

No Stable Self: What Self-Referential Anomaly Defence Can Certify for a Continually Learning System, Why Every Located Guarantee Holds the Benign Distribution Still, and Why Immunology's Own Answer Is a Damage Signal From Outside

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Immune-inspired security proposes that a system can protect itself by learning what it normally is and treating departures from that as threats. Continual learning makes departure the normal case: the system is supposed to change, so a poisoning campaign and a legitimate shift in the data both reach the defender as an unexplained change. This paper asks what a defence that takes its reference from the system's own history can certify in that setting, and finds the answer narrower than either side of the debate states. Three readings of "self" are separated: a frozen reference, a tolerised self that absorbs whatever persists, and a damage-keyed rule that does not need a self at all. The stationarity objection defeats only the first. The second is what many working continual defences use, and it fails differently: retrained anomaly detectors and network coordinate systems have been shown to tolerate change that is slow enough, which is also how immunology's own discontinuity theory describes the biological immune system. That continual-learning and test-time-adaptation defences fail the same way is a prediction carried over from those results, not yet tested. Every positive guarantee located here, from online centroid detectors to Byzantine aggregation to a 2026 theory of poisoning in continual learning, either holds the benign distribution fixed, bounds how far it may move, or assumes attacks are rare; none survives an unbounded legitimate shift together with a persistent minority attacker. The paper argues that the transferable lesson of immunology is the danger model rather than self/non-self discrimination, and that in a learning system a damage signal has to come from supervision held outside the update loop: trusted data, provenance, labels or a restoration point. Where such an anchor has been costed, it was cheap: one study priced its renewal at 50 analyst labels a month, and another gives a static size of 100 clean examples. Whether it can be kept fresh as legitimate data moves, and at what cost, is the open part.

Keywords: continual learning; data poisoning; artificial immune systems; anomaly detection; concept drift; danger theory; Byzantine robustness; test-time adaptation

1. Introduction

The immune system is the oldest metaphor in computer security for a defender that learns. Forrest et al. (1994) cast the problem of protecting a computer as the problem of learning to tell self from other, and built a change detector modelled on the generation of T cells. Two years later the same group defined normal behaviour for a Unix process by short-range correlations in its system calls, and reported in early experiments that this definition stayed stable while the program behaved normally (Forrest et al., 1996). The appeal is plain. A detector that models only the protected system needs no catalogue of attacks, so it can in principle catch attacks nobody has seen.

That appeal is now being transferred to machine learning systems that keep learning after deployment: continual learners, models adapted at test time, language models fine-tuned on user data, and federated models updated from many clients. The security literature on these systems has grown quickly and is mostly a literature of attacks. Poisoning that induces forgetting in a continual learner needs only the current model and the current task's data (Abbasi et al., 2023). A backdoor planted in one task of a sequence survives the learning of later tasks

(Guo et al., 2025). A handful of malicious samples in a test batch can redirect predictions on the benign samples beside them when the model adapts to that batch (Wu et al., 2023). A small fraction of poisoned pretraining text can survive post-training (Zhang et al., 2024).

The defence that suggests itself is the immune one: learn what the model, its updates and its data normally look like, and refuse what departs from that. The objection that motivated this paper is equally plain. Self/non-self discrimination needs a self, and continual learning removes a stable one by design. A learner that is meant to follow a changing world receives a legitimate shift and a poisoning campaign as the same kind of event: data it does not yet explain. If the detector's reference is the system's own past, then the detector's central primitive - departure from normal - does not separate the two cases it most needs to separate.

