INFORMATION SECURITY MANAGEMENT SYSTEM

AI Model Governance Policy

How IrexAI develops, sources, evaluates, deploys, and monitors the AI models inside the IREX Ethical AI Platform: the in-house computer-vision detectors, and the open-weight foundation models behind StreamVLM™ and Ask IREX.

DOCUMENT ID DOC-ISMS-POL-014	VERSION 1.0
EFFECTIVE DATE September 7, 2026	NEXT REVIEW September 7, 2027
DOCUMENT OWNER President, Chief Technology Officer & Co-Founder	APPROVAL AUTHORITY Chief Executive Officer and Chief Technology Officer
STANDARDS ISO/IEC 42001:2023 (aligned) · ISO/IEC 27001:2022 (aligned) · NIST AI RMF 1.0	CLASSIFICATION Public: approved for customer, partner, and tender distribution

Contents

Who This Policy Is For

This is the public statement of how IrexAI Inc. governs the artificial-intelligence models inside the IREX Ethical AI Platform. It is written for the people who have to trust those models before they deploy them: procurement officers, agency privacy and compliance teams, contracting authorities, auditors, and the oversight bodies that review how an agency uses AI. It is also binding on IrexAI's own engineers. It may be supplied to any customer, prospective customer, partner, or contracting authority, and it is the document to send when a tender asks how IrexAI trains, tests, and controls its AI.

The IREX Ethical AI framework, built on the doctrine of **Transparency by Design** and its six pillars, describes what the platform does for an operator. This Policy describes what IrexAI does before a model ever reaches an operator, and what IrexAI keeps doing after it does.

1. Purpose

IrexAI is an AI company. The models are the product. A video-analytics platform for public safety is only as trustworthy as the data its detectors learned from, the tests they passed, the integrity of the files that reach a customer's servers, and the limits placed on what a general-purpose model can do once it is inside the platform. This Policy sets mandatory requirements for every stage of that lifecycle.

The Policy:

- defines the two classes of model that ship in the platform, and governs each with rules fitted to how it is built: **in-house detector models** that IrexAI trains (Part A), and **open-weight foundation models** that IrexAI selects, pins, and confines (Part B);
- restricts training data to four permitted sources, each with a documented permission to use it for model development, and prohibits every other source;
- requires a human to write or review every label a model learns from, including labels first proposed by a machine;
- governs synthetic data produced by IrexAI's 3D artists and generative models, and requires that models are always evaluated on real imagery;
- requires a recorded evaluation, and a demographic bias assessment for any model that detects or identifies people, before release, and disclosure of known limitations to customers;
- protects the integrity of model artifacts from training run to deployment, against poisoning, tampering, and theft;
- confines foundation models to designated endpoints inside the platform, with no tools beyond their designed function and no access to data beyond the permissions of the user they serve;
- ties every model output to the platform's permission model, mandatory Case ID, and non-erasable Logbook, so that a model's action is as accountable as an operator's.

This Policy is a sub-policy of the Information Security Policy (DOC-ISMS-GOV-001) and inherits its governance, enforcement, and review framework. It completes the AI governance set alongside the AI and External Data Provider Policy (DOC-ISMS-POL-013), which governs IrexAI's internal use of third-party AI services and expressly leaves product-model governance to this document.

2. Scope

2.1 What Is in Scope

This Policy applies to every AI model that IrexAI develops, fine-tunes, selects, or integrates and delivers as part of the IREX Ethical AI Platform, in both delivery models: IrexAI-managed instances on the dBrain private cloud, and customer-hosted on-premises instances. It covers the full lifecycle of those models (purpose definition, data sourcing, labeling, synthetic data generation, training, evaluation, packaging, deployment, monitoring, update, and retirement) and every person who takes part in it: IrexAI employees, contracted annotators, 3D artists, and contractors.

Class	What It Is	Examples in the Platform	Governed By
Class A: In-house detector models	Specialized computer-vision models that IrexAI trains and owns. Typically a published, well-understood architecture, fine-tuned on IrexAI datasets and optimized for real-world CCTV conditions and for CPU or GPU inference.	Face recognition, license plate recognition, weapon detection, vehicle and traffic analytics, crowd analytics, fire and smoke, camera tamper and quality control, person re-identification.	Part A (§7 to §12) plus the common sections
Class B: Foundation models	Open-weight vision-language and language models published by third parties, used as published without modification, pinned by version and checksum, and served only inside the platform.	StreamVLM™ prompt-defined detectors; Ask IREX natural-language queries and investigation agents. Both are shipping in beta on selected instances.	Part B (§13 to §17) plus the common sections

2.2 Exclusions

- **IrexAI's internal use of third-party AI services** (assistants, coding tools, agentic tools for staff): governed by DOC-ISMS-POL-013.
- **Sover's own language-model integrations** inside the Sover communications platform. A request that Sover forwards to the IREX platform through Ask IREX is in scope, because the IREX platform executes it.
- **Models developed by customers or third parties** and integrated by a customer alongside the platform.
- **The operator's decisions as data controller**: lawful basis for processing, retention configuration, Case ID discipline, and local data-protection law. These rest with the operating agency, as set out in the Shared-Responsibility Matrix (DOC-ISMS-GOV-009).

2.3 Delivery Models and Shared Responsibility

On IrexAI-managed instances, IrexAI operates the models and the infrastructure they run on. On customer-hosted instances, IrexAI supplies the models, the inference services, and their configuration as licensed software, and the customer operates the environment and owns the data. In both models IrexAI neither owns nor has access to customer video, events, or logs, and this Policy does not change that. Where a customer chooses to help improve a model, it does so by providing a Data Set under its agreement (§8.1), never by granting access to a production instance.

2.4 Proportionality Statement

IrexAI is a technology company of approximately 100 people with a machine-learning team of about 30. Controls in this Policy are implemented in proportion to risk. Where IrexAI departs from a practice recommended by ISO/IEC 42001:2023 or the NIST AI Risk Management Framework, the departure and its compensating control are recorded in the Statement of Applicability (DOC-ISMS-GOV-007). IrexAI does not hold an ISO/IEC 42001 or ISO/IEC 27001 certificate; alignment is stated as such and never as certification.

3. Definitions

Term	Definition
AI Model	A trained set of parameters (weights) plus the code that runs it, which turns an input (a video frame, an image, a text request) into an output (a detection, a match, a structured answer).
Detector Model	A Class A model: an in-house computer-vision model trained by IrexAI for a defined detection, recognition, or classification task.
Foundation Model	A Class B model: a large vision-language or language model trained by a third party for general use and published with downloadable weights ("open-weight"). IrexAI uses such models as published.
Fine-Tuning	Further training of an existing model's weights on IrexAI data. Fine-tuning turns a model into an IrexAI-owned artifact governed by Part A.
Dataset	A defined, versioned collection of images, video frames, or clips with their labels, held for training, validation, or testing.
Held-Out Test Set	Real imagery sealed before training and used only to evaluate a model. Never used for training or tuning.
Synthetic Data	Imagery produced by rendering 3D scenes, by generative image models, or by compositing reference objects onto real backgrounds, rather than captured by a camera.
Label	The annotation a model learns from: a bounding box, a class, an attribute, a caption, or a verdict.
Pre-Labeling	A label proposed by a model rather than written by a person. A pre-label becomes a label only after human review.
Data Set (capitalized)	The term of art in the IREX Distribution and Licensing Agreement, clause 3.4: non-individually-identifiable image data that a licensee or its end user provides to IrexAI for testing, training, and improving the neural networks of the Product.
Development Camera	A camera operated by IrexAI or an IrexAI partner exclusively to collect imagery for model development. Never installed in a public space and never part of a customer deployment.
Model Register	The internal, controlled inventory of every model in development, production, or retirement, with the fields listed in §6.1 (DOC-ISMS-REG-005).
Model Card	The standing description of one model: intended use, data, evaluation results, limitations, and version history (§6.2).
Datasheet	The standing description of one dataset: source, permission, composition, labeling, quality, and retention (§6.3).
Designated Endpoint	An inference service inside the platform's cluster that serves a foundation model to other platform services only.
Tool	In the language-model sense: a function a model may ask the surrounding software to execute, such as a search, a file read, a network call, or code execution.

