

SAFETY TECH REVIEW

Safety Tech Buyer's Toolkit



RFP questions, pilot acceptance criteria, a public evidence checklist and a vendor scorecard for AI video, wearables and EHS software purchases.

Edition 2026-10-05. Free to use inside your organization. General information, not legal advice.
<https://safetytechreview.com>

Contents: 1. Evidence checklist. 2. Vendor scorecard. 3. 40 RFP questions. 4. Pilot acceptance criteria.

PART 1

Public evidence checklist

Check each vendor's public pages before the first call. Anything missing becomes a document to request.

Method: <https://safetytechreview.com/methodology/evidence-checklist>

Item	Why it matters	A	B	C
Security or trust page	Shows how video, accounts and infrastructure are protected before you ask.			
Data processing agreement published	Worker video is personal data in most jurisdictions; the DPA sets roles, retention and transfers.			
ISO 27001 or SOC 2 claimed	A claim tells you which certificate or report to request. We record the claim, not its validity.			
Privacy features documented	Blurring, anonymization and retention controls decide how workers and regulators react.			
Deployment architecture documented	Edge, on-site or cloud processing changes bandwidth, privacy and IT review.			
Pricing or pricing model public	Lets you budget and compare before procurement.			
Accuracy or validation method published	Without a method, accuracy and outcome claims cannot be compared or tested.			
Independent evidence	Analyst, academic, regulator or insurer evidence that the vendor did not publish itself.			
Named customer case study	A named customer with a measured outcome gives you a reference to check.			
Integrations listed	EHS, video and maintenance integrations decide whether events lead to action.			
Company details disclosed	Legal entity, headquarters and leadership tell you who you are contracting with.			
Third-party reviews (5 or more)	Independent user reviews, read for what they actually review.			

PART 2

Vendor scorecard

Agree the weights before responses arrive. Score each answer 0 (no answer or refused), 1 (claim without evidence), 2 (evidence, not from conditions like yours) or 3 (evidence from conditions like yours).

Weighted score = weight x score.

Area	Weight %	Vendor A	Vendor B	Vendor C
Fit with your hazards and use cases				
Detection performance (precision and recall per detection type, on footage like yours)				
Deployment and resilience (edge or cloud, existing cameras, outages)				
Privacy design and worker consultation				
AI governance and documentation				
Security and data protection documents				
Workflow and integrations (EHS system, alerts, reporting)				
Three-year cost (hardware, installation, licences, support)				
References, including one former customer				
Exit terms and data return				
Total	100			

40 questions to ask AI safety vendors: an RFP checklist

From <https://safetytechreview.com/insights/questions-to-ask-ai-safety-vendors> (updated 2026-10-04).

This checklist details 40 questions to ask vendors of AI safety technology, like video analytics that detect unsafe acts and conditions. The list spans eight areas: fit, detection performance, deployment, data protection, AI governance, security, workflow and commercial terms. Each question highlights what a good answer looks like and the warning signs to watch for. Use it to build a request for proposal (RFP) or a shorter request for information. It grounds your comparison in evidence rather than presentation.

These questions assume a buyer in the US, UK or EU evaluating AI video analytics. Most of them also apply to proximity warning systems, wearables and EHS software with AI features.

Before sending out questions, write down the hazard you want to control. Note where it happens and what evidence you have, like incident and near-miss records. Vendors give better answers when they understand the specific problem. You can then score their responses against it.

Three practical rules make answers easier to compare:

- Ask for documents. Test reports, certificates, sample contracts and data processing agreements resist inflation better than prose.
- Fix the format. Set a word limit per answer and provide a pricing table so responses line up.
- Agree on scoring first. Assign weights to each group before responses arrive, and have at least two people score them independently.

Whatever a vendor claims about outcomes at other sites, like reduced incidents, is still just a claim. Record it as such, and test what you can in a pilot.

These questions check whether the product targets your actual hazard or just a nearby one.