This paper tests that objection against the published record and finds that it is right about its conclusion and wrong about its reason. It is right that no defence whose reference comes only from the system's own history can certify a continually learning system against a patient attacker; Section 8 shows that every located guarantee buys its result by holding the benign distribution still,

bounding its movement, or assuming attacks are rare. It is wrong that the trouble is the absence of a stationary self. Most of the defences that work in practice, from the original immune-inspired intrusion detectors to the rolling-window rules and batch-adapted statistics of recent continual and test-time learners, do not assume a stationary self. They assume a self that is updated by tolerance - whatever persists long enough becomes normal. For online centroid detectors and network coordinate systems the resulting failure has been shown: a slow enough attack is tolerated (Kloft and Laskov, 2010; Chan-Tin et al., 2011). That continual-learning and test-time-adaptation defences fail the same way is this paper's prediction from those results, untested for them (Section 13), and not every such defence tolerises: Pawlak et al. (2025) fix their threshold once, from a clean task at the start of the sequence. Immunology describes the biological immune system in exactly these terms. Its discontinuity theory holds that immune responses are triggered by sudden change and that slow or continuous stimulation induces tolerance (Pradeu et al., 2013; Pradeu and Vivier, 2016). The analogy transfers, failure mode included.

What the analogy does not transfer, and what its immunological critics proposed in its place, is the danger model: respond to damage, not to foreignness (Matzinger, 1994; Matzinger, 2002). Artificial immune systems research took that turn two decades ago (Aickelin and Cayzer, 2002; Greensmith et al., 2010) and found that a damage signal has to be supplied from outside the detector. This paper argues that in a learning system the same holds: the damage signal is supervision that the update loop cannot rewrite.

The contributions are four, each stated at the evidence level it has. The first is a separation of three things "self" can mean in a learning defence: a frozen reference, a tolerated self, and no self at all, with a damage signal in its place. The stationarity objection applies only to the first of these (Section 3). The second is a reading of the immunology and artificial immune systems record: prominent immunological critics (Matzinger; Pradeu and colleagues) challenged self/non-self as the criterion of immunogenicity, the second-generation artificial immune systems literature turned to danger signals, and the tolerated replacement carries a slow-change vulnerability on both sides of the analogy (Sections 4 and 5). Two premises of the objection as usually stated are corrected along the way: the exponential cost of negative selection lies in the matching length rather than in dimension as such, and modern robust statistics removes the dimension penalty when the distribution is fixed (Section 5.3). The third is a sorting of the positive guarantees located in the poisoning, online anomaly detection, Byzantine learning and continual-learning literatures by what each assumes about benign data (Table 2, Section 8); the pattern is uniform, since the benign side is fixed, bounded, or outnumbered. The

fourth is a statement of what the danger model asks for once translated, namely an anchor outside the update loop, and of the cost that only one located study has priced, which is keeping that anchor current while legitimate data move (Sections 9 and 11).

The paper builds on four of the author's earlier analyses and should be read as extending rather than repeating them. Mahendrakar (2026a) argued that without labels a drifted class and a new class can generate the same stream; the present paper concerns a different confusion, between legitimate drift and adversarial drift, and its central object is the defender's reference rather than the stream. Mahendrakar (2026b) argued that a record-level detector cannot recover intent in an agent's memory store; the present paper moves from records to weight updates and asks what the update loop itself can certify. Mahendrakar (2026c) examined what the near-constant poisoning result does and does not establish about detection; it is used here only in Section 7.4. The recovery half of the immune proposal - repairing the model once poisoning is detected - meets the suppression-versus-removal problem treated in Mahendrakar (2026d), and Section 10 keeps that part short for that reason. The author's not yet published analysis of acceptance tests in self-modifying systems reaches the conclusion of Sections 9.1 and 9.3 for a different object - an external anchor shrinks but does not disappear, and moving it outside moves the attack surface with it - so Section 9 applies that conclusion to poisoning defences rather than introducing it.

Section 2 states the scope. Section 3 separates the three readings of self. Section 4 reads the analogy against immunology, and Section 5 reads it against the artificial immune systems record. Section 6 states the identifiability problem in the forms the literature has proved it. Section 7 collects the attacks on continual and adaptive learners and what their proposed defences assume (Table 1, Figure 1). Section 8 sorts the guarantees (Table 2). Section 9 translates the danger model and states the decision structure (Algorithm 1). Section 10 treats recovery. Section 11 proposes the evaluations that would settle the open part. Sections 12 and 13 put the strongest case against the paper and state what is not known.