TABLE CONTINUED

Term	Definition
Case ID	The platform's mandatory reference to the lawful grounds for a privacy-sensitive action (a case file, a court order, a missing-person report), recorded in the Logbook with every such action.
Operator	The customer organization running an instance, and the individual users acting within it under its permissions.
Trust Boundary	IrexAI-controlled infrastructure (the dBrain private cloud and IrexAI-managed development systems) and, for a customer-hosted instance, that customer's environment. Data inside a trust boundary does not leave it for inference or labeling.
Model Incident	An event in §18.3: suspected poisoning, tampering, a vulnerability in a model or its serving stack, an unintended capability, systematic bias, a field accuracy regression, or a dataset permission breach.

4. Roles and Responsibilities

Responsibility is assigned by role so that the Policy survives personnel change. The current holders are named in the approval block and, for operational roles, in IrexAI's due-diligence materials on request.

4.1 Chief Executive Officer

Approves this Policy jointly with the Chief Technology Officer; is accountable to customers and contracting authorities for the commitments it makes; receives the annual model-governance report and any Critical model incident.

4.2 Chief Technology Officer (Document Owner)

Owns this Policy. Approves the addition of any new model class, any new foundation model, and any new dataset source category. Approves the Model Register at each quarterly review. Decides release where an evaluation threshold has not been met and records the reason. Authorizes exceptions under §21. Approves the fixed function set available to Ask IREX (§15.2) and any change to it.

4.3 Head of Machine Learning

Operational owner. Maintains the Model Register, model cards, and datasheets. Owns dataset intake, labeling guidelines and labeling quality assurance, the synthetic-data pipelines, the evaluation protocol, and demographic bias testing. Signs the evaluation record for every release. Runs the foundation-model benchmark before any model change under Part B.

4.4 Chief Information Security Officer

Owns model integrity and supply-chain controls (§12), the vulnerability watch for foundation models and the inference stack (§18.2), and the response to model incidents (§18.3). Keeps this Policy aligned with the Statement of Applicability and the Risk Register, in particular the risks of model poisoning and model theft.

4.5 Data Protection Function

The owner of the Data Protection Policy (DOC-ISMS-POL-008) confirms, before intake, that each dataset has a documented permission for model development and a lawful basis where it contains personal data; requires a Data Protection Impact Assessment for any training set containing biometric data; and handles data-subject requests that concern training data.

4.6 Platform Engineering

Implements and operates the designated endpoints, the access-control enforcement that foundation models inherit (§16), network isolation (§14), and request logging. Takes model releases through the Change Management Procedure (DOC-ISMS-PR-004) and the Secure Deployment Procedure (DOC-ISMS-PR-005).

4.7 ML Engineers, Annotators, 3D Artists, and Contractors

Work under a non-disclosure agreement and this Policy. Use only IrexAI's labeling platform, dataset storage, and training infrastructure inside the trust boundary. Never copy imagery to personal devices or to unapproved services. Report any imagery of doubtful provenance, any suspected data contamination, and any anomaly in model behavior to the Head of Machine Learning and the CISO.

4.8 Customers and Operators

Remain data controllers for their instances. Decide whether to provide Data Sets to IrexAI and, if so, under which agreement. Enter a Case ID before privacy-sensitive actions, including those executed through Ask IREX. Verify alerts before acting on them. Report suspected model errors, including false positives with a demographic pattern, through the support process so that they enter the evaluation loop.

5. Principles

Ten principles govern every decision under this Policy. Each carries the pillar of the IREX Ethical AI framework it serves and the NIST AI RMF function it belongs to.

1. **Permissioned data only.** No image enters a training, validation, or test set unless IrexAI holds explicit, documented permission to use it for model development. There is no fifth source. *(Cybersecurity and Privacy Protection · Govern, Map)*
2. **Humans decide what a model learns.** Every label used in training is written by a person or reviewed by a person. A machine-proposed label that no one has checked does not enter training. *(Bias Awareness and Mitigation · Manage)*
3. **Every model has a narrow, stated purpose.** Each model is built for a declared task with a declared list of what it is not for, and the platform's architecture, not just this Policy, keeps operators from turning it to another purpose. *(Narrow Constraints for Law Enforcement · Map)*
4. **Models run where the data lives.** Inference happens inside the trust boundary of the instance. No customer frame, query, or event is sent to a third-party AI service. *(Cybersecurity and Privacy Protection · Manage)*
5. **Least capability for foundation models.** A foundation model receives exactly the functions its design requires and nothing else: no general tools, no memory across requests, no ability to extend itself. *(Narrow Constraints · Manage)*
6. **Permission inheritance.** A model never sees more than the user it serves. Access control is enforced by the platform on every request; the model holds no credentials of its own. *(Permission-Driven User Interface · Manage)*
7. **Evaluate before release, monitor after.** No model reaches production without a recorded evaluation on a sealed real-imagery test set, and a demographic bias assessment where the task involves people. Production behavior is monitored and fed back. *(Bias Awareness and Mitigation · Measure)*
8. **Known limitations are disclosed.** Accuracy is stated with its basis: the dataset, the benchmark, the date. Where systematic bias exists, IrexAI discloses its existence and estimated impact to the customer. No guaranteed-detection claim is made. *(Full Transparency · Measure)*
9. **Provenance and integrity are recorded.** Every model artifact has a lineage: datasets, code version, training run, evaluation, approver, checksum. Artifacts are signed, and the signature is verified before deployment. *(Cybersecurity and Privacy Protection · Manage)*
10. **Every model output is attributable.** Detections, matches, and answers are logged against the operator, the camera, and the Case ID in the non-erasable Logbook, and exported daily in signed form. A model informs a person's decision; it does not take the decision. *(Full Transparency · Govern)*

6. The Model Register and Model Documentation

6.1 Model Register

The Head of Machine Learning maintains the Model Register (DOC-ISMS-REG-005), an internal controlled record of every model from first training run to retirement. It carries, for each model: identifier and class; declared purpose and out-of-scope uses; architecture family and origin (in-house, fine-tuned from a published base, or open-weight as published); version and checksum; license of any third-party base or foundation model; the datasets used, by identifier; the training-run reference; the evaluation record and, where applicable, the bias assessment; the approver and approval date; deployment status (development, staging, production, retired) and the instances it runs on; known limitations; the vulnerability-watch source; and the retirement date. The CTO reviews the Register quarterly. The Register is internal: it names exact model versions and datasets, which are RESTRICTED under the Data Protection Policy §5.4.

