable looks like. It is not a real engagement.

System	Velmarsh Housing Support Eligibility Check (fictitious)
Organisation	Municipality of Velmarsh, Social Services Department (fictitious)
Review screen	The caseworker review screen for new housing-support applications
Date	Sample — not a real engagement
Reviewer	Matthaios Mantzios, UXellence
Service	Oversight readiness review, 5 working days
Framework	Room to Disagree, uxellence.gr/room-to-disagree (DOI 10.5281/zenodo.22079129)

About the system

The Eligibility Check reads each new application for the municipality's housing-support allowance. It proposes an outcome, "eligible" or "not eligible", with a score. A caseworker reviews the proposal on one screen and records the decision. Approved cases wait in a queue until the nightly letter run, which sends the decision letter and creates the payment instruction.

The municipality treats the tool as a high-risk AI system. Access to essential public services and benefits is one of the areas listed in Annex III of the AI Act. For those areas, Article 14 and Article 26(2) apply from 2 December 2027.

We reviewed one interface, as the readiness review sets out: the review screen for new applications. We worked from screen recordings, a test account, the vendor's user manual, the queue settings and four short conversations with caseworkers and their team lead.

This report describes the interface and the conditions it is used under. It never describes individual people.

SAMPLE · fictitious organisation and system

1. Summary verdict

Can the interface be tested yet? Not yet.

On the new-applications path, an approved case can still be stopped before it takes effect, so a seeded-error audit can run there. But one defect would let reviewers spot the seeds (the known-wrong test cases the audit plants in the queue), and that must be fixed before anything is seeded. Two more defects are serious enough that an audit today would mostly confirm what this report already shows. You would pay to learn something you know now.

The profile across the six groups of the 20-point checklist:

Group	Checks met
Understand	1 of 3
Monitor	1 of 4
Resist	0 of 4
Interpret	2 of 3
Intervene	1 of 4
The review screen itself (base condition)	0 of 2

The checklist returns a shape, not a score, and the shape is the finding. Here it is clear. The screen gives caseworkers some help to make sense of a case (Interpret). It gives them almost no protection against simply agreeing (Resist). And a caseworker who uses a screen reader or a keyboard cannot do the job at all (base condition).

Fix first, then test. The three changes in section 5 that matter most can be made in the screen and the queue settings. None needs a new model.

SAMPLE · fictitious organisation and system

2. Can a decision be stopped in time?

Finding: on the new-applications path, an approved case can be stopped before it takes effect. On the renewals path, it cannot.

A seeded case has to be stopped after the reviewer decides and before the decision takes effect. If approval takes effect at once, the method does not apply.

New applications. When a caseworker approves a case, it moves to "Awaiting notification". It stays there until the letter run at 22:00. A user with the supervisor role can hold a case in that state, so it never reaches the letter run or the payment file. This is a clean stopping point. A seeded case can be withdrawn every working day before 22:00.

Renewals ("fast lane"). For renewals with a high score, "Confirm renewal" sends the payment instruction straight to the payments system. Nothing waits in between. There is no point at which a seeded renewal can be stopped. The seeded-error audit cannot run on this path, and no amount of care makes it possible. If the municipality wants evidence about renewals, the fast lane needs a waiting stage first. That is a process decision for the municipality, not a fix we can recommend inside this review.

Blocker: held cases are visible. A held case shows a grey "Held" badge in the team view, which every caseworker can see. If seeded cases are held, caseworkers will soon learn that "Held" cases never proceed. They would then be reviewing a different queue from normal work, and the result would stop describing real conditions. Holding must leave no visible trace in any view a caseworker uses. This is fix 1 in section 5, and the audit cannot start until it is done.

SAMPLE · fictitious organisation and system

3. The 20 checks

Each check is answered yes or no. Where something is only partly in place, the answer is no and the finding says what is missing.

Understand (Article 14(4)(a))

#	Check	Status	Finding
1	Limitations are shown beside the output, in the operator's language.	No	Limitations appear only in the vendor's user manual, not on the review screen.
2	The intended input scope is stated where outputs appear.	No	The model was built on salaried households, but it scores self-employed applicants with no notice on screen.
3	A current "what this system cannot do" statement exists in the flow.	Yes	A "Known limits" panel opens from the case header and was updated with the last model release.

