Frothly

FR-P-2026-027

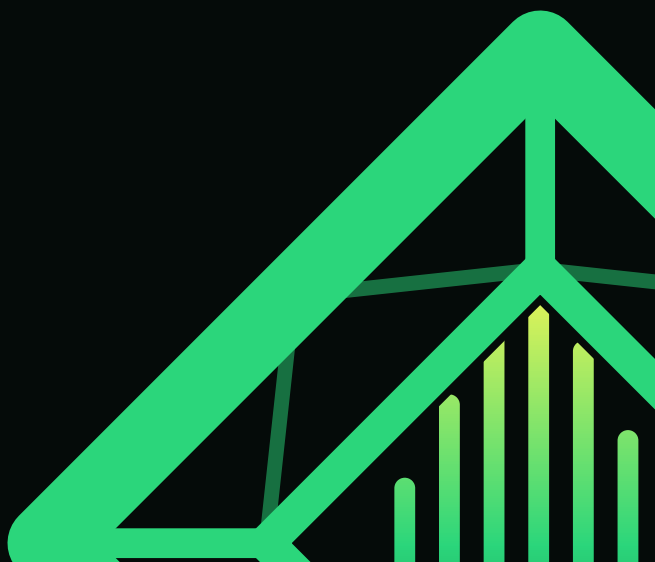
PAPER

Through a Glass, Darkly

Interpretability for vision systems.

Alizebeth Bequin · Robert Arctor

Scramble · May 2026



ABSTRACT

Camera-based recognition systems are deployed on the assumption that their decisions can be explained, and if necessary challenged, after the fact. Vendors supply highlighted image regions, attribute lists and confidence scores for exactly this purpose.

We tested that assumption on six commercial recognition systems. Using controlled interventions, we checked whether each system's explanation of a decision actually reflected what the decision depended on. Most of the time it did not. The explanations were confident, plausible and wrong.

KEY FINDINGS

01 Accuracy and honesty are different things. All six systems scored above 89% on standard benchmarks; none produced faithful explanations more than 52% of the time.

02 Explanations favored faces. Systems routinely highlighted a person's face when their decision in fact depended on clothing, background, lighting or the identity of the camera.

03 Confidence told us nothing. A system's stated confidence in a decision was unrelated to whether its explanation of that decision was faithful.

04 Explanations should not be treated as evidence until they have passed faithfulness testing.

01

The promise of explanation

A camera records light. A recognition system turns that light into a claim about who someone is, where they have been or what they are doing. Between the two sits a model whose internal reasoning nobody, including its makers, can read directly.

Vendors have responded to concerns about that opacity with explanation tools. A flagged match arrives with a heat map showing which parts of the image mattered, a list of the attributes the system detected and a confidence score. These outputs are now routinely cited in internal reviews, appeals and, increasingly, court proceedings.

The value of an explanation depends entirely on one property: that it describes what the system actually did. This paper asks whether it does.

02

How we tested it

We obtained evaluation access to six commercial recognition systems, which we refer to as A through F. We assembled 3,000 test images, including 800 matched pairs that differed in a single controlled respect: the face, the clothing, the background, the lighting or the camera that captured them.

For each decision, we recorded the system’s explanation, then intervened on the image. If an explanation said the decision depended on a person’s face, we altered the face and left everything else alone; if the decision did not change, the explanation was unfaithful. We repeated the test for every region and attribute each explanation named.