6.2 Model Cards

Every production model has a model card, maintained with the model and following the structure of *Model Cards for Model Reporting* (Mitchell et al., 2019): intended use and users; out-of-scope uses; a summary of training data by source category (§8.1); evaluation data and metrics; performance by condition (lighting, distance, weather, camera angle); demographic evaluation where the model detects or identifies people; known limitations and failure modes; and version history. A customer-facing model card summary is available to any customer on request. The internal card additionally carries architecture details, tuning parameters, and internal benchmark figures, which are RESTRICTED and are released only under a non-disclosure agreement where a customer's security review requires them.

6.3 Datasheets for Datasets

Every dataset has a datasheet, created before the first image is used and following *Datasheets for Datasets* (Gebru et al., 2021): the source category (one of the four in §8.1) and the permission instrument (license, agreement clause, or camera-ownership record); collection dates and the type of location; composition (classes, counts, conditions, the proportion of synthetic imagery); the labeling method and quality-assurance result; whether it contains personal or biometric data and the lawful basis if so; known gaps; and the retention and deletion terms.

6.4 Records Retention

Register entries, model cards, datasheets, evaluation records, and training-run records are retained for the life of the model in production plus five years, so that a detection made years earlier can be traced to the model version and data behind it.

PART A

In-House Detector Models

The specialized computer-vision models IrexAI trains and owns: how their purpose is set, where their data comes from, who labels it, how synthetic imagery is used, how they are evaluated for accuracy and bias, and how their integrity is protected from training run to deployment.

7. Model Development Lifecycle

Every Class A model passes through the stages below in order. Each stage has a gate, a record, and an owner. A model that has not passed a gate does not proceed.

#	Stage	Gate	Record	Owner
1	**Purpose and risk assessment.** Define the task, the users, the out-of-scope uses, and the harm if the model is wrong. Assign a tier (§7.1). Decide whether the model will process personal or biometric data.	CTO approves the purpose statement; Data Protection function confirms whether a DPIA is required.	Model Register entry (development); purpose statement	Head of ML
2	**Architecture selection.** Choose a published, well-understood architecture suited to the task and to CPU or GPU inference. Any pretrained base is taken from its publisher's official channel with license reviewed and checksum recorded.	License permits commercial use; checksum verified; safe serialization (§12.1).	Register: architecture family, base, license, checksum	Head of ML
3	**Data assembly.** Assemble training, validation, and held-out test sets from permitted sources only (§8). Seal the test set.	Every dataset has a datasheet and a permission instrument on file.	Datasheets; dataset hash inventory	Head of ML; Data Protection
4	**Labeling.** Label under a written guideline, with human review of every machine-proposed label and sampled quality assurance (§9).	Batch passes its quality-assurance threshold.	Labeling QA record; label provenance	Head of ML
5	**Training.** Train on IrexAI-controlled infrastructure from version-controlled code, recording the code commit, dataset versions, hyperparameters, and random seed so that the run can be reproduced.	Run record complete.	Training-run record	ML Engineering
6	**Evaluation.** Evaluate on the sealed real-imagery test set, by condition, and against the regression tolerance; run the demographic bias assessment for people-related models (§11).	Thresholds met, or a CTO-recorded release decision.	Evaluation record; model card update	Head of ML; CTO
7	**Release.** Package, hash, and sign the artifact; take it through change management and secure deployment (§12); publish the change in release notes.	Signature verified at deployment; Change Advisory Board approval.	Change record; signed artifact; release notes	Platform Engineering; CISO
8	**Monitoring and improvement.** Track field performance and customer-reported errors; generate targeted data for blind spots; retrain and re-evaluate (§11.6, §18.1).	Quarterly review of production models.	Monitoring record; Register update	Head of ML
9	**Retirement.** Withdraw the model from registries, update the Register, and notify customer-hosted instances through release notes.	CTO approval.	Register: retired, date	Head of ML

7.1 Model Tiers

The rigor of stages 1, 6, and 8 scales with the tier assigned at stage 1.

Tier	Models	Additional Requirements
Tier 1: person-identifying	Face recognition, person re-identification, person attribute models	DPIA before training on biometric data; demographic bias assessment mandatory before every release; robustness testing against spoofing and presentation attacks; results disclosed to customers per §11.3.
Tier 2: safety-critical detection	Weapon detection, fire and smoke, falls and track intrusion, dangerous crowding	Condition-stratified evaluation (distance, lighting, occlusion); false-positive review of field reports each quarter; targeted synthetic data for rare cases (§10).
Tier 3: object, traffic, and camera quality	Vehicle and license plate recognition, traffic violations, two-wheelers, camera tamper and quality	Standard evaluation protocol; regional adaptation (for example, a new plate format) goes through the full lifecycle as a new model version.

8. Dataset Governance

8.1 Four Permitted Sources

IrexAI trains its models only on imagery that was provided to IrexAI with explicit permission for that use. Exactly four sources qualify. The permission instrument for each is named in the datasheet and kept on file.

Source	What Qualifies	Permission Instrument	Conditions
1. Public open-source datasets	Research and community datasets published for reuse.	The dataset's license, reviewed and recorded before intake.	The license must permit commercial use and model training. Datasets withdrawn or disputed by their authors, or assembled by scraping identifiable people without a documented basis, are excluded. IrexAI does not itself scrape images of people from the web or social media.
2. Synthetic datasets	Imagery rendered by IrexAI's 3D artists, produced by generative image models, or composited from reference objects onto IrexAI-owned backgrounds.	IrexAI authorship; licenses of the 3D assets and generative models used, reviewed and recorded.	Governed by §10. Tagged as synthetic in the dataset and the datasheet.
3. Customer Data Sets under a signed agreement	Non-individually-identifiable image data that a licensee or end user provides to IrexAI to fix or improve performance in its deployment.	Clause 3.4 (Neural Network Training) of the IREX Distribution and Licensing Agreement, or an equivalent written clause in the customer's agreement.	Used only for testing, training, and improving the neural networks of the Product. Never disclosed, published, shared with another customer, or used for any other purpose. Provided by the customer; IrexAI never extracts it from a production instance.
4. Development cameras	Imagery from cameras operated by IrexAI or IrexAI partners exclusively for model development.	Ownership or partner agreement for the camera and the site, recorded in the datasheet.	Never installed in a public space. Never part of a customer deployment. The notice and consent arrangements in force at the site are recorded in the datasheet.

8.2 Prohibited Sources

The following never enter a dataset, and no exception under §21 can admit them:

- video, frames, events, watchlist images, or exports from any customer production instance, managed or on-premises. IrexAI has no access to this data and does not seek it;
- feeds from cameras installed in public spaces, including cameras in customer deployments;
- images of people scraped from the web, social media, or news media;
- datasets whose license forbids commercial use or model training, or whose authors have withdrawn them;
- imagery bought from data brokers without a documented and lawful basis for the people it shows;
- any imagery whose provenance cannot be established.

8.3 Intake, Storage, and Access

- A datasheet is created and the permission instrument filed before any image is used for training, validation, or testing.
- Datasets containing faces or otherwise identifiable people are classified RESTRICTED under the Data Protection Policy and stored inside the trust boundary with access limited to named personnel under the Access Control Policy (DOC-ISMS-POL-001). Access is logged.
- Every dataset version carries a content hash inventory, so that additions, removals, and substitutions are detectable (§12.3).
- Training, validation, and test splits are recorded. Held-out test sets are sealed before training and are never used for training or hyperparameter tuning.

