Building Safe AI Systems with Model Cards

Samuel Lefcourt, PhD

MAY 1-5, 2026 · DSAI MT. WASHINGTON, BALTIMORE

Bio

Samuel Lefcourt, PhD obtained his doctorate from Johns Hopkins University in Computer Science, focusing on Contextual AI for Safety-Critical Applications. His mindset is grounded in building safe AI systems, following his experience with NIST in the field of AI Forensics. Applying that knowledge, his work spans defense missions ranging from terrestrial nuclear weapon detection to space systems cybersecurity across the DOD, DHS, AFRL. His work in healthcare revolves around precision medicine: biomarker segmentation and radiomic fingerprinting for clinical modeling.

He is the Co-Chair of the AI/ML COI at the Space ISAC, the threat intelligence source for the public and private sectors. Through this, Samuel has led collaborative efforts to establish and release AI standards for aerospace. Additionally, Samuel is the Founder & CEO of Lefcourt Medical AI, building explainable AI solutions for diagnostics.

AI Model Card

Flourish-A-Thon

“

the *Machine Learning Technology Readiness Levels* framework defines a principled process to ensure robust, reliable, and responsible systems while being streamlined for machine learning workflows

---

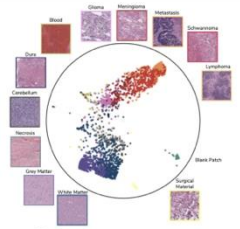
Lavin, A., Gilligan-Lee, C.M., Visnjic, A. *et al.* Technology readiness levels for machine learning systems. *Nat Commun* 13, 6039 (2022).

Example (TRL) Model Card

TECHNOLOGY NAME	Neuropathology Copilot v1.0
TRL	4 <link to previous cards>
R&D OWNER / REVIEWER	A. Lavin / G. Renard
PROD OWNER / REVIEWER	S. Wozniak / S. Jobs
COMPONENT CODES	1.1, 4.2, 4.3

TL;DR	Analyze WSI of brain tissue in 3 main steps: (1) unsupervised CV model produces Poincare manifold viz (Naud & Lavin '20), (2) domain expert selects data points, (3) U-Net classifier

Data considerations	3 datasets have been used to train and validate the system: 1. Open dataset (Naud & Lavin '20) 2. Pilot dataset provided by BioLab, v1.0 3. Simulated datasets (w/ structured domain randomization), v2.3
Ethics	Note the demographics info on specific Dataset Cards. Datasets anonymized, pipeline runs w/o metadata. The Latent Sciences Ethics Checklist has been completed.

Model / alg details	The SP-VAE model runs unsupervised on neurological whole-slide images (WSI), producing a latent manifold that represents a hierarchical organization of tissue types. An medical expert identifies several data points to inspect.  *Example visualization of the latent organization of brain tissue types.*
Metrics, results	Classification accuracy >0.97 on the 5 main brain cancer types. Inference per WSI runs ~1.0s on 2-GPU. Full quantitative reports: < link to experiments wiki >
Caveats, known edge cases, recommendations	Changing imaging sources will require retraining the full model (notably the SP-VAE annealing parameter). Whenever possible it is recommended that users provide feedback annotations. Non-tissue material is correctly flagged as anomalous.
Key assumptions	The training and production images are equivalent, specifically from the exact same sensor(s).
Intended use	The model must include human expert in the loop, and it has not yet been validated for other disease areas.

Model Card Components

1

Model Info
Model Data

2

Evaluation
Limitations

3

Ethics
Safety

Component 1

Model Information

- Inputs
  - Book, MRI image, foreign audio
- Outputs
  - Text summary, disease diagnosis, native audio
- Architecture (ResNet50)

Model Data

- Number of Datasets
  - Dataset Title
  - Dataset Origin
  - Number of Samples

Take 10 minutes to put your model information and data together

Performance Testing Strategies

Flourish-A-Thon

How to pick a performance metric(s)?