2. Scope, and what this paper does not claim

The paper is an analysis of published work. It consolidates the published measurements, sorts the located guarantees by what they assume, states the decision structure as an algorithm and proposes the evaluations that would settle the open part; every number in it is a number some cited paper reports, attributed on the spot.

Its object is the class of defences that decide whether to accept an update, a sample, a task or a client contribution by comparing it with a reference derived

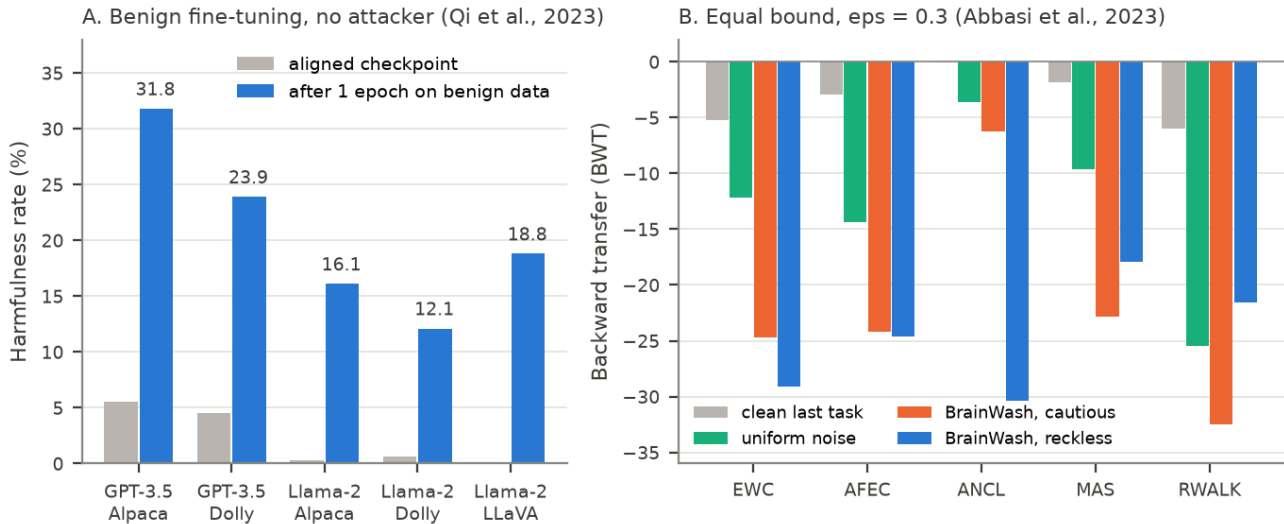


Figure 1. Panel A: harmfulness rate of aligned chat models before and after one epoch of fine-tuning on benign instruction datasets, from Qi et al. (2023), Table 3 (GPT-3.5 Turbo was not tested on LLaVA-Instruct). Panel B: backward transfer on CIFAR-100 after the last task is learned clean, with uniform noise, or with the two BrainWash variants, all at L-infinity bound 0.3, from Abbasi et al. (2023), Table 1; more negative means more forgetting. Panel A shows benign data producing an attack-like effect; panel B shows inputs of identical perturbation size producing different and overlapping damage. Each panel re-plots values reported by the cited sources.

from the system's own data or parameter history - "self-referential" defences. That class includes immune-inspired detectors, drift detectors used as gates, outlier-based sanitisation with statistics estimated from incoming data, median-type aggregation in federated learning, and the detection rules inside recent continual-learning defences. It excludes defences whose reference is held outside the loop (a trusted dataset, provenance metadata, human labels, a cryptographic commitment), except as the alternative the paper argues for.