8.4 Minimization and De-Identification

- Where the task does not need identity, faces and plates are blurred or cropped before the imagery enters training. Only Tier 1 models train on face-identity data.
- No dataset links an image to a name, a case, or a watchlist. Customer watchlist images are never used for training.
- Demographic attribute labels collected for bias evaluation are held separately from training labels and used for evaluation (§11.3).

8.5 Retention and Deletion

A dataset is retained while a model trained on it is in production, plus the records-retention period in §6.4, unless its permission instrument sets a shorter term. When a permission instrument expires or is validly withdrawn, the affected imagery is removed from the dataset, the datasheet is updated, and the next training run excludes it.

9. Labeling and Annotation

9.1 Who Labels, and Where

Labels are written by IrexAI staff and by contracted annotators bound by non-disclosure agreements and trained on the task's labeling guideline. All labeling takes place in IrexAI's labeling platform, operated inside the trust boundary. Imagery is not exported to annotators' personal devices, to consumer file-sharing services, or to any labeling service outside IrexAI's control.

9.2 Labeling Guidelines

Each task has a written, version-controlled labeling guideline before labeling starts: class definitions, box conventions, edge cases, ambiguity rules, and what not to label. Changes to a guideline mid-project are recorded, and batches labeled under the old rule are re-checked.

9.3 Machine Pre-Labeling

Large vision-language models may propose labels (boxes, classes, attributes, captions) to speed up annotation. Three rules apply:

- Pre-labeling models run on IrexAI-controlled infrastructure. Submission of dataset imagery to any external AI provider is governed by DOC-ISMS-POL-013 and is prohibited for customer Data Sets, for imagery of identifiable people, and for biometric imagery.
- Every machine-proposed label is reviewed by a person before it enters a training set. The reviewer accepts, corrects, or rejects it. Acceptance and correction rates are recorded per batch, and a batch with a low acceptance rate is re-labeled from scratch.
- Label provenance (human-written, or machine-proposed and human-verified) is stored with each label.

9.4 Quality Assurance

- A sample of every batch is independently re-labeled by a second annotator; inter-annotator agreement is measured against a threshold set in the guideline.
- Batches below threshold are re-labeled. Systematic disagreements trigger a guideline revision.
- Label-distribution checks flag batches whose class balance or box statistics differ unexpectedly from the rest of the dataset, as a poisoning control (§12.3).

9.5 Sensitive Attributes

Demographic attributes (gender, age group, skin tone) are labeled where they are needed to evaluate a Tier 1 model's performance across groups, under a documented protocol, and are held separately from training labels. Where an attribute is itself a product function, that function is governed by the platform's permission model and Case ID requirement, not by this section.

10. Synthetic Data

10.1 Why IrexAI Uses Synthetic Data

The events a public-safety platform must detect are rare, and real footage of them is scarce, sensitive, and often unusable. Synthetic imagery lets IrexAI train detectors for falls, track intrusions, and firearms without collecting footage of real people in danger, cover conditions (distance, lighting, occlusion, weather) that real data under-represents, and close a blind spot found in evaluation within days instead of waiting for rare footage. It also reduces the amount of real imagery of people that IrexAI needs to hold.

10.2 Methods

- **3D rendering.** IrexAI's 3D artists build scenes (stations, platforms, streets, sites) and render randomized variations of the target event with varied actors, poses, cameras, and lighting.
- **Generative models.** Generative image models produce or vary backgrounds, textures, weather, and object appearance.
- **Reference compositing.** Shape-accurate reference objects (for example, a firearm rendered from standardized reference images) are composited onto IrexAI-owned CCTV backgrounds with randomized scale, angle, finish, and occlusion, so that the model learns the object in real noise, blur, and compression.

10.3 Rules

- Every synthetic sample is tagged as synthetic in the dataset and the datasheet, and the proportion of synthetic to real imagery is recorded for every training set.
- Generated samples pass human quality control before training: reviewers check shape, scale relative to people and hands, lighting and shadow integration, and compositing artifacts. The rejection rate is recorded. A high rejection rate is accepted as the cost of a clean set.
- Synthetic data is never used to construct the likeness or identity of a real person, and never serves as the sole training basis for a Tier 1 face-identity model.
- Generative models used for synthesis run inside the trust boundary, or are approved providers under DOC-ISMS-POL-013 receiving no customer-derived or biometric input.
- Licenses for 3D assets, rendering engines, and generative models are reviewed and recorded like any other third-party component.
- **Evaluation is on real imagery.** Held-out test sets used for release decisions contain real imagery only. A synthetic-only test set may be used for development diagnostics and is labeled as such.

10.4 Improvement Loop

Models are pretrained on curated synthetic sets and fine-tuned on real annotated imagery from permitted sources. Evaluation surfaces blind spots (a small object at distance, low light, an object confused with a weapon); IrexAI generates targeted synthetic examples for those cases, retrains, and re-evaluates on real imagery. The loop repeats until the release threshold is met.

11. Evaluation, Accuracy, and Bias Testing

11.1 Evaluation Protocol

Every release is evaluated on a sealed, real-imagery held-out test set with metrics fitted to the task: precision, recall, and false-positive rate per class for detectors; mean average precision for multi-class detection; verification and identification accuracy at stated false-match rates for face recognition; character and plate accuracy for license plates. Results are stratified by condition (lighting, weather, distance and object size in pixels, occlusion, camera angle) so that the model card can say where the model works and where it degrades.

11.2 Release Thresholds and Regression Gate

Thresholds are set per model before evaluation, not after. A new version must not degrade on the sealed test set beyond a defined tolerance relative to the version it replaces. Where a threshold is missed, only the CTO may approve release, with the reason recorded in the evaluation record and the limitation carried into the model card and release notes.

11.3 Demographic Bias Assessment

All face recognition systems, including IrexAI's, show some variation in accuracy across demographic groups. IrexAI's rules follow from the Bias Awareness and Mitigation pillar:

- Every Tier 1 model is evaluated across gender, age group, and skin tone or ethnic group before release, on labeled evaluation sets that include public benchmarks built for this purpose and IrexAI's own sets. The spread between the best- and worst-performing groups is recorded.
- Mitigation techniques used in training (dataset balancing across groups, color and geometric augmentation, combining datasets of different origin) are recorded in the model card.
- Where systematic bias exists, IrexAI discloses its existence and estimated impact to customers. Results of the most recent assessment are available to any customer on request, and the last published study is summarized in Appendix C.

11.4 Independent Evaluation

IrexAI submits models to independent benchmarks where the benchmark represents the conditions the model is built for, and states results with the benchmark, dataset, and date. IrexAI's face recognition engine was evaluated in the NIST Face Recognition Technology Evaluation (FRTE, then named FRVT); the result and IrexAI's reasons for not entering later rounds are stated in Appendix C. IrexAI does not imply an evaluation it has not undergone, and does not claim guaranteed detection.

11.5 Robustness Testing

Tier 1 and Tier 2 models are tested for robustness before release: face recognition against presentation attacks (printed photos, screens, masks) through the platform's liveness checks; detectors against compression, resolution loss, and partial occlusion; and, where a realistic threat exists, against adversarial patterns. Results are recorded in the evaluation record.

11.6 Field Validation and Feedback

Pilots and production feedback are part of evaluation. Customer-reported false positives and misses are triaged by the ML team, and Tier 2 false-positive reports are reviewed each quarter. A pattern of errors becomes a targeted data request (§10.4) or a Data Set request to the customer under its agreement (§8.1). A customer report that suggests a demographic pattern is escalated to a bias assessment.