Define your task

1. Classification
2. Regression
3. Segmentation
4. Generation

Metric

- F1, Accuracy, Sensitivity, Specificity, ROC-AUC
- MAE, MSE, RMSE, R2
- Dice, IoU, Hausdorff Distance
- FID, SSIM

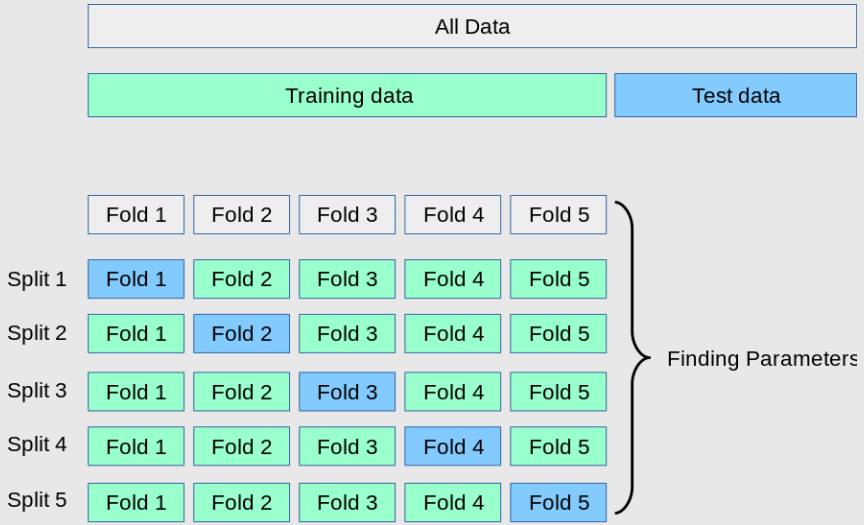
AI Validation Plan

Flourish-A-Thon

What is a validation plan?

How will you demonstrate that your model results are "trustworthy" and "believable"

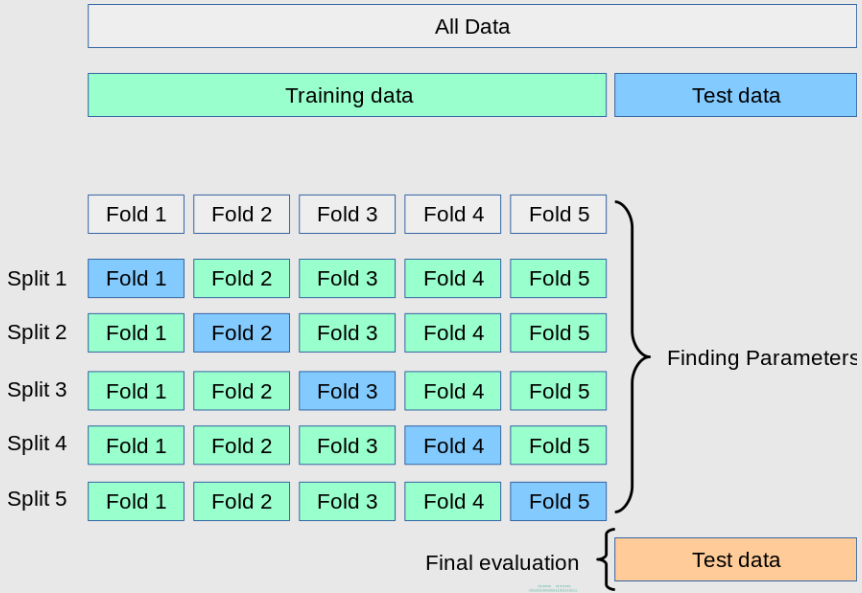
1. Single Dataset
   - Train, eval, test
   - Cross-Validation
   - Nested Cross-Validation



What is a validation plan?