Monitor (Article 14(4)(a))

#	Check	Status	Finding
4	Uncertainty is displayed as uncertainty, never as a flat verdict.	No	A green "Suggested: Eligible" badge looks the same at a score of 0.51 and at 0.98; the score itself is hidden.
5	Anomalies and drift appear where the operator is already looking, in-flow.	No	Drift reports go to a monthly IT dashboard that caseworkers never see.
6	Base-rate or normal-range context travels with the output.	No	Nothing tells the caseworker how often cases like this one are eligible.
7	The operator can see when the model or its data last changed.	Yes	The screen footer shows the model version and its release date.

Resist (Article 14(4)(b))

#	Check	Status	Finding
8	Friction symmetry: rejecting costs no more clicks, fields or time than accepting.	No	Approve takes one click. Reject opens a form with three required fields, a free-text reason of at least 200 characters and a supervisor's name.
9	No pre-selected approval, auto-advance on accept, or batch-accept for consequential items.	No	The screen jumps to the next case after Approve, and the queue view has "Approve all in view".
10	Queue volume and pacing permit genuine scrutiny per item — the arithmetic is stated.	No	The daily target is 120 cases in about 390 working minutes: 3.25 minutes per case, for files that average 14 pages. Nobody has written this down.
11	No penalty, formal or workflow, attaches to rejection.	No	A weekly "overturn report" lists rejected proposals by caseworker name and goes to team leads.

Interpret (Article 14(4)(c))

#	Check	Status	Finding
---	-------	--------	---------

#	Check	Status	Finding
12	The explanation offers the evidence needed to disagree, never persuasion.	No	The explanation lists "top factors" in words such as "income stability: high", with no link to the figures in the documents.
13	Output is framed as input to a human decision, consistently.	Yes	Labels read "Suggested outcome" and "Record your decision" throughout the flow.
14	Interpretation aids are reachable without leaving the decision.	Yes	The applicant's documents open in a side panel next to the proposal.

Intervene (Article 14(4)(d)–(e))

#	Check	Status	Finding
15	Per-case override and reversal are reachable in-flow.	Yes	A caseworker can override before approval, and reverse an approved case while it waits for the letter run.
16	Declining to use the system for a case is an available path.	No	Every case must carry a score before it can be decided; there is no "assess without the tool" route.
17	A real stop control exists; the operator knows it and its consequences.	No	A "Pause scoring" control exists in the admin console, but none of the caseworkers we spoke to knew about it.
18	Overrides feed back visibly — the operator can see they mattered.	No	Overrides are logged, but nothing is ever shown back to caseworkers.

The review screen itself (the base condition)

#	Check	Status	Finding
19	The review interface meets WCAG 2.2 AA.	No	Approve and Reject are icon-only buttons with no accessible name (4.1.2); focus is not visible in the case list (2.4.7); scanned documents have no text layer (1.1.1).
20	An operator using assistive technology can perform every check above.	No	With keyboard only, the Reject form cannot be opened (2.1.1). With a screen reader, the outcome badge and the documents cannot be read.

The numbers in brackets in checks 19 and 20 are WCAG 2.2 success criteria. This review does not include a full WCAG audit of the screen. These are the failures that stop the oversight checks above from working.

SAMPLE · fictitious organisation and system

4. The five capabilities

Rating scale: **Supported, Partly supported, Not supported**. The base condition is rated too, because all five capabilities stand on it.

Understand — Partly supported.

The "Known limits" panel is a good start, and it sits in the flow where it belongs. But the most important limit is missing from the screen. The model was not built for self-employed applicants, and nothing tells the caseworker when one appears. A limitation stated only in a manual does not count.

Monitor — Not supported.

The caseworker cannot tell a confident output from a correct one. The badge shows a flat verdict and hides the score. Drift is reported to IT, not to the person reviewing the case. The model version in the footer is the only signal the caseworker gets.