11.7 Stating Accuracy

Every accuracy figure IrexAI publishes carries its basis: the dataset or benchmark, the metric, and the date. Figures are not rounded up, and false-positive rates are named alongside detection rates. The platform's alerts are presented to operators as detections to verify, never as findings.

12. Model Integrity and Supply Chain

12.1 Third-Party Bases and Components

- Pretrained bases, open-source architectures, and training libraries are obtained from their publishers' official channels. The license and the checksum are recorded in the Model Register before use.
- Model files are stored in safe serialization formats where the ecosystem provides them, and every downloaded artifact is scanned for embedded executable content before it is loaded.
- Training and inference dependencies and container images are scanned under the Vulnerability Management Procedure (DOC-ISMS-PR-003), with remediation to its severity timelines.

12.2 Training Pipeline Integrity

- Training code and configuration live in version control under the Secure Development Policy (DOC-ISMS-POL-005): every change to a pipeline or a model is reviewed by at least two approvers before merge.
- Training runs execute on IrexAI-controlled infrastructure from a recorded commit, dataset versions, and seed, so that a production model can be reproduced and audited.

12.3 Poisoning Controls

Data and model poisoning (inserting samples so that a model misses a target or flags a benign object) is a recorded risk in IrexAI's Risk Register. Controls: the permitted-source rule and datasheets (§8.1) establish provenance; content-hash inventories (§8.3) detect substitution; label-distribution checks (§9.4) and second-annotator sampling detect anomalous batches; sealed test sets that the training team does not curate detect unexpected behavior; access to dataset stores is logged; and the two-approver rule covers every pipeline change.

12.4 Artifact Signing and Deployment

Every model artifact released to production is hashed and signed. Artifacts are held in IrexAI's registry under access control, and the signature is verified before deployment under the Secure Deployment Procedure (DOC-ISMS-PR-005). A signature mismatch blocks deployment and is a model incident (§18.3). Release goes through the Change Management Procedure (DOC-ISMS-PR-004), with a rollback path to the previous signed version.

12.5 Model Confidentiality

Model weights, architectures, tuning parameters, and internal benchmarks are RESTRICTED (Data Protection Policy §5.4). Registry access is limited and logged, and network egress from clusters is constrained. On customer-hosted instances, model artifacts are delivered as part of the licensed software and remain IrexAI intellectual property; the customer's agreement prohibits decompiling, reverse engineering, and extraction. Model theft is a recorded risk with its own treatment in the Risk Register.

PART B

Foundation Models in StreamVLM™ and Ask IREX

The open-weight vision-language and language models that IrexAI selects, pins, and confines: how they are chosen, where they may run, what they may and may not do, and how they inherit the platform's permissions and Case ID.

13. Selection and Approval of Foundation Models

13.1 What They Are, and How They Are Used

StreamVLM™ turns an operator's plain-text description into a working detector, and Ask IREX turns a plain-language request into an investigation across the archive. Both are built on large open-weight models published by third parties: vision-language models that understand an image and a text instruction together, and language models that interpret a request. IrexAI uses these models **as published**. It does not fine-tune or otherwise alter their weights. Their behavior inside the platform is shaped by IrexAI's own system prompts, output schemas, function definitions, and platform logic, all of which are IrexAI code under version control. Both capabilities are shipping in beta on selected instances and are described that way on every external surface.

13.2 Selection Criteria

A foundation model may be considered for the platform only if it meets all of the following. The CTO records the assessment in the Model Register.

Criterion	Requirement
License	Permits commercial use and self-hosting, including on customer-hosted instances, without usage reporting to the publisher and without terms that conflict with the customer's agreement.
Offline operation	Weights are downloadable and run fully offline. The model and its serving software make no network calls of their own and carry no telemetry.
Provenance and integrity	Obtained from the publisher's official channel; checksum recorded and verified on every download; safe serialization; scanned before loading (§12.1).
Documentation	The publisher's model card, training-data description, and safety evaluation are reviewed and filed. Models without such documentation are not selected.
Task performance	Meets IrexAI's benchmark for the intended function: detection accuracy on IrexAI's StreamVLM™ evaluation set of real CCTV frames and prompts; function-selection and grounding accuracy on IrexAI's Ask IREX evaluation set.
Resource fit	Runs within the GPU node specified for the instance at the required frame rate and latency.
Security history	Known vulnerabilities and jailbreak history reviewed; a vulnerability-watch source identified (§18.2).
Replaceability	The platform's model interface allows the model to be swapped for another that meets these criteria, so that a license change or vulnerability never leaves the platform without a path forward.

13.3 Why This Policy Does Not Name the Models

This Policy, and IrexAI's public materials, describe the class of model and not the specific model or version. Two reasons. Model identities change with each release cycle, and a named version in a public document is stale within months. More important, the exact model and version are part of the platform's attack surface: an adversary who knows them knows which published jailbreaks and weaknesses to try. Customers whose security review requires it receive the current model identity, version, license, publisher documentation, and checksum under a non-disclosure agreement, and are notified of changes through release notes.

13.4 No Fine-Tuning

IrexAI does not modify foundation-model weights. Should IrexAI ever fine-tune a foundation model, the result becomes an IrexAI-owned artifact governed in full by Part A, the Model Register records it as such, and the regulatory analysis in §20 is redone, because a substantial modification can make the modifier the provider of the model.

13.5 Version Changes

Replacing a foundation model, or moving to a new version of the same model, is a change under the Change Management Procedure. The new version is benchmarked against the current one on IrexAI's evaluation sets (§13.2), must meet the regression tolerance of §11.2, is deployed to staging before production, and is recorded in the Model Register with its new checksum. Customers are notified through release notes.

14. Deployment Boundary: Designated Endpoints Only

Foundation models are served exclusively by IrexAI's inference services on designated endpoints inside the platform. The following rules are enforced by the platform's configuration and network policy, not left to operator discretion.

- **Inside the instance.** The inference service runs in the instance's own cluster on a GPU node owned by IrexAI (managed instances) or by the customer (customer-hosted instances). No frame, prompt, query, or result leaves the instance.
- **No third-party inference.** IrexAI does not use a cloud AI service for StreamVLM™ or Ask IREX inference, on any instance, in any region.
- **Reachable only by the platform.** Designated endpoints accept requests only from authenticated platform services. They are not exposed to end users, to the internet, to other instances, or to the Sover messenger. A request from Sover reaches the IREX platform's API as the authenticated user, and the platform, not Sover, calls the model.
- **No egress.** Inference services have no outbound network access. The model cannot call out, and nothing it emits can be sent anywhere except back to the platform service that asked.
- **Bounded consumption.** Requests are scheduled through a priority queue with fair allocation across camera channels; stale frames are dropped rather than queued without bound. StreamVLM™ detectors carry confidence thresholds and cooldown intervals that suppress duplicate alerts.
- **Logged.** Every request to a designated endpoint is logged with the platform request identifier, the calling service, the operator, and the Case ID where one applies.

15. Capability Constraints: No Tools Beyond the Design

A general-purpose model is capable of far more than the task IrexAI gives it. The platform is designed so that the extra capability is unreachable.

