

Semi and Fully Autonomous AI Systems

Part 3: Earning Autonomy

2026-10-26 09:00:00 -0500 · unpublished draft · review PDF

DRAFT · banner: assets/images/autonomous-pt3-banner.png (1920 × 1080), not made yet

Part 3: Earning Autonomy

A model earns autonomy one decision class at a time — and loses it the same way.

On Optimum Recognition Error and Reject Tradeoff.

— C. K. Chow, the title of a 1970 paper [1]



The letter

In 2018 a tenant-screening company paid \$3 million to settle FTC charges over the records it attached to applicants: an exact match on the last name, “soft” matches on first name, middle name and date of birth, and people denied housing over someone else’s record [6].

That system was a matching rule. Put a model in its place and the stakes don’t change; the questions do. Six months after a report goes out, a letter comes back disputing one line. Under the Fair Credit Reporting Act the clock is thirty days, and the person can ask how the line was checked [4]. So reconstruct it. Which model, which version, trained on what? How sure was it? Which rule turned its answer into a line on a report? Did a person look, and who?

If the answer is a shrug, the system was never autonomous. It was unsupervised.

This is Part 3 of a series. [Part 1](#) put a human in the loop as gatekeeper; [Part 2](#) moved the human onto the loop as watchdog, and promised to take the watchdog away too [8]. This part keeps the promise, one decision class at a time.

Where Parts 1 and 2 left us

Part 1 found the labels inside the loop: every expert correction is a training example. Part 2 took the gates out and put people on the rail, and ended on deciding **per gate, not per system**. Both left one question open: what evidence moves a decision up the ladder, and what moves it back down? [Autonomy is a](#)

spectrum answered it for tasks: “earned” means measured, on your own work [9]. Here is the measurement for a model’s decisions.

Accuracy is the wrong dial

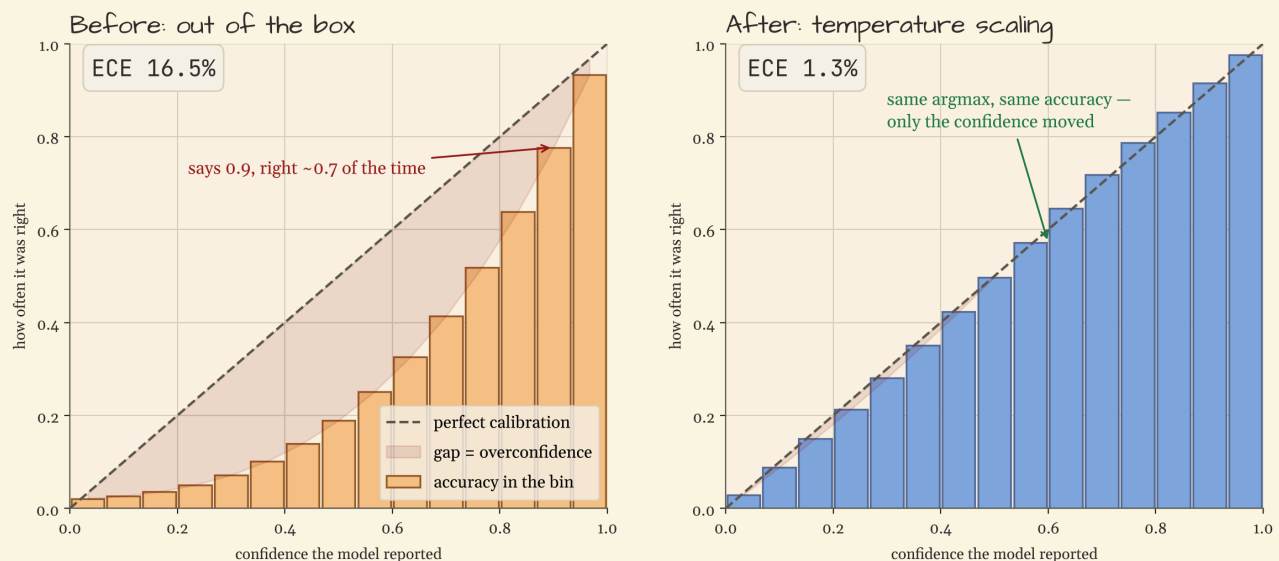
Headline accuracy is one number for the whole system, averaged over lopsided traffic. A rare, high-harm class can score near zero while the average looks fine. Autonomy is granted per class, so accuracy can’t tell you which decisions to hand over.

