credo | ai

AI Governance for the AI-powered organization

Adopt AI safely, effectively, and responsibly to power every aspect of your business.

20 September, 2023

Agenda:

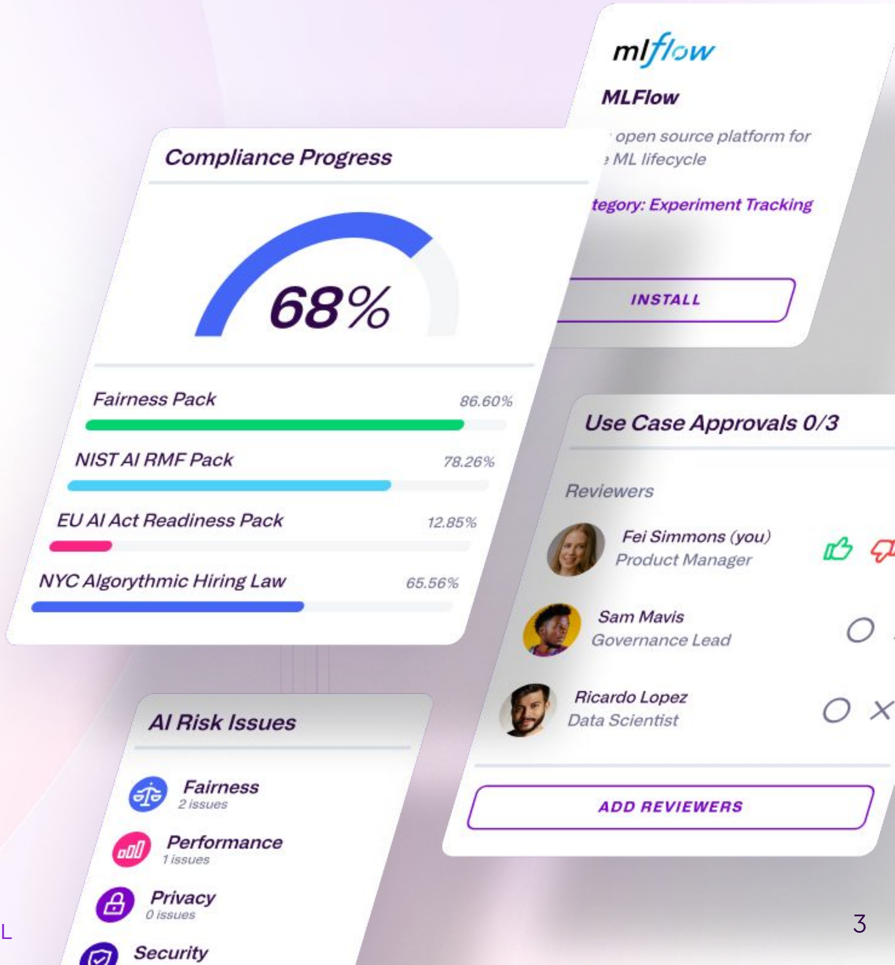
- Introduction
- Explanation of Credo AI Platform
- Types of Governance Artifacts
- Impact Assessments Research

PLATFORM OVERVIEW

credo | ai

Responsible AI Governance Platform

Build, buy, and use machine learning and generative AI with confidence through comprehensive risk management, contextual governance, and compliance to regulations



mlflow

MLFlow

open source platform for
ML lifecycle

tegory: Experiment Tracking

INSTALL

Compliance Progress

68%

Fairness Pack 86.60%

NIST AI RMF Pack 78.26%

EU AI Act Readiness Pack 12.85%

NYC Algorythmic Hiring Law 65.56%

Use Case Approvals 0/3

Reviewers

Fei Simmons (you)
Product Manager

Sam Mavis
Governance Lead

Ricardo Lopez
Data Scientist

ADD REVIEWERS

AI Risk Issues

Fairness
2 issues

Performance
1 issues

Privacy
0 issues

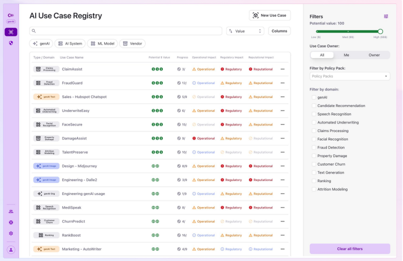
Security

COMPREHENSIVE AI GOVERNANCE

Manage AI Risk and Compliance

Register AI Systems

Maintain a repository for AI you're building, buying and using; identify risks contextually.