15.1 StreamVLM™

The platform selects frames from a camera stream (at a fixed interval or when the scene changes), crops them to the region of interest, and sends each frame, or a before-and-after pair, to the model together with a detector prompt. The prompt is assembled by the platform from the operator's plain-text description and IrexAI's fixed system prompt. The model returns a structured verdict (match or no match, a confidence, a short description) that the platform validates against a schema before it becomes an event. The model:

- makes **no tool calls** of any kind;
- keeps **no memory** between requests; each frame is evaluated on its own;
- has **no access** to the archive, the event database, watchlists, other cameras, or the network;
- has **no biometric function**. StreamVLM™ evaluates scenes and conditions (a person lying down, flooding, graffiti, an unattended object). It cannot identify a person, and it has no watchlist to match against. Identification on the platform exists only in the face recognition module, under its own database-size limit and Case ID gate.

15.2 Ask IREX

Ask IREX interprets an operator's natural-language request and plans the searches that answer it. To do so it may call only a **fixed set of platform functions** that IrexAI designed for it: search people, vehicles, and events in the archive; retrieve results and context; and compose a summary of what was found. The rules:

- The function set is defined in IrexAI code, versioned, and approved by the CTO. It cannot be extended at runtime by the model, by an operator, or by content in a request.
- Every function is **executed by the platform**, never by the model, through the same IREX API a human operator uses, under the calling operator's session and permissions (§16).
- The model has **no code execution, no web browsing, no file-system access, no outbound network, no messaging or e-mail, no configuration or permission management, and no ability to create tools or alter its own instructions**.
- Functions with side effects beyond search and retrieval (for example, saving a result set or exporting media) require the operator's explicit confirmation in the interface before the platform executes them, and are logged as the operator's actions.

15.3 Output Handling

Model output is data, not instructions. The platform validates every output against its schema, rejects malformed output, never executes or interprets output as code or commands, and renders Ask IREX summaries alongside the platform records they were drawn from, marked as AI-generated. A StreamVLM™ alert is a detection for a person to verify, subject to its confidence threshold. An Ask IREX answer is a guide to the underlying video and event records, which remain the evidence.

16. Access-Control Inheritance and Case ID

The platform's permission model and Case ID requirement apply to a model exactly as they apply to a person, because the platform enforces them on every request and the model has no way around them.

- **No credentials.** Foundation models hold no tokens, keys, or database connections. There is nothing in the model's possession that could be leaked or misused.
- **The operator's permissions, every time.** Each function Ask IREX calls is executed under the requesting operator's own session. The platform evaluates role-based access control for that call as it would for a manual search, and returns to the model only the results that operator is entitled to see. A model cannot reach a camera, a recording, or a database that the operator cannot reach.
- **Camera scope for detectors.** A StreamVLM™ detector processes frames only from cameras its owner is authorized for, and the events it generates are visible under the same user-group permissions as any other event.
- **Case ID before the search.** An Ask IREX investigation requires a Case ID before any search runs, and the platform records that Case ID on every function call the agent triggers. Creating or changing a StreamVLM™ detector is alarm-monitor management, which requires a Case ID.
- **Logbook.** Every request, function call, and result is recorded in the non-erasable Logbook against the operator, the camera, and the Case ID, and exported daily in digitally signed form, so that the operator's supervisors and independent oversight bodies can review what the AI did and on whose authority.
- **Prompt injection is contained by design.** Text that appears inside a frame (a sign, a printed instruction) or inside an operator's request is treated by the platform as data. The model has no tool that an injected instruction could invoke, no data beyond the operator's permissions that it could disclose, and no network on which to send anything. The residual effect of an injection is a wrong verdict or an unhelpful answer, which the confidence threshold, human verification, and the Logbook address.

17. Prompt, Output, and Content Safety

- **System prompts are IrexAI code.** The instructions that confine a model to its task are versioned, reviewed, and not editable by operators. Detector prompts and Ask IREX requests are operator input and are logged as such.
- **Task confinement.** System prompts and output schemas restrict the model to the designed function. A request outside the function (for example, a request to identify a person, to change a setting, or to reach an external system) has no function to call and returns no result.
- **Grounding.** Ask IREX answers are built from the results the platform's functions return. The summary is shown with those records so that an operator can check it, and it carries an AI-generated label.
- **Hallucination handling.** StreamVLM™ verdicts are filtered by confidence threshold and cooldown, and reach an operator as alerts to verify. Ask IREX cannot invent a record: it can only summarize records the platform returned.
- **Retention of prompts.** Operator prompts and model outputs are retained with the Logbook under the operator's configured retention, as part of the audit record, not as training data. IrexAI does not use operator prompts, frames, or results to train or evaluate any model.

18. Monitoring, Vulnerability Management, and Model Incidents

18.1 Production Monitoring

For every production model, IrexAI monitors what the instance's operator permits it to see in aggregate (detection volumes, confidence distributions, inference latency, and error rates) and what customers report through support. Drift indicators (a shift in detection volume or confidence distribution without a corresponding change in the scene) trigger review. On customer-hosted instances, monitoring runs locally and is reviewed with the customer.

18.2 Vulnerability Watch

The CISO maintains a vulnerability watch for each foundation model, each pretrained base, the inference engine, and the serving stack: publisher advisories, vulnerability databases, and the adversarial-technique catalog MITRE ATLAS. Findings are assessed and remediated under the Vulnerability Management Procedure (DOC-ISMS-PR-003) at its severity timelines. Because foundation models are pinned and replaceable (§13.2), a model with an unfixable weakness is replaced under §13.5.

18.3 Model Incidents

The Incident Response Policy (DOC-ISMS-POL-006) applies. The following categories are recognized as model incidents and recorded in the Incident Register with the prefix **MODEL-**.

Category	Example	Severity Floor and First Action
Dataset permission breach	Imagery found in a dataset without a permission instrument on file	**High.** Quarantine the dataset; identify models trained on it; Data Protection function assesses notification duties.
Suspected data or model poisoning	Anomalous label batch; a model that systematically misses a specific target	**High.** Freeze the pipeline; re-evaluate on the sealed test set; trace provenance through hash inventories.
Artifact integrity failure	Signature or checksum mismatch at deployment	**High.** Block deployment; investigate the registry and pipeline; rotate signing material if compromise is suspected.
Foundation-model or serving-stack vulnerability	A published jailbreak or a serving-stack vulnerability affecting a pinned model	**Per CVSS under DOC-ISMS-PR-003.** Assess exposure given §14 to §16; patch or replace the model under §13.5.
Unintended capability	A model reaches a function, dataset, or destination outside its design	**Critical.** Disable the capability; treat any data exposure under the Data Protection Policy breach procedure.
Systematic bias discovered	Field reports or an assessment show a demographic performance gap above the recorded spread	**High.** Bias assessment; disclosure to affected customers per §11.3; mitigation plan.
Field accuracy regression	A module's false-positive rate rises materially after a release	**Medium.** Rollback to the previous signed version if warranted; targeted data and retraining.
Model theft or exfiltration	Unauthorized copy of model artifacts from a registry or node	**High.** Revoke access; legal notice under the customer's agreement where applicable; Risk Register update.

Customers affected by a model incident are notified within the timeframes of the Incident Response Policy and their agreements, with the model version, the effect on detections, and the corrective release.

19. Transparency to Customers and Oversight Bodies

19.1 What IrexAI Discloses

- Which platform capabilities use in-house detector models and which use foundation models, and that StreamVLM™ and Ask IREX are in beta on selected instances.
- The dataset source categories in §8.1 and the prohibited sources in §8.2. On request, how a customer's own Data Sets were used.
- Accuracy with its basis, known limitations, and the results of the latest demographic bias assessment for Tier 1 models (§11.3), plus the published record in Appendix C.
- A customer-facing model card summary for any production model, on request.
- Model changes, through release notes for every platform version.
- Under a non-disclosure agreement, where a customer's security review requires it: the identity, version, license, publisher documentation, and checksum of each foundation model in use; the internal model card; and evaluation records.