Worse, the training that buys accuracy spends something else. Guo and colleagues put a 1998 LeNet next to a 2016 ResNet: the old network was less accurate and its confidence matched its accuracy; the new one was more accurate and “substantially” surer than it had any right to be [2]. The dial that matters first is **calibration**: a model that says 0.9 should be right nine times in ten.

Temperature scaling fixes it with one scalar T fitted on held-out data: divide the logits by T and confidence starts meaning what it says. Their ResNet-110 went from 16.5% calibration error to 1.3%, and the fix “does not affect the model’s accuracy”: the argmax doesn’t move [2]. The number gets honest, which is what lets a threshold on it mean something.

Calibration: a model that says 0.9 should be right nine times in ten

One scalar T , fitted on held-out data, moves the bars onto the diagonal without changing a single decision.



Schematic bin shapes. ECE before/after from Guo et al., ICML 2017, Table 1 (ResNet-110 on CIFAR-100: 16.53% → 1.26%).

Semi and Fully Autonomous AI Systems, Part 3 | Steven Miers

One caveat that matters for anything that reads text: text classifiers calibrate less cleanly. On 20 Newsgroups the same fix only got from 8.0% to 4.1% [2]. Calibrate, then measure. Don’t assume.



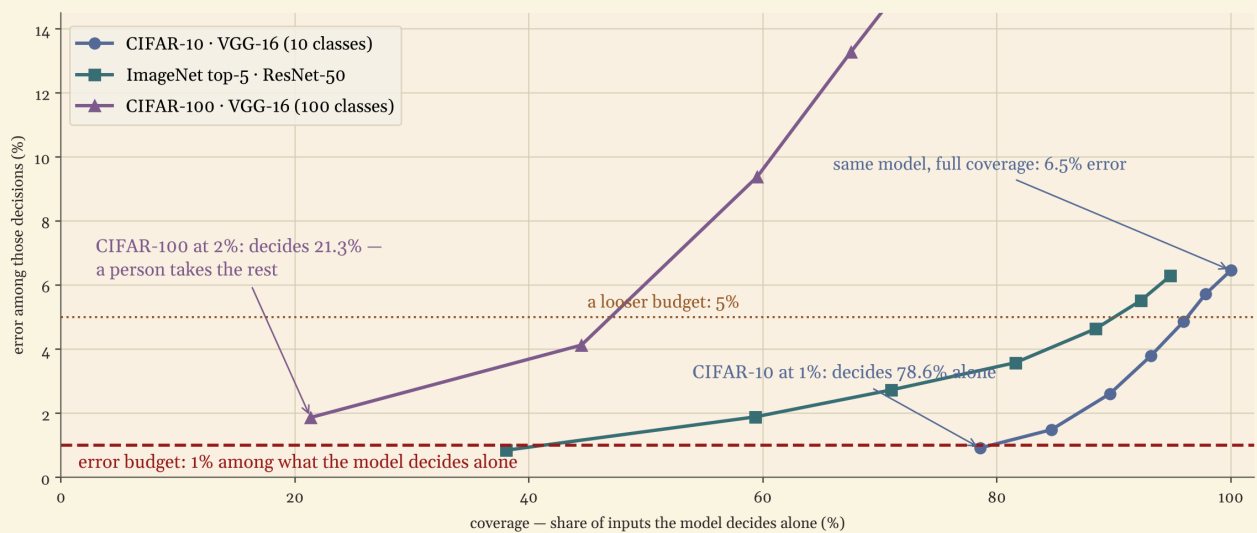
The abstention threshold is the autonomy dial

Chow priced the trade in 1957 and again in 1970: a classifier allowed to say *I don't know* buys lower error with lower coverage [1]. The reject option is older than almost everyone now rediscovering it. **Selective classification** brings it to deep networks [3]: pick a threshold τ on calibrated confidence. Above τ the model decides; below it, a person does.

Plot error among the decided items against the share decided and you get a **risk–coverage curve**. Draw the error budget across it, 1% here, and the curve reads off how much traffic the model may handle alone.

Risk-coverage: the dial that grants autonomy

Raise the threshold (τ) and the model decides less but errs less. The budget line reads off how much it may decide alone.



Test risk and coverage from Geifman & El-Yaniv, NeurIPS 2017, Tables 1, 2 and 6 (softmax response, SGR). Image benchmarks, shown for the mechanism.