#	Question	A good answer includes	Red flags
1	Which of our listed hazards can your product detect today, and which detections are in beta or on the roadmap?	A clear split between generally available, beta and planned, mapped to your hazards	Roadmap items presented as available
2	Which customers use these detections in operations similar to ours (sector, site size, indoor or outdoor)?	Named or anonymized examples with comparable conditions	Only demo footage or unrelated sectors
3	What does your product not do well, and in which conditions do you advise against it?	Specific limits, such as long distances, heavy occlusion or low light	"No limitations"
4	How are site-specific rules, zones and exceptions configured, and who does it?	A described process, with customer and vendor roles	Every change needs a paid vendor request
5	What changes on our side (layout, procedures, staffing) have customers needed to make to get value?	Honest examples of supervisor time and process changes	"Plug and play" with no workload

Precision measures the share of alerts that turn out to be real events. Recall tracks the share of real events the system actually catches [\[13\]](#). Pushing one up usually pulls the other down. A single accuracy figure can look high even if a system misses rare events [\[13\]](#). Ask for both metrics, broken down by detection type.

#	Question	A good answer includes	Red flags
6	What are precision and recall for each proposed detection, and on what test data (hours, sites, lighting, camera angles)?	Figures per detection with a described dataset and date	One headline "accuracy" number
7	How does performance change at night, in rain, dust, glare, at long range or with partial occlusion?	Results broken down by condition, or a plan to test them in your pilot	No condition-level data
8	Have you tested whether detection rates differ by body size, skin tone, clothing, headwear, PPE color or mobility aids? What did you find?	A described test and its results, including any gaps found	"Our AI has no bias"
9	What alert volume per camera per shift should we expect, and how is it tuned down without losing important events?	A typical range from comparable sites and a tuning method	No figures
10	Will you support a pilot measured on our own footage, with our reviewers labeling alerts and staged tests for recall?	Agreement, plus help with test design	Results reported only from the vendor's dashboard
11	How often do models change, how is each update validated, and can we defer or roll back an update?	Release notes, regression testing and customer control	Silent updates with no notice

Question 8 matters because demographic differences in error rates are well documented in some computer vision tasks. A 2019 NIST study of 189 face recognition algorithms found demographic differentials in the majority of them, including higher false positive rates for some demographic groups [\[12\]](#). Safety detections such as person or PPE detection are different tasks, but the study shows why vendors should test for such gaps instead of assuming they are absent.

#	Question	A good answer includes	Red flags
12	Where does processing run (camera, edge device, on-premises server, cloud), and what data leaves the site, in what form?	A data flow diagram showing video, clips, images and metadata	Vague "secure cloud" answer
13	What are the minimum camera specifications, and which of our existing cameras qualify?	A site survey method and clear specifications	All cameras qualify without a survey
14	What bandwidth, power, rack space and network changes are required?	Figures per camera and per site	Requirements discovered after contract
15	How do you detect camera outages, blocked views, moved cameras or degraded image quality?	Automatic health monitoring and alerts	Customer must notice problems
16	What happens to detection and alerting if the internet connection or cloud service is down?	Local buffering or edge alerting, and stated recovery behavior	Total loss of alerts with no notice

In the UK and EU, a data protection impact assessment (DPIA) is required before processing that is likely to result in a high risk, which workplace video analytics will often be, and the ICO's guidance uses location

tracking of employees as an example of processing that requires one [\[2\]](#). Under the GDPR, biometric data used to uniquely identify a person is a special category with stricter conditions, and vendors acting as processors must sign terms meeting Article 28 [\[1\]](#).

#	Question	A good answer includes	Red flags
17	What personal data does the system process at each stage, and does any feature use biometric identification such as face matching?	A data inventory per feature, with identity features clearly marked	Unclear whether faces are matched
18	Can faces and bodies be blurred or anonymized before data leaves the device, and can identity features be switched off entirely?	Configurable anonymization at the edge and a way to verify it	Anonymization only in the user interface
19	Do you use customer footage to train or improve models? Is it opt-in, how is it de-identified, and can we withdraw?	Off by default, written opt-in, described de-identification, withdrawal process	Training rights buried in terms of service
20	What are the default and configurable retention periods for video, clips, images and event data?	Per data type, configurable by the customer	One fixed retention period
21	Where is data hosted, which sub-processors are used, and will you sign our data processing agreement?	Region options, a current sub-processor list, Article 28 terms [1]	Refusal to share sub-processors
22	What documentation will you provide for our DPIA and worker consultation?	A DPIA support pack and plain-language worker materials	"That is the customer's job"