19.2 What IrexAI Does Not Publish, and Why

Model architectures, tuning parameters, internal benchmark figures, exact dataset inventories, and the identity of foundation models are RESTRICTED. Publishing them would expose the platform's attack surface and IrexAI's intellectual property without improving a customer's ability to verify the platform's behavior, which the Logbook already provides.

19.3 Independent Oversight

The signed daily Logbook export gives ethics committees, inspectors general, and court-appointed auditors an independent record of every AI action on an instance, including actions taken through Ask IREX and detections raised by StreamVLM™, without access to the platform and without IrexAI's involvement.

20. Alignment with External Frameworks

IrexAI states alignment, not certification. The table maps this Policy to the frameworks customers most often ask about.

Framework	How This Policy Aligns	Sections
NIST AI Risk Management Framework 1.0 (AI RMF, NIST AI 100-1)	Govern: roles, register, review. Map: purpose statements, tiers, source rules. Measure: evaluation protocol, bias assessment, monitoring. Manage: integrity controls, incident categories, capability constraints.	§4 to §7, §11, §12, §18
NIST Generative AI Profile (NIST AI 600-1)	Applied to Part B: provenance and supply chain of foundation models, confinement of capability, output handling, human oversight, incident response for generative components.	§13 to §18
ISO/IEC 42001:2023 (AI management systems)	This Policy is IrexAI's artifact for the developer role for its own models and the integrator role for foundation models: AI policy, roles, risk and impact assessment, data governance, lifecycle, documentation, monitoring, and improvement. Departures are recorded in the Statement of Applicability.	Whole document
ISO/IEC 23894:2023 (AI risk management), **ISO/IEC 42005:2025** (AI system impact assessment), **ISO/IEC 5259** series (data quality for ML), **ISO/IEC TR 24027:2021** (bias), **ISO/IEC 24029** (robustness of neural networks)	Risk tiers and purpose-and-harm assessment; datasheets, provenance, and quality assurance; demographic evaluation protocol; robustness testing.	§7, §8, §9, §11.3, §11.5
EU Artificial Intelligence Act (Regulation (EU) 2024/1689)	The platform's biometric functions are high-risk under Annex III. This Policy addresses the provider obligations most relevant to models: data and data governance (Article 10: permitted sources, datasheets, bias examination), technical documentation (Article 11: model cards, register), record-keeping (Article 12: Logbook), transparency to deployers (Article 13: §19), human oversight (Article 14: §15.3, §16), and accuracy, robustness, and cybersecurity (Article 15: §11, §12, §14). Foundation models are used unmodified, so the general-purpose AI model provider obligations of Chapter V rest with their publishers; IrexAI reviews and files their documentation (§13.2). Because IrexAI does not fine-tune them, it does not become their provider by substantial modification (Article 25). The Act's own data-governance article recognizes synthetic data as a way to detect and correct bias while limiting the processing of sensitive personal data (Article 10(5)), which is how §10 uses it. Under the 2026 amending regulation, the obligations for Annex III high-risk systems are scheduled to apply from December 2027; IrexAI applies this Policy now rather than waiting for that date. The platform is user-hosted, so compliance is established per deployment; this Policy supports it and does not claim it.	§8, §11, §13, §15, §16, §19
OWASP Top 10 for LLM Applications (2025)	Prompt injection (LLM01), sensitive information disclosure (LLM02): §16. Supply chain (LLM03): §13. Data and model poisoning (LLM04): §8, §12.3. Improper output handling (LLM05): §15.3. Excessive agency (LLM06): §15. System prompt leakage (LLM07): §17. Misinformation (LLM09): §15.3, §17. Unbounded consumption (LLM10): §14. Vector and embedding weaknesses (LLM08): not applicable; the platform does not retrieve from a document store for these functions.	§13 to §17

TABLE CONTINUED

Framework	How This Policy Aligns	Sections
MITRE ATLAS	Threat modeling reference for poisoning, model theft, evasion, and prompt injection; source for the vulnerability watch.	§12, §18.2
CISA and NSA guidance: *Deploying AI Systems Securely* (2024); *AI Data Security* (2025)	Secure deployment environment, integrity verification of model artifacts, data provenance and integrity, and monitoring reflect the joint guidance for organizations deploying AI systems.	§8, §12, §14, §18
FBI CJIS Security Policy	Model governance supports the auditing and accountability, configuration management, and system integrity areas for agencies handling criminal justice information: attributable outputs, signed artifacts, controlled change.	§12.4, §16, §18
US federal AI use and acquisition (OMB memoranda M-25-21 and M-25-22, April 2025)	IrexAI can supply the documentation federal agencies are directed to obtain from vendors: model cards, data provenance by source category, evaluation results, and change history; models run inside the agency's environment, which avoids dependence on a vendor-hosted model.	§6, §19
GDPR, CCPA/CPRA, and biometric-privacy laws	Training data containing personal or biometric data is handled under the Data Protection Policy: lawful basis confirmed at intake, DPIA for biometric training sets, data-subject requests, no training on customer personal data outside a signed Data Set clause.	§4.5, §8, §17

21. Exceptions

Exceptions to this Policy shall be requested in writing to the CTO, identify the section concerned, the reason, the compensating control, and an expiry date not more than six months out, and be recorded in the Exception Register and reviewed at the next quarterly Model Register review. The following exceptions **shall not** be granted:

- training, validating, or testing on imagery from a source other than the four in §8.1, or on any source in §8.2;
- admitting an unreviewed machine-proposed label to a training set;
- releasing a Tier 1 model without a demographic bias assessment;
- serving a foundation model outside a designated endpoint, exposing it to end users or the internet, or using a third-party cloud inference service for StreamVLM™ or Ask IREX;
- granting a foundation model a function, tool, or data access beyond the design in §15, or credentials of its own;
- deploying a model artifact whose signature does not verify.

22. Policy Review and Maintenance

22.1 Review Cycle

- **Annual:** full review by the CTO, approved with the CEO, on the anniversary of the effective date, informed by the annual model-governance report (models released, evaluations, bias assessments, incidents, exceptions).
- **Quarterly:** Model Register review by the CTO; Tier 2 false-positive review; vulnerability-watch review by the CISO.
- **Ad hoc:** on a new model class, a foundation-model change, a Critical or High model incident, a material regulatory change, or a change in a permission instrument.

22.2 Change Process

Material changes (a new dataset source category, a relaxation of a capability constraint, a change to the exceptions that shall not be granted) require approval by the CEO and the CTO. Minor changes (clarifications, editorial corrections, new references) may be approved by the CTO with notice to the CEO.

22.3 Document History

Version	Date	Author	Changes
1.0	September 7, 2026	CTO	Initial issue. Two-part structure (in-house detector models; foundation models in StreamVLM™ and Ask IREX); four permitted dataset sources; human review of all labels; synthetic-data rules; evaluation and bias-testing protocol; integrity and supply-chain controls; designated-endpoint, capability, and access-control constraints for foundation models; framework alignment.