Semi and Fully Autonomous AI Systems, Part 3 | Steven Miers

The published numbers make the point. A CIFAR-10 model that errs 6.5% at full coverage keeps error under 1% on the 78.6% of inputs it is surest about. Give the method 100 classes instead of 10, and at a 2% budget it decides 21.3% of the traffic; a person takes the rest [3].

Coverage at budget: one model, four decision classes

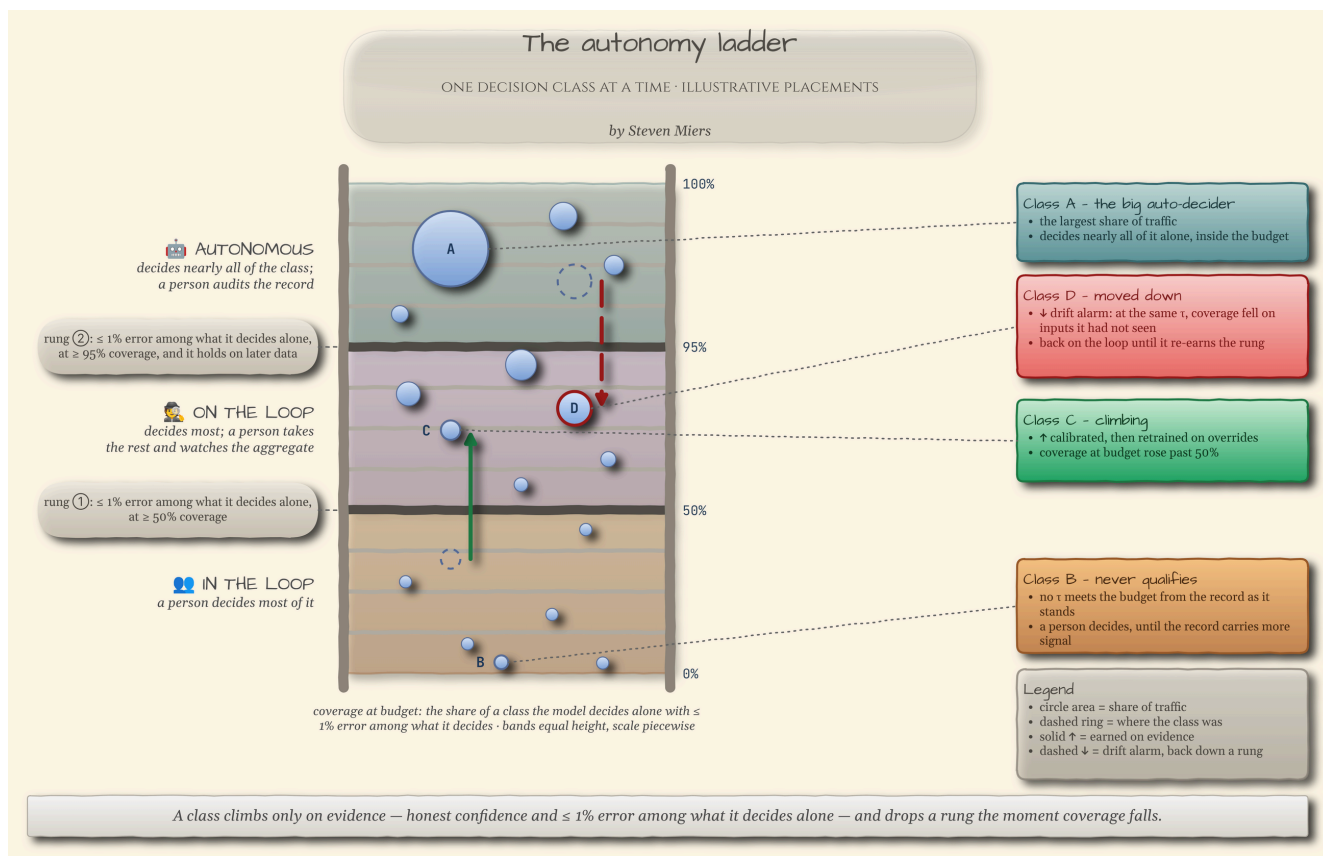
Softmax-response abstention on published benchmarks. Autonomy is what the model may decide alone inside the budget.