The ICO's monitoring guidance says employers must make sure workers are aware of what personal information monitoring collects and how, and must be clear about its purpose [\[3\]](#). Vendor materials written for workers, as requested in question 22, help meet that expectation.

The EU AI Act already prohibits AI systems that infer the emotions of people at work, except for medical or safety reasons; that prohibition has applied since 2 February 2025 [\[4\]](#)[\[6\]](#). Where a system is high-risk under the Act, employers deploying it must inform workers' representatives and affected workers before putting it into service and assign competent human oversight [\[5\]](#). Under the amended application dates in Article 113, obligations for high-risk systems listed in Annex III apply from 2 December 2027 [\[6\]](#). Ask these questions now, because contracts signed today will run into that period.

#	Question	A good answer includes	Red flags
23	Have you assessed whether any part of the product is a high-risk AI system under the EU AI Act? What is the conclusion and reasoning?	A written assessment per feature	"The AI Act does not apply to us" with no reasoning
24	Does any feature infer emotions, stress, fatigue or engagement from faces, voice or body language?	A clear yes or no per feature, with the safety rationale for any exception claimed	Fatigue or attention scoring with no legal analysis
25	What instructions for use, intended purpose statement and logging will you provide to support our deployer duties?	Documents available now, plus log export	Promised "when required"
26	Do you follow a recognized AI governance framework, such as the NIST AI Risk Management Framework or ISO/IEC 42001?	Named framework and evidence of use, such as a certificate or internal policy [7] [8]	Framework named with no evidence

#	Question	A good answer includes	Red flags
27	How do customers report errors or bias concerns, and how do you investigate and tell affected customers?	A defined process with response times	No process

#	Question	A good answer includes	Red flags
28	Which security certifications or attestations do you hold, and what is their scope?	Current ISO/IEC 27001 certificate or SOC 2 report, with scope covering the product and hosting [9] [10]	Certification of a data center only
29	How are edge devices hardened, patched and monitored, and how fast are critical vulnerabilities fixed?	Stated patch timelines and secure boot or equivalent	Manual patching on request
30	Who at your company can access our systems or footage remotely, and how is that access approved and logged?	Named roles, customer approval, full logging	Standing remote access for support staff
31	How is the system separated from our operational technology networks, and do you follow ISA/IEC 62443 guidance?	Network segmentation design and familiarity with industrial security standards [11]	Request for flat network access
32	How quickly will you notify us of a security incident affecting our data?	A fixed number of hours in the contract	"Promptly"

SOC 2 reports cover controls relevant to security, availability, processing integrity, confidentiality or privacy, and a given report need not cover all five [\[10\]](#). Always check the report's scope and read any exceptions the auditor flagged.

#	Question	A good answer includes	Red flags
33	How are alerts routed, reviewed, dismissed and escalated, and can we cap alert volume per person?	Configurable routing and caps, with review logged	All alerts to everyone
34	Which video management and EHS systems do you integrate with, and can alerts create actions whose status flows back?	Named integrations and a documented API	Integration "on the roadmap"
35	Can we export all event data and clips in an open format at any time, at no extra cost?	Self-service export in standard formats	Export only at exit, for a fee
36	What onboarding, training and support do you provide, in which languages and time zones?	Named service tiers and response times	Support by email only

#	Question	A good answer includes	Red flags
37	What is the full three-year cost, including licenses, hardware, installation, support and annual price increases?	A complete pricing table with caps on increases	Hardware or support priced later
38	What are the pilot terms, and how do pilot fees convert into a production contract?	Written pilot terms, acceptance criteria and conversion pricing	Free pilot with no terms

#	Question	A good answer includes	Red flags
39	Can you give three references with similar operations, one live for more than a year, and one customer that stopped using the product?	All three, including the former customer	Only recent or hand-picked references
40	What happens to our data and service if we leave, or if you are acquired or stop trading?	Export, certified deletion, transition support and change-of-control terms	No exit provisions

The European Commission's model contractual AI clauses, published in high-risk and non-high-risk versions for public buyers, offer a useful source of contract wording on transparency, human oversight and accuracy [\[14\]](#). They leave out liability and payment terms. So, pair them with your standard commercial agreements.