Resist — Not supported.

This is the weakest group and the most urgent. Approving costs one click. Rejecting costs a form, a long reason and a supervisor's name, and the rejection then appears on a report with the caseworker's name on it. The queue target leaves 3.25 minutes per case. Each of these pushes the caseworker towards agreeing. Together they price disagreement far above agreement. The friction-symmetry check (8) fails by a wide margin.

Interpret — Partly supported.

The framing is right: the tool suggests and the caseworker decides, and the documents are one click away. The explanation is the gap. "Income stability: high" makes the proposal feel reasonable but gives the caseworker nothing to check. The explanation should point to the figure and the document it came from, so a caseworker can see where the system may have misread a case.

Intervene — Partly supported.

Override and reversal work well inside the waiting window. But there is no way to decide a case without the tool. The stop control exists but is unknown to the people who would need it. And overrides vanish without trace, so caseworkers have no reason to think their disagreement changes anything.

Base condition — Not supported.

A caseworker who uses a keyboard or a screen reader cannot reject a proposal and cannot read the documents. For that person, none of the five capabilities is available. If the interface fails the person assigned to oversight, the oversight obligation fails with it.

SAMPLE · fictitious organisation and system

5. What to fix first

In order. Fix 1 is required before any audit. Fixes 2 and 3 are what we would want in place before you spend money on one.

1. Make holds invisible.

Remove the "Held" badge from every view a caseworker can open. A held case should look exactly like any other case awaiting the letter run. Without this, the audit cannot run. (*Section 2.*)

2. Make rejecting cost no more than approving.

Replace the reject form with one required reason code, chosen in the same number of clicks as Approve. Make the free-text note optional. Remove the supervisor's name. Stop sending the overturn report by caseworker name; if team leads need the data, report it at team level. (*Checks 8 and 11.*)

3. Fix keyboard and screen-reader access on the decision controls.

Give Approve and Reject accessible names, make the Reject path work by keyboard, and show focus in the case list. Add a text layer to scanned documents, or offer the extracted text beside them. (*Checks 19 and 20.*)

4. Remove batch approval and auto-advance.

Take "Approve all in view" off the queue for housing-support cases. After a decision, return the caseworker to the queue instead of opening the next case. (*Check 9.*)

5. Show uncertainty and context.

Show the score band beside the suggested outcome, and add one line on how often similar cases are eligible. Flag self-employed applicants as outside the model's intended scope. (*Checks 2, 4 and 6.*)

We also recommend writing down the queue arithmetic (check 10) and briefing caseworkers on the stop control (check 17). Both cost little and do not need a release.

SAMPLE · fictitious organisation and system

6. What happens next

When fixes 1 to 3 are in, the new-applications path is ready for a **seeded-error audit** (€8,250 + VAT, about 4 weeks). The full fee for this review, €1,650, is credited against it.

What the audit involves:

- **Ground rules first.** Written authorisation and the ethics terms are agreed before anything is seeded. That includes workforce-level notice that seeded-item audits take place, and no disciplinary use.
- **Seeds from real errors.** We build the seeds from the errors the Eligibility Check actually makes, at severities that matter. Trivial seeds produce flattering numbers.
- **Your threshold, in writing, before the first seed.** The municipality records the detection rate below which it will change or withdraw the review step, and justifies that figure against its risk assessment. We do not set the number for you.
- **Two weeks in the live queue.** Seeds go into the real queue, unannounced, at realistic frequency, and are held in "Awaiting notification" before the letter run.
- **A report on the interface, never on people.** Results describe the screen and its conditions. Below roughly thirty seeds, we report raw counts, not percentages.

One limit to keep in mind. A low detection rate is strong evidence that oversight is not happening on this screen. A high one is weaker evidence that it is, because seeds come from failures you already know about.

The renewals fast lane stays out of scope until it has a waiting stage. We will say the same at the scoping call.

Matthaios Mantzios leads every engagement himself, and we take on a limited number at a time.

UXellence · uxellence.gr · matthaios@uxellence.gr

SAMPLE. Fictitious organisation and system. Not a real engagement.