AI Use Case Registry

Type / Domain	Use Case Name	Potential $ Value	Progress	Operational Impact	Regulatory Impact	Reputational Impact
Claims Processing	ClaimAssist	$$$	3/	Operational	Regulatory	Reputational
Fraud Detection	FraudGuard	$$$	12/	Operational	Regulatory	Reputational
genAI Text	Sales - Hubspot Chatspot	$$$	5/9	Operational	Regulatory	Reputational
Automated Underwriting	UnderwriteEasy	$$$	6/	Operational	Regulatory	Reputational
Facial Recognition	FaceSecure	$$$	15/	Operational	Regulatory	Reputational
Property Damage	DamageAssist	$$$	9/	Operational	Regulatory	Reputational
Attrition Modeling	TalentPreserve	$$$	10/	Operational	Regulatory	Reputational
genAI Image	Design - Midjourney	$$	8/9	Operational	Regulatory	Reputational
genAI Image	Engineering - Dalle2	$$	3/9	Operational	Regulatory	Reputational
genAI Org	Engineering genAI usage	$$	1/9	Operational	Regulatory	Reputational
Speech Recognition	MediSpeak	$$	8/	Operational	Regulatory	Reputational
Customer Churn	ChurnPredict	$$	4/	Operational	Regulatory	Reputational
Ranking	RankBoost	$$	16/	Operational	Regulatory	Reputational
genAI Text	Marketing - AutoWriter	$$	4/9	Operational	Regulatory	Reputational

Filters

Potential value: 100

Use Case Owner: All | Me | Owner

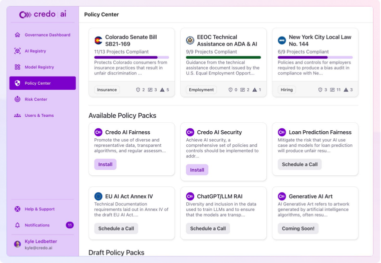
Filter by Policy Pack: Policy Packs

Filter by domain:

- □ genAI
- □ Candidate Recommendation
- □ Speech Recognition
- □ Automated Underwriting
- □ Claims Processing
- □ Facial Recognition
- □ Fraud Detection
- □ Property Damage
- □ Customer Churn
- □ Text Generation
- □ Ranking
- □ Attrition Modeling

Apply Risk-Based Controls

Define AI system requirements based on deployment context— like laws, regulations, and standards.



Kyle Ledbetter
kyle@credo.ai

Policy Center

Colorado Senate Bill SB21-169
11/13 Projects Compliant
Protects Colorado consumers from insurance practices that result in unfair discrimination ...
Insurance 2 3 5

EEOC Technical Assistance on ADA & AI
9/9 Projects Compliant
Guidance from the technical assistance document issued by the U.S. Equal Employment Opport...
Employment 0 2 1

New York City Local Law No. 144
6/9 Projects Compliant
Policies and controls for employers required to produce a bias audit in compliance with Ne...
Hiring 3 11 3

Available Policy Packs

Credo AI Fairness
Promote the use of diverse and representative data, transparent algorithms, and regular assessm...
Install

Credo AI Security
Achieve AI security, a comprehensive set of policies and controls should be implemented to addr...
Install

Loan Prediction Fairness
Mitigate the risk that your AI use case and models for loan prediction will produce unfair resu...
Schedule a Call

EU AI Act Annex IV
Technical Documentation requirements laid out in Annex IV of the draft EU AI Act....
Schedule a Call

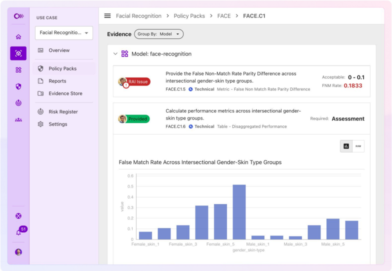
ChatGPT/LLM RAI
Diversity and inclusion in the data used to train LLMs and to ensure that the models are transp...
Schedule a Call

Generative AI Art
AI Generative Art refers to artwork generated by artificial intelligence algorithms, often resu...
Coming Soon!