	Error budget	Error among what it decided	Coverage at budget	Verdict on the ladder
CIFAR-10, VGG-16 · 6.5% error at full coverage	1%	0.92%	78.6%	on the loop · decides most
same model, looser budget	5%	4.86%	96.0%	autonomous · decides nearly all
CIFAR-100, VGG-16 · 100 classes	2%	1.87%	21.3%	in the loop · a person decides most
same model, looser budget	10%	9.38%	59.5%	on the loop · decides most
ImageNet top-1, ResNet-50	2%	1.64%	25.9%	in the loop · a person decides most
ImageNet top-5, ResNet-50	1%	0.85%	38.1%	in the loop · a person decides most
same model, looser budget	5%	4.64%	88.5%	on the loop · decides most

Test risk and coverage from Geifman & El-Yaniv, NeurIPS 2017, Tables 1, 2, 5, 6. Rungs: $\geq 50\%$ coverage at budget $\rightarrow$ on the loop; $\geq 95\%$ $\rightarrow$ autonomous.

Semi and Fully Autonomous AI Systems, Part 3 | Steven Miers

Now do it **per class**. Each class gets its own τ , the lowest one that keeps it inside the budget, and its coverage at that τ is its evidence. That number puts it on a rung. The rungs are a proposal: a class decides alone at 95% coverage inside the budget, held on later data; it moves onto the loop at 50%; below that, a person decides.



Some classes never qualify. When the same free-text line turns up with different labels in the record, no model reading only that line can be right 99 times in 100 on it, and no τ will get it there. That isn't a tuning problem. It's a data problem, and the honest answer is a person, until the record carries more signal.

Rules are not weights

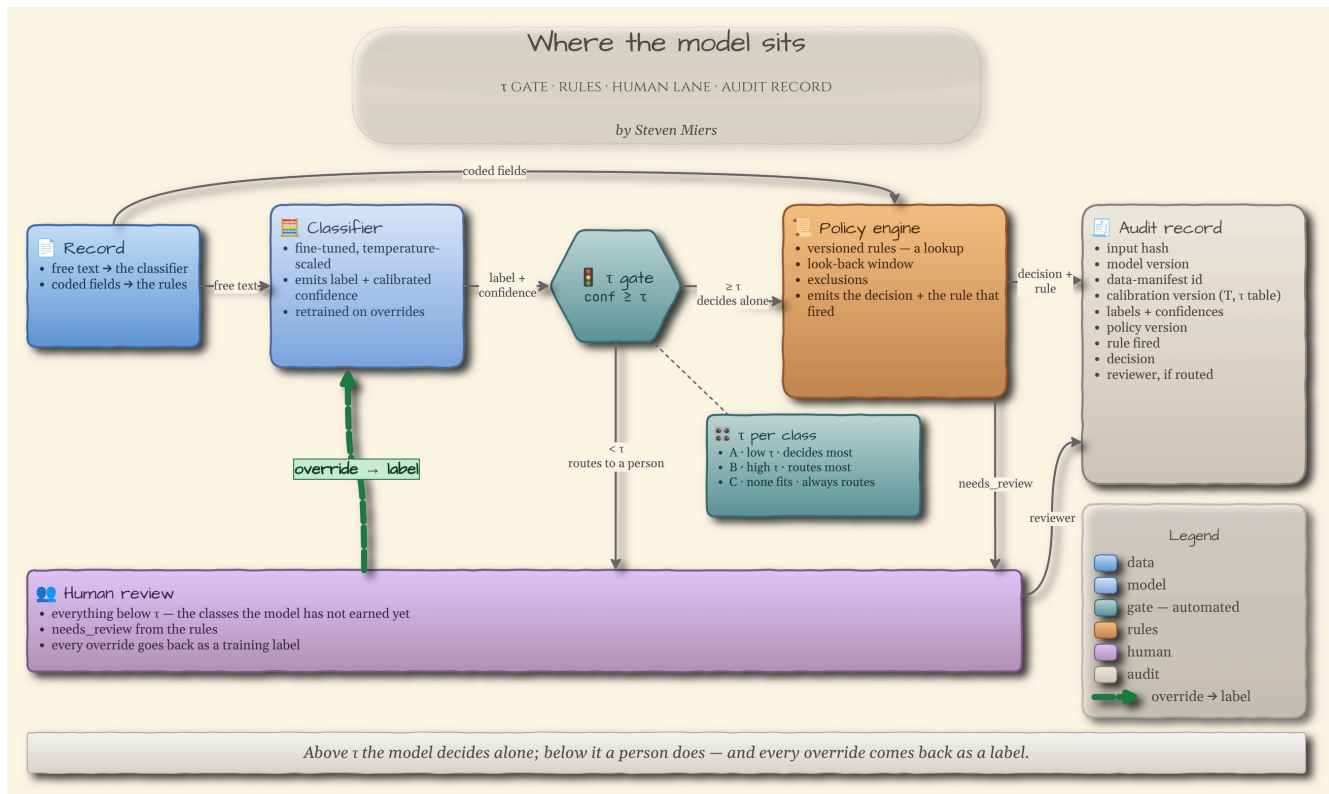
Not everything belongs in the model. What a category *means* for the decision (how far back to look, which outcomes are excluded, what gets reported) is a lookup a person writes, versions and signs. Rules change by statute and by client, not by gradient. **And whoever writes the letter is owed the rule that fired, not a logit.**