The paper does not claim that self-referential defences are useless. Several work well under conditions the paper names, and Section 12 argues those conditions are often realistic. It does not claim that poisoning a continual learner is easy in practice; Su et al. (2025) show that for test-time adaptation the risk was overstated under unrealistic threat models, and that result is carried as a counterweight. It does not claim that supervision anchors are sufficient; Section 9.3 states how they fail.

"Continual learning" is used broadly, to cover task-incremental and class-incremental learning, task-free streaming learning, test-time adaptation and repeated fine-tuning of deployed models. The arguments that depend on a particular setting say which.

3. Three readings of "self" in a learning defence

The word "self" in an immune-inspired proposal can name three different things, and most of the confusion in the debate comes from running them together. The definitions below are this paper's.

Reading 1: a frozen reference. Self is a fixed sample of normal behaviour, collected once. Detectors are

generated against it and then used unchanged. This is negative selection in its original form: candidate detectors that match any self string are discarded during a maturation phase, and the survivors flag anything they match later (Forrest et al., 1994). The reference is the training sample. As Aickelin and Cayzer (2002) put it from the immunological side, self in practice is limited to what the detectors saw while they matured.

Reading 2: a tolerised self. Self is whatever the system has recently observed and has not been told is harmful. The reference moves, by a rule that absorbs persistent observations into normal. In immune-inspired intrusion detection this took the form of detector lifespans, tolerisation periods and rolling coverage (Hofmeyr and Forrest, 2000; Kim et al., 2007). In modern continual learning it takes the form of rolling-window statistics on updates (Hu and Duan, 2026), consolidation triggered when the loss plateaus (Aljundi et al., 2019), and test-time adaptation, which re-estimates normalisation statistics from each incoming batch (Wu et al., 2023).

Reading 3: no self, a damage signal instead. The decision to respond is keyed to evidence of harm rather than to foreignness. In immunology this is the danger model (Matzinger, 1994). In artificial immune systems it is the dendritic cell algorithm, which combines signals that its designer has designated in advance as indicating anomaly, damage or safety (Greensmith et al., 2010). In a learning system it is any rule that asks whether an update made the model worse at something that matters, measured on a reference the update cannot touch.

The stationarity objection - continual learning has no stable self, so self/non-self discrimination is undefined - is decisive against Reading 1 and only Reading 1. A

frozen reference in a system designed to change will flag every legitimate shift, and a defender who responds by widening the reference has moved to Reading 2. Pawlak et al. (2025) come closest to Reading 1, fixing a detection threshold once from a clean task at the start of the sequence, but under a threat model with a single poisoned task; no located work defends a frozen reference for a learner whose legitimate data move without bound. The live question is therefore not whether a stationary self exists, but what Reading 2 can certify, and whether Reading 3 is available without importing exactly the supervision immune-inspired methods were meant to do without. Sections 6 to 8 answer the first question and Section 9 the second.

4. The analogy, read against immunology

4.1 Immunology's critics challenged self/non-self first

The objection is sometimes put as if self/non-self discrimination were an unchallenged account of biological immunity, so that its failure in machine learning is a failure of transfer alone. It is still the majority view - Pradeu and Carosella (2006) call it "the consensus view", and Matzinger (2002) says immunologists have built on it "for over 50 years" - but prominent immunological critics have challenged it. Matzinger (1994) proposed that the immune system "does not care about self and non-self", that what drives it is guarding the body against danger, and that it acts on signals from the surrounding tissues rather than on its own judgement alone. She restated the case in 2002, arguing that decades of close study had exposed problems inherent in the self/non-self paradigm, and that harm, not foreignness, is what the immune system chiefly attends to (Matzinger, 2002). Pradeu and Carosella (2006) proposed replacing the self/non-self criterion of immunogenicity with a criterion of continuity, concluding that immunologists can keep talking about the self only if they accept that self/non-self discrimination does not decide when an immune response occurs.