A simple scheme works well:

Score	Meaning
0	No answer, refusal or irrelevant answer
1	Answer given with no supporting evidence
2	Clear answer with partial evidence
3	Clear answer backed by documents, test results or contract terms

Weight the groups to match your priorities. For a first deployment of video analytics, many buyers give the most weight to detection performance, data protection and three-year cost, and less to breadth of features. Treat some answers as pass or fail regardless of score, for example a refusal to switch off identity features or to put a breach notification period in the contract.

Shortlist two or three vendors and test them in a structured pilot with written acceptance criteria. The RFP tells you what a vendor says it can do. The pilot tells you whether it does so on your site.

A good RFP for AI safety technology asks for evidence in every area: fit with your hazards, precision and recall per detection in conditions like yours, deployment and resilience, data protection and worker privacy, AI governance and EU AI Act readiness, security, workflow and integration, and commercial and exit terms. The 40 questions here, each with what a good answer includes and the red flags to watch for, can be used in full for a formal tender or cut down to the most important for a smaller request. Agree weighted scoring before answers arrive, treat vendor outcome figures as claims, ask for a reference that left, and confirm the shortlist through a pilot measured on your own footage.

Sources

1. Regulation (EU) 2016/679 (General Data Protection Regulation), EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
2. ICO, When do we need to do a DPIA?. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias/when-do-we-need-to-do-a-dpia/>
3. ICO, Monitoring workers: data protection and monitoring workers. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/employment/monitoring-workers/data-protection-and-monitoring-workers/>
4. AI Act, Article 5: Prohibited AI practices. <https://artificialintelligenceact.eu/article/5/>
5. AI Act, Article 26: Obligations of deployers of high-risk AI systems. <https://artificialintelligenceact.eu/article/26/>

6. AI Act, Article 113: Entry into force and application. <https://artificialintelligenceact.eu/article/113/>
7. NIST, AI Risk Management Framework. <https://www.nist.gov/itl/ai-risk-management-framework>
8. ISO/IEC 42001 Information technology, Artificial intelligence, Management system. <https://www.iso.org/standard/42001>
9. ISO/IEC 27001 Information security management systems. <https://www.iso.org/standard/27001>
10. AICPA, SOC 2 examinations. <https://www.aicpa-cima.com/topic/audit-assurance/audit-and-assurance-greater-than-soc-2>
11. ISA, ISA/IEC 62443 series of standards. <https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards>
12. NIST, NIST study evaluates effects of race, age, sex on face recognition software (2019). <https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software>
13. Google for Developers, Classification: accuracy, recall, precision and related metrics. <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall>
14. European Commission Public Buyers Community, EU model contractual AI clauses. <https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/eu-model-contractual-ai-clauses-pilot-procurements-ai>

PART 4

AI safety pilot acceptance criteria: how to define pass and fail

From <https://safetytechreview.com/insights/ai-safety-pilot-acceptance-criteria> (updated 2026-10-04).

Acceptance criteria are the written tests an AI safety pilot must pass before a buyer commits to a wider rollout. Strong criteria are set before go-live. They cover detection quality, operational fit, governance, and cost. Each measure needs a clear counting method, a pass threshold, and an owner. This article shows how to write them for AI video analytics and similar monitoring systems, with a template you can adapt.

This guidance applies to most safety monitoring technology, including proximity warnings and wearables. But the examples focus on video analytics, where accuracy claims are hardest to verify. For the wider buying process, including RFPs and contracts, see the chapter on buying and piloting.