The law is written the same way. The FCRA requires “reasonable procedures to assure maximum possible accuracy” [4], and its reporting-period limits cap most adverse items at seven years while leaving convictions out of the limit: a look-back window and an exclusion, written by Congress. Employment decisions weigh the *Green* factors, offense, time since, and job, which courts have applied since 1975 and which outlive the EEOC’s 2012 guidance now headed for rescission [5]. The CFPB withdrew its advisory opinions on name-only matching and background screening in May 2025 [7]. *The guidance was withdrawn; the duty was not.* And when Medicare regulators were asked about algorithms in coverage decisions, they gave this section in one sentence: “artificial intelligence cannot be used to shift the coverage criteria over time” [10].



The audit record

Every decision emits one record: **input hash · model version · data-manifest id · calibration version (T and the τ table) · labels and confidences · policy version · rule fired · decision · reviewer, if it was routed**. Nine fields, none exotic; most systems log half of them in four different places. The EU’s AI Act asks for the same thing of high-risk systems: “automatic recording of events (logs) over the lifetime of the system” [11].



Three slots and a lane. The classifier reads the free text and emits a label with a calibrated confidence. The τ gate, one dial per class, passes what the class has earned and drops the rest to a person. The policy engine reads the coded fields through versioned rules and emits the decision with the rule that fired. And the loudest arrow on the page runs backward: a reviewer’s override comes home as a label. Part 1’s thesis, still working.

Back to the letter. Look up the input hash; read off the model, the confidence, the rule, the reviewer. Replay it. The thirty-day clock becomes an afternoon, and if the line was wrong, the correction goes back as a label like any other override.

The loop closes — and the alarm

Overrides return as labels. Retraining runs on a trigger, not a calendar, behind the automated evaluation Part 2 put in place of the governance meeting. And the drift alarm watches one number first: **coverage at a fixed τ** .

Coverage falls before accuracy does. Accuracy needs labels, and labels arrive late: after a review, a dispute, a quarter. Coverage needs nothing. When inputs turn unfamiliar, confidence drops below τ and more traffic routes to people. The queue grows before anyone can measure the error rate.

The drift alarm: what moves when the inputs change

BERT fine-tuned in one domain, tested in another. The temperature it needs rises with the shift — before anyone has labels.

Trained on → tested on	Accuracy, in → out	Calibration error, in → out	Temperature it needs, in → out	What a fixed τ sees
SNLI → MNLI (inference)	90.0 → 73.5	2.5 → 7.0	1.20 → 1.41	confidence drifts down, coverage falls a little
QQP → TwitterPPDB (paraphrase)	90.3 → 87.6	2.7 → 8.5	1.34 → 2.91	accuracy barely moves; confidence is what's wrong
SWAG → HellaSWAG (commonsense)	79.4 → 34.5	2.5 → 12.6	0.99 → 3.61	coverage drops; still not enough for a 34% model

Out of domain, temperature scaling fitted in-domain leaves 7–13% calibration error on the table. Coverage is the early alarm, not the whole alarm.

Desai & Durrett, EMNLP 2020, Tables 2 and 4 (BERT, out of the box; ECE in %). τ = the abstention threshold.

Semi and Fully Autonomous AI Systems, Part 3 | Steven Miers

The published record supports the alarm and bounds it. Fine-tune BERT on one dataset, test it on a shifted one, and the temperature it needs rises with the shift, from about 1.0 to as much as 3.6 [12]: the dial moving before anyone has labels. But a model calibrated at home stays too sure abroad. In a question-answering study, an answer given at confidence 0.6 was right 80% of the time in-domain and 45% out of it [13]. It abstained more. It did not abstain enough.

So the alarm is two-sided: coverage leaving its band at a fixed τ , *and* the inputs themselves changing, a vocabulary the model hasn't seen, a new jurisdiction's format. Either one moves a class back down a rung. That's not a failure. That's the system working.