Aickelin and Cayzer (2002), writing for artificial immune systems practitioners, listed the anomalies that motivated the shift: useful autoreactivity against stressed cells, the fact that self in practice is confined to what lymphocytes saw during maturation, and the fact that the body changes over a lifetime, so that defences learned early "might be autoreactive later". The last point is the stationarity objection, stated about the organism two decades before it was stated about continual learners.

4.2 Immunology's replacement has a slow-change vulnerability

The more recent theory makes the analogy sharper rather than weaker. Pradeu, Jaeger and Vivier (2013) proposed a discontinuity theory. Innate cells such as macrophages and natural killer cells react to short-lived

alterations of the tissue around them and adjust to alterations that persist; on this view an effector response is set off by an abrupt change in the molecular patterns the immune cells meet. Pradeu and Vivier (2016) put the principle in one line: abrupt shifts in antigenic input provoke a response, while gradual or sustained input produces tolerance. They point to data on immune checkpoints in viral infections, cancers and allergies as support.

Two consequences follow for the transfer. First, the biological immune system is a Reading 2 system: it maintains a tolerised self, and what counts as normal is whatever has persisted. Second, it carries the corresponding vulnerability. If tolerance follows slow change, then a threat that changes slowly enough is tolerated, and the essays name viral infection and cancer among the settings where their checkpoint evidence applies. The paper does not claim more than the abstracts of these essays state; it draws only the structural inference that a rule keyed to the rate of change can be evaded by keeping the rate low. That inference is the frog-boiling attack of the security literature (Section 6.3), and it is the property of Reading 2 that Sections 6 to 8 find in every learning defence of that kind.

4.3 What the biological answer actually is

The danger model does not say that damage detection is easy. It says that the signal to respond comes from tissue under stress, not from the immune cell's comparison of an antigen with its record of self. In the organism that signal is produced by cells whose damage is the thing to be prevented. The machine analogue is a measurement of the harm an update does, taken on something the update did not shape. Section 9 argues that in a learning system this is supervision held outside the update loop, and that its cost has gone almost entirely unreported.

5. What the artificial immune systems record shows

5.1 The first generation and its scaling problem

The negative selection algorithm worked on small problems. Kim et al. (2007), in the field's main review of intrusion detection, report that Forrest's original experiment reached a 100 percent detection rate with 125 detectors on a file encoded as 655 binary strings of 32 bits, and call this a relatively small-scale problem. The same review summarises Kim and Bentley's attempt on network traffic: non-self detection rates below 16 percent, an estimate that 80 percent detection would need about 6×10^8 detectors and 1,429 years to generate from twenty minutes of data, and the authors' conclusion that scaling issues made the algorithm perform poorly on real problems. The review also reports that every proposed variant of detector generation with r-chunk matching has runtime exponential in the

matching length r , and that Stibor and colleagues showed the generated detector set underfits for small r , pushing r towards the string length.

Real-valued variants did not escape. Kim et al. (2007) report a comparison by Stibor and colleagues in which real-valued negative selection with variable-sized detectors reached at most 2.6 percent detection on a high-dimensional problem, while a one-class support vector machine reached high detection at acceptable cost. In the GECCO paper that asked whether negative selection suits anomaly detection at all, Stibor et al. (2005) reported that real-valued negative selection, and a simpler classifier built from self detectors, reached high accuracy only when trained on examples of both normal and anomalous data, while a one-class support vector machine managed with normal data alone. That finding matters more for this paper than the cost figures. The method whose rationale was learning from self alone needed examples of non-self to work, which is to say it needed supervision.

The review is also explicit about the conceptual limit. It reports the view of other researchers that the founding assumption of negative selection - that foreign patterns are intrusions - is "possibly the most controversial problem" of the method: a pattern can be unfamiliar without being hostile, so many false alarms may be unavoidable (Kim et al., 2007). Freitas and Timmis (2007), reviewing artificial immune systems for data mining, argued for problem-specific design over generic immune algorithms and listed seven limitations of the field as it stood.