Without agreed criteria, a pilot ends in a meeting where everyone reads the same dashboard differently. The vendor points to the number of unsafe events detected. Operations points to the supervisor hours spent clearing alerts. Finance asks what any of it means for injury costs. Each view is reasonable, and none of them answers whether the system should be scaled.

Written criteria fix three problems. They stop the goalposts moving after results arrive. They force a conversation, before money is spent, about what the system is for and how its alerts will be used. And they give procurement something to put in the contract, so that conversion to a paid production deployment depends on evidence rather than enthusiasm.

ISO 45001, the international standard for occupational health and safety management systems, expects organizations to manage changes that affect health and safety performance and to control the procurement of products and services (clauses 8.1.3 and 8.1.4) [5]. A documented pilot with predefined acceptance tests is one practical way to show that a new monitoring system went through that process. The NIST AI Risk Management Framework makes a similar point from the AI side. Its Measure function asks organizations to test AI systems against defined metrics in conditions similar to those where they will be deployed [6].

Each criterion needs six parts. If one is missing, the criterion will be disputed at the end of the pilot.

Part	What it means	Example
Definition	Exactly what is being counted	"A true alert is one where a reviewer confirms a person was inside the marked forklift exclusion zone while a forklift was within it"
Method	How the data is collected	Weekly random sample of 50 alerts, labeled by two trained reviewers
Threshold	The number that counts as a pass	At least 85% of sampled alerts are true
Sample size	How much data is needed before the result counts	At least 300 labeled alerts over the active phase

Part	What it means	Example
Conditions	Where and when the threshold must hold	Day and night shifts separately
Owner	Who measures and reports	Site EHS advisor, reviewed by the steering group

It helps to split criteria into two buckets. Gates are must-pass conditions. If a gate fails, the pilot fails, or it is extended with a written fix plan. Scored measures, on the other hand, guide the decision on how far and fast to scale. They can also feed into commercial negotiations. Privacy conformance and minimum recall on high-severity hazards are typical gates. User satisfaction and integration effort are common scored measures.

Precision is the share of alerts that are real events. Recall is the share of real events the system catches ^[1]. They pull against each other. Tuning a model to alert more readily tends to raise recall and lower precision, and the reverse also tends to hold ^[1]. A single "accuracy" figure hides this trade-off. Google's machine learning course gives the standard example. When the event of interest is rare, a model that never flags anything can still score 99% accuracy ^[1]. Unsafe events are rare compared with the hours of footage where nothing happens. So ask any vendor quoting accuracy what their figure actually counts.

Set thresholds per detection type, based on what each alert triggers:

Alert use	What matters most	Practical stance
Real-time intervention that stops work or sounds a horn	Precision	False alarms erode trust quickly; set a high precision gate
Real-time warning to a supervisor's phone	Both	Balance; cap alert volume per supervisor
Near-miss capture for investigation	Recall	Missed events defeat the purpose; accept more noise
Trend reporting and heat maps	Consistency	Errors matter less if they are stable over time and across areas

There is no industry standard threshold. A figure you set should come from how alerts will be handled and how much reviewer time the site can afford. Agree interim thresholds for the calibration phase, when zones and sensitivity are still being tuned, and final thresholds for the active phase.

Precision measured on a handful of alerts is mostly noise. As a rough guide, if reviewers label 100 randomly sampled alerts and 90 are true, the 95% confidence interval runs from about 84% to 96%. With 400 labeled alerts at the same rate, that range narrows to about 87% to 93%. This is ordinary sampling arithmetic. It means a pilot target of "90% precision" needs several hundred reviewed alerts per detection type before anyone can say it was actually met.

Sample randomly across shifts, cameras, and weeks. Reviewing only the alerts a supervisor happened to open will skew the data toward busy periods and obvious events. Put two reviewers on part of the sample and compare their labels. If they disagree often, your definition of a true event needs tightening before the numbers mean anything.