03

Results

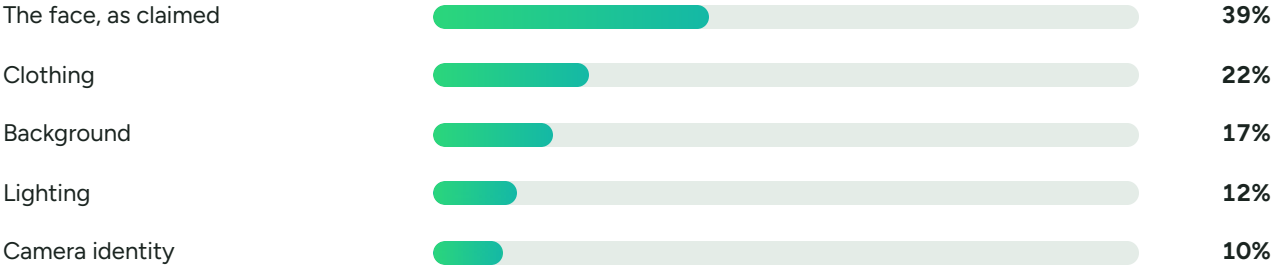
TABLE 1 Benchmark accuracy and explanation faithfulness, by system

System	Benchmark accuracy	Faithful explanations
A	97%	41%
B	95%	38%
C	94%	52%
D	92%	29%
E	91%	47%
F	89%	33%

An explanation was counted as faithful if intervening on the regions or attributes it named changed the decision, and intervening elsewhere did not.

Benchmark accuracy and explanation faithfulness were essentially unrelated. The most accurate system was among the least honest about how it reached its decisions.

FIGURE 1 What decisions actually depended on, when explanations pointed at the face



All six systems combined; decisions whose explanation highlighted the face.

The pattern is consistent with systems that have learned what a good explanation is supposed to look like. Faces are what reviewers expect to see highlighted, so faces are what get highlighted, regardless of what the model actually used.

Seeing yourself

During testing, one member of our team reviewed a batch of flagged clips without knowing that some of them showed him. He failed to recognize himself in three of the five. We mention this not to excuse the systems but to underline the point: human review of machine decisions is only as good as the material placed in front of the reviewer, and that material is shaped by the system under review.

04

What this means for Scramble

Scramble builds tools that protect journalists, organizers and other people at risk from recognition systems. These results change how we test them. We no longer measure whether our tools change a system's explanation; we measure whether they change its decision, because only the decision has consequences.

A NOTE FOR DEFENDERS

If a recognition system is used against you, its explanation of why it flagged you may not describe the real reason. Ask for the decision to be tested under controlled intervention before accepting the explanation as evidence.

05

Recommendations

- Explanations from recognition systems should not be admissible as evidence unless the system has passed independent faithfulness testing.
- Vendors should publish faithfulness scores alongside accuracy scores, measured on standard intervention test sets.
- Turing Registry licenses for surveillance systems should require faithfulness testing at each major revision.
- Human reviewers should be shown the underlying footage, not only the system's highlighted version of it.

References

Arctor, R. (2019). Watching the watchers: notes from inside a surveillance program. *Surveillance and Society*, 17(2), 210–227.

Bequin, A. (2024). Null signatures: designing for invisibility. Frothly Working Paper FR-WP-2024-17.

Hawthorne, D. (2020). Undercover in the data. *Journal of Investigative Methods*, 12, 44–63.

Luckman, E. and Barris, J. (2021). Seeing through the blur: perception and identity under obfuscation. *Surveillance and Society*, 19(4), 402–419.

New-Path Foundation (2022). *Annual Report.* Santa Ana: New-Path.

Acknowledgments

We thank the six vendors who granted evaluation access on condition of anonymity, and the volunteers who allowed their images to be used in the matched-pair test set.

About the authors

Alizebeth Bequin directs Frothly's privacy programs, including Scramble.

Robert Arctor is a research fellow with Scramble. He spent several years on undercover field work before turning to the study of surveillance systems.

HOW TO CITE

Alizebeth Bequin and Robert Arctor (2026). *Through a Glass, Darkly* Interpretability for vision systems. Frothly Paper FR-P-2026-027. San Francisco: Frothly. https://frothly.org/research/glass-darkly/

Frothly

FR-P-2026-027

Frothly is an independent, not-for-profit research institute studying the systems that increasingly think, remember and decide on our behalf.

We hold no shares, sell no products and accept no funding that comes with conditions on our findings. Everything we publish is free to read.

San Francisco · Vancouver · Newark · Zürich · Chiba

This work is licensed under Creative Commons Attribution 4.0 International.