The record is not uniformly negative. Hofmeyr and Forrest (2000) report that their distributed system, LISYS, detected intrusions while keeping false positives low, and a later review by Gupta and Dasgupta (2021) argues that negative selection has improved over the following decade and outperforms other methods in some settings. Neither result concerns a reference that must keep moving while an adversary tries to move it.

5.2 The second generation moved the supervision rather than removing it

LISYS already used a tolerised self. Kim et al. (2007) describe its detectors as matured against self over a tolerisation period, given finite lifespans, promoted to memory after confirmed alarms, and costimulated by a user who confirms that an alert is genuine; they report an analysis in which rolling coverage handled concept drift. Two features matter here. The self moved, by replacement of detectors over time, so LISYS was a Reading 2 system. And the loop that decided which alarms were real ran through a human.

Aickelin and Cayzer (2002) made the human dependency the motivation for danger theory. They observed that immune-inspired detectors respond only after infection passes a threshold and a human confirms

it, that deciding what is self and what is not is frequently unclear and may take an expert, and that it is "inefficient, if not impossible, to map the entire non-self universe", since that universe changes over time. Their proposal was to ground responses in danger signals instead.

The dendritic cell algorithm implemented that proposal, and its own descriptions show where the supervision went. Greensmith et al. (2010) define its inputs as signals fixed before deployment: a pathogen-associated signal whose increase is "a definite indicator of anomaly", a danger signal that probably indicates damage, and a safe signal that probably indicates normality. They state that while the algorithm "removes the need to define self", it still requires choosing an antigen representation and pre-classifying the input signals, which are heuristic approximations of normal and anomalous. The designer who maps system measurements onto these signals supplies the knowledge of what damage looks like. Stibor et al. (2009) then rewrote the algorithm's signal processing as dot products and showed that it amounts to a set of linear classifiers whose boundaries are all parallel, which led them to judge the method of little use on harder problems.

Immune-inspired defences in modern deep learning have mostly targeted a different threat. RAILS, for example, uses an immune-inspired evolutionary process at inference time to resist adversarial examples on fixed benchmarks, reporting, among its results, a robustness improvement of at least 5.62 percent on MNIST under PGD (Wang et al., 2021). No located immune-inspired method addresses a learner whose reference moves while an adversary feeds it.

5.3 Two premises of the objection that need correcting

The objection as usually put has two further premises, and both need narrowing.

The first is that negative selection needs detector counts exponential in feature dimension, and that neural representations are the worst case. The exponent the artificial immune systems literature documents is in the matching length r of string-based detectors (Kim et al., 2007); the real-valued failure Stibor reports is an empirical failure on a high-dimensional problem rather than a proved exponential law. Dimension does matter for poisoning: Steinhardt et al. (2017) attribute the vulnerability of the IMDB sentiment data to its high dimension and its many irrelevant features, and Wang, Jha and Chaudhuri (2020) find that poisoning defences in online learning work for low-dimensional, well-separated data and fail more easily for high-dimensional, low-margin data. Kloft and Laskov (2010) find that the slope of an optimal attack on an online centroid detector grows with input dimension, logarithmically, levelling off once dimension exceeds the number of training points.

But high dimension is not an impossibility in itself. Diakonikolas et al. (2016) give computationally efficient estimators whose error under an adversarially corrupted fraction of samples does not depend on dimension, for Gaussians, product distributions and some mixtures, and Tran et al. (2018) use such robust statistics to find backdoor examples through their spectral signature. What those results require is a fixed distribution from which the clean samples come. The binding obstacle for a continual learner is therefore the one this paper concentrates on - movement of the benign distribution - and the dimension argument should be dropped from the case.

The second premise is that legitimate shift and poisoning "present the identical signal". In the limit, and in particular instances, they do (Section 6). In general they do not: Paudice et al. (2018) found that optimal poisoning points generated without a detectability constraint lay well away from genuine data and could be removed by outlier detection, and Pawlak et al. (2025) detect most single-task poisoning with a threshold set on one clean task. The accurate claim is that separation always rests on an assumption about the benign data or the attacker that the stream itself cannot confirm. Section 8 makes that precise.