Recall is harder to measure. You have to know about the events the system missed. Three methods work in practice:

- **Staged events.** With worker agreement and proper controls, walk a person into a marked zone, park a vehicle in a pedestrian aisle, or remove a hard hat in view. Do this across different cameras, times of day, and weather conditions. Never stage anything that creates real risk.
- **Independent observation.** Trained observers record target events in a set area for a set period, using a checklist. The results are then matched against system detections.
- **Footage review.** Reviewers watch randomly selected footage segments and log every target event. They then compare their log with what the system flagged.

Be careful with small numbers. Hanley and Lippman-Hand showed in 1983 that when no events occur in n trials, the upper 95% confidence limit for the true rate is roughly 3 divided by n [2]. Applied to recall testing, 30 staged events with no misses means the true miss rate could still be as high as about 10%. A recall gate on a high-severity hazard needs dozens of staged events per condition, spread across the cameras that matter.

A system that detects well can still fail in operation. These criteria test whether the site can actually live with the technology.

Criterion	Definition	Typical method
Latency	Time from event to alert reaching the right person	System logs, checked against staged events
Uptime	Share of scheduled hours each camera and the analytics are working	Health monitoring logs
Alert load	Alerts per supervisor per shift needing review	System logs
Review time	Minutes per shift spent reviewing alerts	Time logs or short diary study
Time to action	Time from alert or trend to an assigned corrective action	EHS action log
Action closure	Share of resulting actions closed by their due date	EHS action log
Running cost	License fees plus internal hours, scaled to full deployment	Invoices and time logs

Alert load deserves a firm cap. If supervisors cannot review every high-priority alert during their shift, the system is generating a record of unaddressed hazards, which is a liability after an incident as well as a wasted cost. Time to action and closure tie the technology to the leading indicators OSHA describes as measures that drive change, as opposed to lagging indicators that count injuries after they happen [4].

Most buyers want the pilot to show that unsafe behavior fell. That is a reasonable goal, but it requires careful design.

First, collect a baseline. Running the system in silent mode for several weeks provides a baseline measured in the exact same terms as the active phase. The system records detections but sends no alerts.

Structured manual observations during that same window provide a second baseline that does not rely on the vendor.

Second, use a comparison area if you can. Areas picked because they had a bad recent month tend to improve on their own. This effect is called regression to the mean, and Barnett and colleagues describe randomly allocated comparison groups and multiple baseline measurements as ways to reduce it when designing a study [3]. Without a comparison area, a drop in detected events could reflect the system's impact. But it could also stem from the extra attention a pilot brings, seasonal volume, or plain chance.

Third, agree on the expected direction and approximate size of change before go-live, and explicitly state which indicators count. Examples include pedestrian entries into vehicle zones per hundred vehicle movements, or the share of observed tasks with required equipment worn. Normalize the data by activity level so a quiet month doesn't look like a safety improvement.

Injury rates rarely belong in pilot acceptance criteria. Over a few months in one area, injuries are too infrequent to separate a real change from noise. Track them over the longer term instead.

Governance failures can kill a deployment regardless of system performance. Treat these checks as gates:

- **Privacy conformance.** Settings such as face blurring, retention periods, access permissions and disabled features match the data protection impact assessment (DPIA). In the UK and EU, a DPIA is required before processing likely to result in high risk, and the ICO's guidance uses the tracking of employees as an example of processing that can meet that test [8]. Audit the configuration and the access logs at the end of the pilot.
- **Use policy compliance.** Alerts were used for coaching and engineering fixes, as agreed, and not for discipline outside the agreed policy. The ICO's monitoring guidance expects workers to be told what monitoring is for [9].
- **Worker information.** If the system is a high-risk AI system under the EU AI Act, employers acting as deployers must inform workers' representatives and affected workers before putting it into service or using it at the workplace, and must assign human oversight to people with the necessary competence, training and authority, and the necessary support [7]. Even where the Act does not apply, or a system may only later become high-risk, both are good practice for a pilot.
- **Human oversight in practice.** Named people reviewed alerts on every pilot shift, and their decisions were logged.

Adapt the table below for each detection type in scope. The thresholds shown are just placeholders for illustration. Set your own based on alert use, staffing, and risk.