23. References

- NIST AI 100-1, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, January 2023
- NIST AI 600-1, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, July 2024
- NIST AI 100-2 E2025, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, March 2025
- NIST SP 800-218A, *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models*, July 2024
- ISO/IEC 42001:2023, *Artificial intelligence management system: Requirements*
- ISO/IEC 23894:2023, *Artificial intelligence: Guidance on risk management*
- ISO/IEC 42005:2025, *Artificial intelligence: AI system impact assessment*
- ISO/IEC 5259 series, *Data quality for analytics and machine learning (ML)*, Parts 1 to 4 (2024) and Part 5 (2025)
- ISO/IEC TR 24027:2021, *Bias in AI systems and AI aided decision making*; ISO/IEC TR 24029-1:2021 and ISO/IEC 24029-2:2023, *Assessment of the robustness of neural networks*
- ISO/IEC 27001:2022 and ISO/IEC 27002:2022, information security management systems and controls
- Regulation (EU) 2024/1689, the *Artificial Intelligence Act*, Articles 9 to 15, 25, and Chapter V, as amended by the 2026 Digital Omnibus regulation on AI
- Regulation (EU) 2016/679, the *General Data Protection Regulation*; California Consumer Privacy Act as amended by the California Privacy Rights Act
- OWASP, *Top 10 for Large Language Model Applications*, 2025 edition
- MITRE, *ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems*
- CISA, NSA, and partners, *Deploying AI Systems Securely*, April 2024; *AI Data Security: Best Practices for Securing Data Used to Train and Operate AI Systems*, May 2025
- NIST, *Face Recognition Technology Evaluation (FRTE)* and *Face Analysis Technology Evaluation (FATE)*, the programs that succeeded the Face Recognition Vendor Test (FRVT) in 2023
- Office of Management and Budget, Memoranda M-25-21 and M-25-22, April 2025
- FBI, *Criminal Justice Information Services (CJIS) Security Policy*, current version
- Mitchell, M. et al., *Model Cards for Model Reporting*, FAT* 2019; Gebru, T. et al., *Datasheets for Datasets*, Communications of the ACM, 2021

24. Related IREX Documents

- IREX Ethical AI framework: Transparency by Design, the six pillars, and Case ID, published at irex.ai/ethical-ai
- DOC-ISMS-GOV-001, Information Security Policy
- DOC-ISMS-GOV-007, Statement of Applicability
- DOC-ISMS-GOV-009, Shared-Responsibility Matrix
- DOC-ISMS-POL-001, Access Control Policy
- DOC-ISMS-POL-004, Infrastructure and Cloud Security Policy
- DOC-ISMS-POL-005, Secure Development Policy
- DOC-ISMS-POL-006, Incident Response Policy
- DOC-ISMS-POL-007, Logging and Monitoring Policy
- DOC-ISMS-POL-008, Data Protection Policy
- DOC-ISMS-POL-011, Vendor Security Policy
- DOC-ISMS-POL-013, AI and External Data Provider Policy
- DOC-ISMS-PR-003, Vulnerability Management Procedure; DOC-ISMS-PR-004, Change Management Procedure; DOC-ISMS-PR-005, Secure Deployment Procedure
- DOC-ISMS-REG-005, AI Model Register (internal); DOC-ISMS-REG-002, Risk Register (internal)
- IREX Distribution and Licensing Agreement, clause 3.4 (Neural Network Training); IREX SaaS Agreement

Appendix A: Dataset Sources Quick Reference

May Be Used for Training, Validation, or Testing	May Never Be Used
Public open-source datasets whose license permits commercial use and model training	Any customer production data: video, frames, events, watchlists, exports, logs
Synthetic imagery produced by IrexAI (3D rendering, generative models, reference compositing), tagged as synthetic	Feeds from cameras in public spaces, including cameras in customer deployments
Customer Data Sets provided under clause 3.4 of the Distribution and Licensing Agreement or an equivalent written clause, non-individually identifiable, for improving the Product only	Images of people scraped from the web, social media, or news media
Imagery from development cameras operated by IrexAI or its partners, not in public spaces, used only for model development	Datasets with licenses that forbid commercial use or training, or withdrawn by their authors; broker data without a lawful basis; imagery of unknown provenance

Every label a model learns from is written or reviewed by a person. Every dataset has a datasheet and a permission instrument on file before its first use.

Appendix B: Foundation-Model Capability Boundary

Capability	StreamVLM™	Ask IREX
Where it runs	Designated endpoint inside the instance, on the instance's GPU node	Designated endpoint inside the instance, on the instance's GPU node
Third-party cloud inference	Never	Never
Input	Selected frames from authorized cameras plus the platform-assembled detector prompt	The operator's request plus results returned by platform functions
Output	Schema-validated verdict: match, confidence, short description	Function selections executed by the platform; a summary shown with its source records
Tool calls	None	Fixed, CTO-approved set of platform search and retrieval functions only
Code execution, web browsing, file system, outbound network, messaging	None	None
Memory across requests	None	Within one investigation session only; nothing retained for training
Credentials held by the model	None	None
Data access	The frame it is given	Only what the calling operator's permissions allow, returned by the platform
Biometric identification	None	Only by invoking the platform's face recognition function under its Case ID and database-size limits, as the operator could manually
Case ID	Required to create or change a detector	Required before any search; recorded on every function call
Logging	Every request and event in the Logbook	Every request, function call, and result in the Logbook
Weights	As published, pinned, checksum-verified, not fine-tuned	As published, pinned, checksum-verified, not fine-tuned
Availability	Beta, selected instances	Beta, selected instances

Appendix C: Published Evaluation Evidence

The record below is public and dated. Current figures for any model are available to customers on request under §19.

Date	Evaluation	Result as Published
March 2021	NIST Face Recognition Vendor Test (now FRTE), 1:N identification, Border and Kiosk datasets	IrexAI's face recognition algorithm ranked first in accuracy among US companies and in the top 10 of 268 algorithms worldwide on these datasets. This is IrexAI's most recent submission. IrexAI has not entered later rounds because the other FRTE datasets (visa and mugshot photos) do not represent the angle, resolution, and distortion of real CCTV, and the benchmark's offline per-image runtime allowance does not reflect real-time operation.
June 2024	Synthetic data in railway incident detection	Adding 5,000 synthetic images to a fine-tuning set of 6,000 real captures, on a 50,000-image real pretraining set, raised mean average precision by 4.6%.
July 2024	Demographic bias study of the suspect recognition engine on the Racial Faces in-the-Wild benchmark (four groups, approximately 3,000 individuals and 6,000 verification pairs each)	Verification accuracy between 97.20% and 99.15% across groups for the GPU version and between 96.67% and 98.55% for the CPU version, a spread of 1.95% and 1.88% respectively. Findings were shared with all platform customers.
July 2026	Human-reviewed synthetic data for weapon detection	Roughly 75% of generated samples rejected in human quality control; approximately 300 verified examples per weapon class enter training; models pretrained on the curated synthetic set and fine-tuned on real annotated CCTV.

Source: IREX news archive at irex.ai/news.

Approval and Signatures

The undersigned approve this AI Model Governance Policy on behalf of IrexAI Inc. and confirm that it states the requirements IrexAI applies to every AI model delivered in the IREX Ethical AI Platform.

Role	Name	Signature	Date
Co-Founder & Chief Executive Officer	Calvin Yadav	_____	September 7, 2026
President, Chief Technology Officer & Co-Founder (Document Owner)	Nik Ptitsyn, PhD	_____	September 7, 2026

Classification: Public · **Next Review Date:** September 7, 2027 · **Last Updated:** September 7, 2026