6. The identifiability problem, in the forms it has been proved

6.1 Continual learning

Hu and Duan (2026) give the most direct statement. Modelling attack and defence in regularisation-based continual learning as an online zero-sum game, they prove that there exist problem instances in which, under frequent attacks that shift the data pattern, no regularisation-based method can be guaranteed to converge, and they name the reason: widespread distribution-shifting attacks create "an identifiability crisis, where malicious perturbations become statistically indistinguishable from legitimate task shifts". Their construction is deliberately simple - a one-dimensional continual linear regression with a uniform task distribution. The authors point out that a convergence failure in so easy a setting also rules out convergence in the general, non-uniform case.

The same paper states the attribution problem that confronts any self-referential detector after a large update. A large distance between consecutive models may come from a benign task correcting a model that earlier attacks corrupted, or from a malicious task pulling a benign trajectory away; the authors write that the system "cannot determine" which.

6.2 Federated and distributed learning

The federated literature proved the spatial version of the same problem earlier. Karimireddy et al. (2022) observe that a worker whose update looks unusual may be

malicious or may simply hold unrepresentative data, that the update should be ignored in the first case and is valuable in the second, and that the server cannot tell the two apart. They formalise this as an information-theoretic lower bound: with a fraction δ of Byzantine workers and heterogeneity ζ squared among the good ones, there are problems on which any algorithm suffers error at least of order δ times ζ squared (divided by the strong-convexity constant for the function-value form), however long it runs. Their mimic attack, in which Byzantine workers copy one chosen good worker, cannot be filtered because it is indistinguishable from that worker. Allouah et al. (2023) open from the same point - under heterogeneous data a Byzantine machine can hardly be told apart from an honest outlier - and give aggregation that achieves the optimal error under heterogeneity, not zero error. Baruch et al. (2019) showed that small changes crafted to stay within the spread of honest updates evade the existing defences, with 20 percent of workers sufficient to cut CIFAR-10 accuracy by half.

The correspondence with continual learning is direct. Heterogeneity across clients at one time plays the role that legitimate drift plays across one learner over time. A defence that cannot separate an attacker from an honest outlier in space has no better prospect of separating an attacker from an honest change in time.

6.3 Adaptive anomaly detection and the frog-boiling attack

The security literature on retrained anomaly detectors reached the same conclusion before deep learning. Barreno et al. (2010) note that the ability to follow a changing situation, which recommends learning to security engineers, is exactly what an attacker exploits. Rubinstein et al. (2009) made the point for network intrusion detection: once normal is learned from traffic and attacks are flagged as departures from it, the attacker is invited to tamper with the traffic the model learns normal from, so that later attacks fall inside it.

Kloft and Laskov (2010) analysed this precisely for an online centroid detector, and their result is the clearest formal statement of what Reading 2 costs. When the detector keeps all past data, an attacker needs a number of points exponential in the desired displacement. When it keeps a finite window and drops old points - when it adapts - the displacement grows linearly in the number of injected points, which the authors judge cannot be considered secure; removing the nearest neighbour of each incoming point does not help. Adaptation is what turns an exponential cost into a linear one.

Chan-Tin et al. (2011) named the strategy. Their frog-boiling attack defeated outlier-detection defences for network coordinate systems, including one that required part of the network to be trusted and one based on reputation. Their conclusion was that filtering inputs statistically, or by reputation, is not enough to make a

network coordinate system secure. Dey et al. (2020) describe the same threat for continually retrained behavioural detectors in enterprise security: such a detector must be retrained to follow drift, and that opens the way to an adversary who distils attack traces into the model gradually until the full attack looks normal. Their mitigation brings an analyst into the loop to annotate alerts.