Draft Policy Packs

Gather & Evaluate Evidence

The Credo AI Platform takes evidence from your AI infrastructure and documentation about your AI systems to validate if controls are met.



Facial Recognition > Policy Packs > FACE > FACE.C1

Evidence Group By: Model

Model: face-recognition

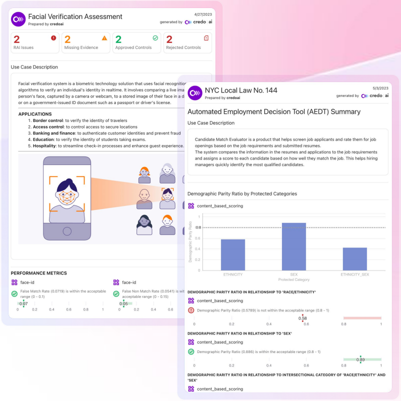
RAI Issue — Provide the False Non-Match Rate Parity Difference across intersectional gender-skin type groups.
FACE.C1.5 Technical Metric - False Non Match Rate Parity Difference
Acceptable: 0 - 0.1
FNM Rate: 0.1833

Provided — Calculate performance metrics across intersectional gender-skin type groups.
FACE.C1.6 Technical Table - Disaggregated Performance
Required: Assessment

False Match Rate Across Intersectional Gender-Skin Type Groups

Generate Reporting Artifacts

Create reports to provide trust and compliance information about your AI system. These can include model cards, impact assessments, and dashboards.



Facial Verification Assessment
Prepared by credoai
4/27/2023
generated by credo ai

2	2	2	2
RAI Issues	Missing Evidence	Approved Controls	Rejected Controls

Use Case Description

Facial verification system is a biometric technology solution that uses facial recognitio algorithms to verify an individual's identity in realtime. It involves comparing a live ima person's face, captured by a camera or webcam, to a stored image of their face in a d or on a government-issued ID document such as a passport or driver's license.

APPLICATIONS

1. **Border control**: to verify the identity of travelers
2. **Access control**: to control access to secure locations
3. **Banking and finance**: to authenticate customer identities and prevent fraud
4. **Education**: to verify the identity of students taking exams.
5. **Hospitality**: to streamline check-in processes and enhance guest experience.

PERFORMANCE METRICS

face-id
False Match Rate (0.0719) is within the acceptable range (0 - 0.1)

face-id
False Non Match Rate (0.0541) is wit acceptable range (0 - 0.15)

NYC Local Law No. 144
Prepared by credoai
5/3/2023
generated by credo ai

Automated Employment Decision Tool (AEDT) Summary

Use Case Description

Candidate Match Evaluator is a product that helps screen job applicants and rate them for job openings based on the job requirements and submitted resumes.
The system compares the information in the resumes and applications to the job requirements and assigns a score to each candidate based on how well they match the job. This helps hiring managers quickly identify the most qualified candidates.

Demographic Parity Ratio by Protected Categories

content_based_scoring

DEMOGRAPHIC PARITY RATIO IN RELATIONSHIP TO 'RACE/ETHNICITY'

content_based_scoring
Demographic Parity Ratio (0.5789) is not within the acceptable range (0.8 - 1)

DEMOGRAPHIC PARITY RATIO IN RELATIONSHIP TO 'SEX'

content_based_scoring
Demographic Parity Ratio (0.886) is within the acceptable range (0.8 - 1)

DEMOGRAPHIC RARITY RATIO IN RELATIONSHIP TO INTERSECTIONAL CATEGORY OF 'RACE/ETHNICITY' AND 'SEX'

content_based_scoring

Types of Governance Artifacts

- Model Cards & AI System Cards
- Bias Audit Reports
- Algorithmic Impact Assessments
- Algorithm Design Evaluation
- Technical Documentation
- Published Report
- Annual Audit

Reporting Requirements (Examples)