How will you demonstrate that your model results are "trustworthy" and "believable"

1. Multiple Datasets
   - Train, eval, test
     - Test is now a "holdout" set
   - Cross-Validation
     - Test is now a "holdout" set
   - Cross-Validation
     - Each fold is a different dataset



Take 10 minutes to put your model information and data together

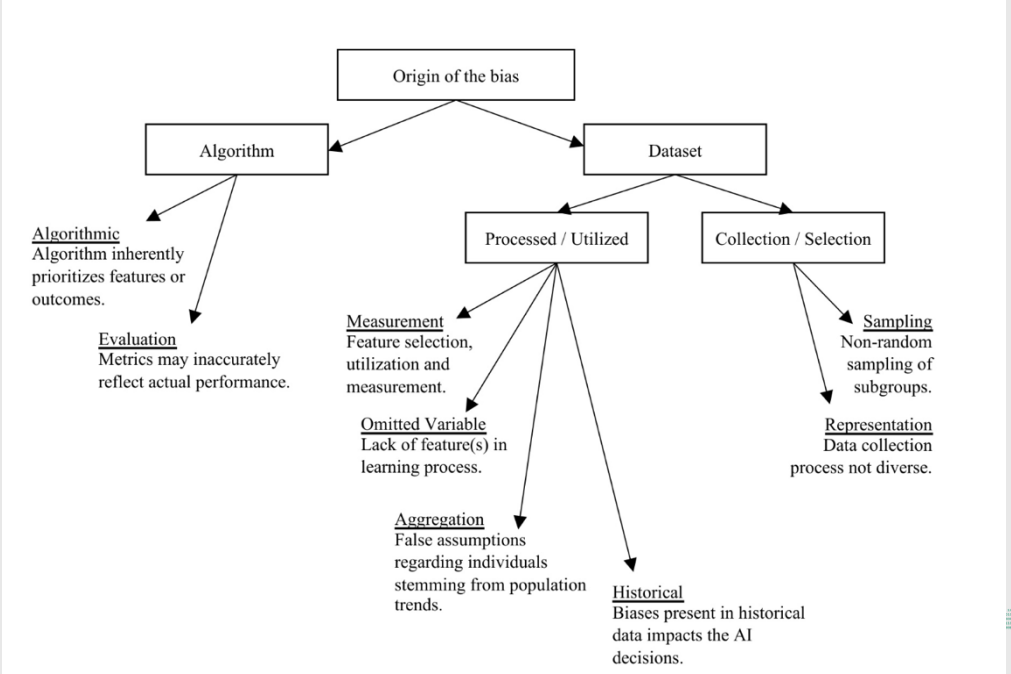
Evaluation

Approach: Gemini 3.1 Flash-Lite was evaluated across a range of benchmarks, including speed, reasoning, multimodal capabilities, factuality, agentic tool use, multi-lingual performance, coding, and long-context. Benchmark details on approach, results, and their methodologies can be found at: https://deepmind.google/models/evals-methodology/gemini-3-1-flash-lite

AI System Limitations

Limitations are important!

It is vital to understand the bounds of your AI system: where might it fail, how does your validation system fall short?



Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." ACM computing surveys (CSUR) 54.6 (2021): 1-35.

Flourish-A-Thon

Scope is important!

It is vital to understand the bounds of your AI system: where does it succeed?

Intended Usage and Limitations

Benefit and Intended Usage: Gemini 3.1 Flash-Lite is well suited for applications that require high volume, cost-efficient and low latency tasks.

Known Limitations: For more information about the known limitations for Gemini 3.1 Flash-Lite, see the Gemini 3 Pro model card.

Acceptable Usage: For more information about the acceptable usage for Gemini 3.1 Flash-Lite, see the Gemini 3 Pro model card.

Take 10 minutes to put your intended use and limitations together

Intended Usage and Limitations

Benefit and Intended Usage: Gemini 3.1 Flash-Lite is well suited for applications that require high volume, cost-efficient and low latency tasks.

Known Limitations: For more information about the known limitations for Gemini 3.1 Flash-Lite, see the Gemini 3 Pro model card.

Acceptable Usage: For more information about the acceptable usage for Gemini 3.1 Flash-Lite, see the Gemini 3 Pro model card.