Same shape, other domains

Prior authorization, benefits, content moderation: the ladder, the rules layer and the audit record carry over; only the taxonomy and the budget change. A Medicare Advantage insurer is defending a suit alleging its coverage algorithm's denials were reversed about 90% of the time on appeal; the contract claims survived because the insurer's own documents say clinicians decide [14]. The Dutch government fell in January 2021 over a benefits-fraud model whose "use, workings and effect" were hidden from the public [15]: the shrug, at national scale. Meta's Oversight Board, reviewing a handful of the hundreds of thousands of appeals it receives, overturned the company 90% of the time in 2023 [16]. Every one of those is the letter by another name.

Key Takeaways

- **Autonomy is granted per decision class, on evidence — and revoked the same way.** Nothing starts on the top rung.
- **The evidence is calibrated confidence and error among the decisions the model takes alone, under a budget.** Headline accuracy can't see it.
- **τ is the autonomy dial.** Coverage at budget places a class on the ladder; a class whose text carries mixed labels never qualifies, and that's a data problem.
- **Rules are not weights.** Versioned, signed, outside the model, so the disputant gets the rule that fired.
- **Coverage at a fixed τ is the early alarm, not the whole alarm.** It falls before accuracy can be measured, and not by enough, so watch the inputs too.

This is **Part 3** of a series: [Part 1: HITL](#) · [Part 2: HOTL](#).

References

1. Chow, C. K. "An Optimum Character Recognition System Using Decision Functions." *IRE Transactions on Electronic Computers* EC-6(4), 1957; and "On Optimum Recognition Error and Reject Tradeoff." *IEEE Transactions on Information Theory* 16(1), 1970.
2. Guo, Chuan, et al. "On Calibration of Modern Neural Networks." *ICML 2017*. Figure 1; Table 1; Section 4.2.
3. Geifman, Yonatan, and Ran El-Yaniv. "Selective Classification for Deep Neural Networks." *NeurIPS 2017*. Tables 1, 2, 5 and 6.
4. Fair Credit Reporting Act: 15 U.S.C. § 1681e(b) (accuracy); § 1681c(a), FCRA § 605(a) (reporting-period limits); § 1681i(a) (30-day reinvestigation; description of the procedure on request).

5. *Green v. Missouri Pacific Railroad Co.*, 523 F.2d 1290 (8th Cir. 1975); U.S. Equal Employment Opportunity Commission, *Enforcement Guidance on the Consideration of Arrest and Conviction Records in Employment Decisions under Title VII*, No. 915.002, 25 April 2012 (draft rescission under OMB review as of October 2026).
6. Federal Trade Commission. “Texas Company Will Pay \$3 Million to Settle FTC Charges That It Failed to Meet Accuracy Requirements for Its Tenant Screening Reports,” 16 October 2018 (RealPage, Inc.).
7. Consumer Financial Protection Bureau. “Interpretive Rules, Policy Statements, and Advisory Opinions; Withdrawal.” 90 Fed. Reg. 20084, 12 May 2025.
8. Miers, Steven. *Semi and Fully Autonomous AI Systems*, [Part 1: HITL](#) and [Part 2: HOTL](#), 2026.
9. Miers, Steven. [Autonomy is a spectrum, not a switch](#), 2026.
10. Centers for Medicare & Medicaid Services. *Frequently Asked Questions related to Coverage Criteria and Utilization Management Requirements in CMS Final Rule (CMS-4201-F)*, 6 February 2024.
11. Regulation (EU) 2024/1689 (the AI Act), Article 12(1) (record-keeping) and Article 14 (human oversight).
12. Desai, Shrey, and Greg Durrett. “Calibration of Pre-trained Transformers.” *EMNLP 2020*. Tables 2 and 4.
13. Kamath, Amita, Robin Jia, and Percy Liang. “Selective Question Answering under Domain Shift.” *ACL 2020*. Section 5.3.
14. *Estate of Lokken v. UnitedHealth Group*, No. 0:23-cv-03514 (D. Minn.): complaint filed 14 November 2023 (the 90% figure is the plaintiffs’ allegation); order on the motion to dismiss, 13 February 2025.
15. Amnesty International. *Xenophobic Machines: Discrimination through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal*, EUR 35/4686/2021, 25 October 2021.
16. Oversight Board. *2023 Annual Report*, 2024 (53 decisions; 398,597 appeals received).

Steven Miers — git-steven.github.io