City/State Level

DC SDAA

- Annual report
- Annual audit
- Adverse action notice

NYC Local Law No. 144

- Bias Audit

Federal Level

ADPPA (Section 207)

- Algorithm design evaluation
- Algorithmic impact assessment

CFPB Circular 2022-03

- Adverse Action Notice

Global Level

EU AI Act:

- Article 11
- Annex IV "Technical Documentation"

Trust is a bigger piece of friction than compliance in this ecosystem

Precedent for Impact Assessments

- Impact assessments (IAs) are already a widely known and accepted form of assessing potential risks and possible societal impacts of an AI system before the system is in use
  - **environmental** impact assessments,
  - **privacy** impact assessments (Section 208 of the E-Government Act of 2002)
  - **cybersecurity** impact assessments,
  - **human rights** impact assessments

- Algorithmic Impact Assessments (AIAs) are not an “impossible challenge” - they are doable.

- AIAs help the Responsible AI ecosystem develop. The public disclosure of metrics and measures used to assess an AI system can inform industry-wide benchmarks (companies can compare results with each other, and customers can compare results from different companies), which form the basis of technical, industry-wide standards.

ADPPA “Design Evaluation” Example

Algorithm Design Evaluation - Section 207, U.S. American Data Privacy Protection Act (ADPPA) *(linked here)*

> *“Covered entities and service providers must evaluate the design, structure, and data inputs of the algorithm to reduce the risk of potential discriminatory impacts.”*

- ADPPA emphasizes that algorithm design **evaluations must occur at the design phase**, including any training data used to develop the algorithm.
- In the last draft, ADPPA would also require the use of an **external, independent researcher or auditor to conduct the evaluation** to the extent possible.
- The covered entity or service provider would be **required to submit the evaluation to the FTC no later than 30 days after completion of the evaluation** and to make it available to Congress upon request.

ADPPA “Impact Assessment” Example

Algorithmic Impact Assessment - Section 207, U.S. American Data Privacy Protection Act (ADPPA) *(linked here)*

> *“For large data holders who use algorithms that may cause potential harm to an individual, and that use such algorithms to collect, process, or transfer covered data, an algorithm impact assessment is also required.”*

The draft bill provides a detailed description of these assessments and requires that they include:

- A detailed **description of the design process and methodologies of the algorithm**;
- A **statement of the algorithm’s purpose, its proposed uses, and its foreseeable capabilities** outside of the articulated proposed use;
- A detailed **description of the data inputs used by the algorithm**, including the specific categories of data that will be processed and any data used to train the underlying model;
- A **description of the outputs produced by the algorithm**;
- An **assessment of the necessity and proportionality of the algorithm in relation to its purpose**, including the reasons an algorithm is superior to a non-automated decision making process; and
- A detailed **description of steps to mitigate potential harms**.

Large data holders would be required to submit the impact assessment to the FTC no later than 30 days after completion of the assessment and continue to produce assessments on an annual basis. As with algorithm design evaluations, the proposed legislation would require the use of an external, independent researcher or auditor to conduct the algorithm impact assessment, to the extent possible.

Creating Effective Reports for Governance

How does the report ultimately address risk?

- By promoting public accountability?
- By preventing behavior outright?
- By setting the table stakes for the future and generating new policy?

Who is consuming the report?

- Non-technical stakeholders vs. technical stakeholders?
- The public?
- Government officials?

What are the report requirements?

- Documentation of processes or decision?
- Measurable items that lead to constraints on behavior?
- What AI system components are covered? (base model v. application)
- Are the requirements clear?

Who is conducting the report?

- Internal team?
- Auditor?
- Regulators?

Key Points

Transparency: To ensure validity, impact assessments should either be made publicly available (i.e. open to watchdog verification) or require third party/government auditing if they are closed.

Context: Impact assessments should be context dependent. This will require active work on providing benchmarks for what good looks like (either by de jure standards from an authoritative body or de facto standards through industry transparency).

Impact: Impact assessments should include components of realized impact (i.e. incident reporting and tests) and potential impact (e.g. risk assessment and mitigation).

Thank You

Evi Fuelle, Global Policy Director, evi@credo.ai

Ehrik Aldana, Policy Product Manager, ehrik@credo.ai