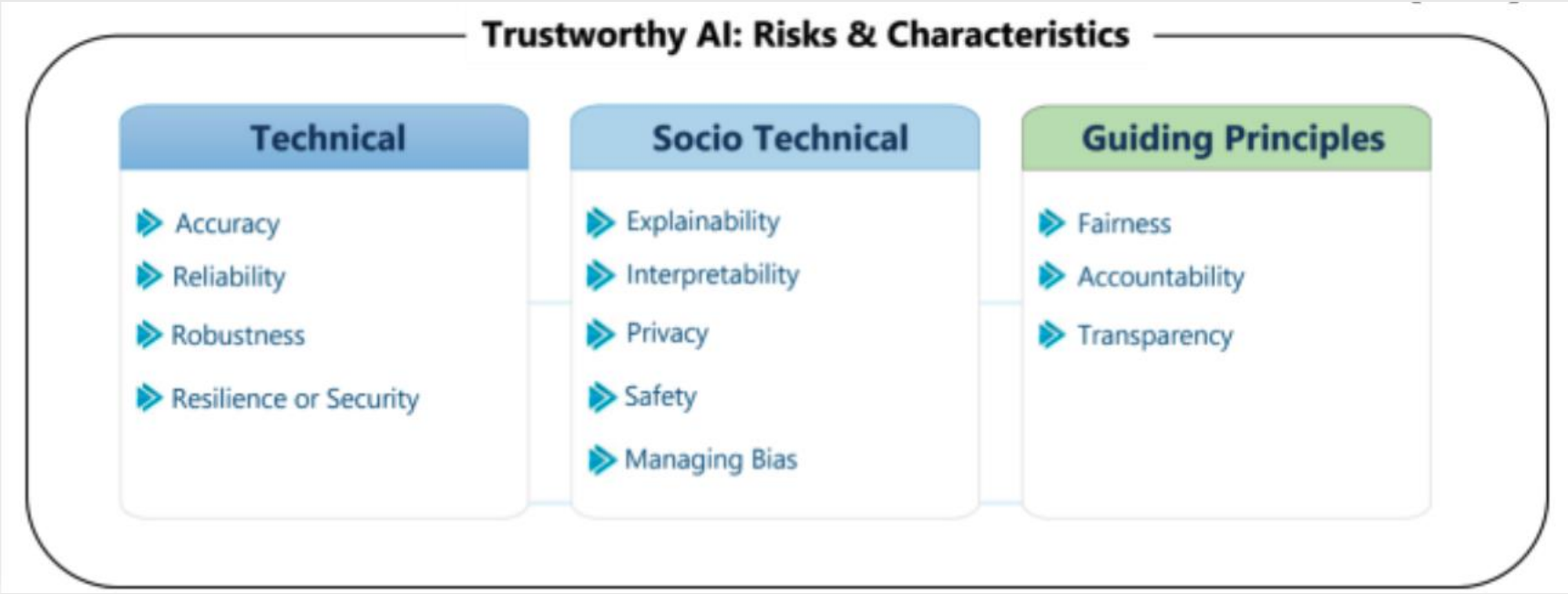
AI System Safety: A Sociotechnical Approach

Flourish-A-Thon

Sociotechnical design bridges the gap between societal needs and technological capabilities, treating them as one coherent system (Chen and Metcalf, 2024). As opposed to the traditional mindset of tackling ethics as a purely technical challenge, sociotechnical ethics approach safety as a combination of technical design and societal factors.

Lefcourt, Samuel ENHANCING TRUST IN SAFETY-CRITICAL AI: THE ROLE OF CONTEXTUAL INTEGRITY IN SPACE SYSTEMS CYBERSECURITY AND RADIOMICS

NIST Risk Management Framework (Ethics & Safety)



Take 10 minutes to put your ethics and safety together

Final Model Card

Flourish-A-Thon

Take the remainder of time to cohesively format your model card!

Flourish-A-Thon

Thank you!

Feel free to reach out!

www.linkedin.com/in/samuel-lefcourt

Flourish-A-Thon