#	Criterion	Type	Threshold (example only)	Sample or period	Owner
1	Precision, vehicle zone intrusion	Gate	Agreed % of sampled alerts true, day and night separately	300+ labeled alerts	EHS advisor
2	Recall, vehicle zone intrusion	Gate	Agreed % of staged events detected	60+ staged events across cameras and shifts	EHS advisor

#	Criterion	Type	Threshold (example only)	Sample or period	Owner
3	Latency, real-time alerts	Gate	Alert received within agreed seconds	All staged events	IT
4	Uptime	Gate	Agreed % of scheduled hours	Full active phase	IT, vendor
5	Alert load	Gate	Below agreed alerts per supervisor per shift	Full active phase	Operations
6	Time to action, high severity	Scored	Action assigned within one shift	All high-severity alerts	Operations
7	Leading indicator change	Scored	Agreed direction and size versus baseline and comparison area	Baseline plus active phase	EHS lead
8	Privacy and use policy conformance	Gate	Full conformance with DPIA and policy	End-of-pilot audit	Data protection lead
9	Supervisor and worker experience	Scored	No unresolved serious concerns	Survey and interviews	HR, EHS
10	Projected running cost	Scored	Within business case range	Invoices, time logs	Finance

Attach the completed table to the pilot agreement. The European Commission's public buyers community publishes model contractual AI clauses for public organizations procuring AI systems, with versions for high-risk and non-high-risk systems and articles on technical documentation, transparency, human oversight and accuracy. They are not a complete contract and leave out matters such as acceptance, payment and liability ^[10]. Private buyers can still borrow elements from them when drafting the performance and documentation obligations around a pilot.

At the end of the active phase, produce a short evaluation report. Give each criterion one line: result, sample size, pass or fail, and comments. Then apply the rules agreed at the start:

- All gates passed: proceed to a scale-up plan, using scored measures to set the scope and pace.
- A gate failed for a fixable reason, such as a poorly placed camera: extend the pilot for a fixed period with a written fix plan and the same criteria.
- A gate failed for a structural reason, such as recall that cannot be raised without an unmanageable alert volume: stop entirely, or scale only the detection types that passed.

Record the decision and the reasoning behind it. If the system goes into production, keep measuring those same criteria at intervals. Model updates, layout changes, and seasonal conditions can shift performance long after the pilot ends.

Acceptance criteria turn an AI safety pilot from a product demonstration into a test with a clear answer. Write them before go-live, attach them to the contract and give each one a definition, method, threshold, sample size, conditions and owner. Separate must-pass gates from scored measures. Set precision and recall per detection type according to what each alert triggers, and gather enough reviewed alerts and staged events for the results to be meaningful. Add operational criteria for latency, uptime, alert load, time to action and cost, judge leading indicator change against a baseline and a comparison area, and treat

privacy conformance, use policy compliance, worker information and human oversight as gates. At the end, apply the decision rules you agreed at the start and keep measuring after rollout.

Sources

1. Google for Developers, Classification: accuracy, recall, precision and related metrics. <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall>
2. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA, 1983. <https://jamanetwork.com/journals/jama/article-abstract/385438>
3. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology, 2005. <https://academic.oup.com/ije/article/34/1/215/638499>
4. OSHA, Using leading indicators to improve safety and health outcomes. <https://www.osha.gov/leading-indicators>
5. ISO 45001:2018 Occupational health and safety management systems. <https://www.iso.org/standard/63787.html>
6. NIST, AI Risk Management Framework. <https://www.nist.gov/itl/ai-risk-management-framework>
7. AI Act, Article 26: Obligations of deployers of high-risk AI systems. <https://artificialintelligenceact.eu/article/26/>
8. ICO, When do we need to do a DPIA?. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias/when-do-we-need-to-do-a-dpia/>
9. ICO, Employment practices and data protection: monitoring workers. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/employment/monitoring-workers/>
10. European Commission Public Buyers Community, EU model contractual AI clauses. <https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/eu-model-contractual-ai-clauses-pilot-procurements-ai>