Korycki and Krawczyk (2020) carried the question into stream mining and state the stationarity objection in their own words: what could be considered adversarial "may become a valid instance from a drifted distribution", so a defender cannot know in advance whether a change comes from drift or from attack. They report that existing drift detectors cannot differentiate real from adversarial drift, and argue that poisoned instances arranged into a coherent concept of their own will pass outlier, noise and novelty filters alike.

6.4 The converse direction: legitimate updates that look like attacks

The confusion runs both ways, and the second direction is the one a defender pays for in false alarms. Qi et al. (2023) fine-tuned aligned chat models for one epoch on widely used benign instruction datasets and observed safety degrade in every case they tested: GPT-3.5 Turbo's harmfulness rate rose from 5.5 to 31.8 percent after Alpaca and from 4.5 to 23.9 percent after Dolly, and Llama-2-7b-Chat's from 0.3 to 16.1 percent after Alpaca (Figure 1, panel A). He et al. (2024) went further: selecting the 100 benign examples most similar to known harmful data raised the GPT-judged attack success rate after fine-tuning to 71 percent on Alpaca, against 13 percent for a random 100, and the selected examples were mostly lists, bullet points and mathematics.

These are updates with no attacker. Judged by effect, they resemble a safety poisoning attack; judged by content, they are ordinary data. A self-referential defence that gates on effect would block legitimate customisation, and one that gates on content would admit the harmful subset. He et al. find the harmful subset only by comparing candidate data with a set of known harmful examples, which is a reference held outside the update.

7. Attacks on continual and adaptive learners, and what their defences assume

7.1 Task-sequential continual learning

Table 1 collects the measured values the argument rests on. Each row is one result, with its source named on the row.

The first continual-learning poisoning results showed that a continual learner's main weakness - it cannot revisit old data - is also a security weakness. Han et al. (2023) showed that regularisation-based learners, which

keep no previous data and so cannot track performance on earlier tasks, can be made to lose a chosen earlier task through poisoned data for the new one. Li and Ditzler (2023) showed that label-flipping and a clean-label variant they call PACOL cause forgetting in generative-replay and regularisation-based learners, and tested sanitisation defences against them: an L2-distance defence on deep features, using a feature extractor trained on the first task and the true labels, removed nearly all label-flipped points in a simple subset but failed against the imperceptibly perturbed ones. A one-class SVM on deep features removed about half of the stealthy samples in their CIFAR subset; isolation forest and local outlier factor did not cope with the hard cases.

BrainWash (Abbasi et al., 2023) needs only the current model and the current task's data, and it is the source of panel B of Figure 1. On CIFAR-100 with EWC, backward transfer was -5.2 when the last task was learned clean, -12.2 with uniform noise of L-infinity norm 0.3, and -29.1 with the reckless BrainWash variant at the same bound. Figure 1 shows a detail worth keeping: the attack is not always worse than noise of the same size. For RWALK, uniform noise gave -25.5 against -21.6 for reckless BrainWash, though the cautious variant gave -32.5. Damage is therefore not itself a fingerprint of malice, and a perturbation-size gate sees the same thing in every column.

Guo et al. (2025) showed persistence. Their latent task backdoor touches the training of a single task and leaves all other tasks benign, yet reaches attack success rates above 90 percent in most configurations, with a mean of 90.6 percent against regularisation-based and 90.7 percent against replay-based learners for their CNN architecture. With components added to evade SentiNet and I-BAU, the success rate fell by only 0.56 and 2.2 percent in relative terms under those defences, from 0.9 to 0.895 and 0.88. The point for this paper is the direction of the persistence. A continual learner's regularisers exist to protect what was learned earlier; once a backdoor has been learned, the same machinery protects it.

That point generalises to task-free learning. Aljundi et al. (2019) decide when to consolidate - when to update the importance weights that protect current knowledge - by detecting plateaus in the training loss: a stable low loss signals a good time to protect what was learned. The rule is self-referential in the sense of Section 3. Speculatively, a poisoner who can produce a stable plateau on poisoned data would have that knowledge consolidated and protected by the learner's own mechanism; no located study tests this.

7.2 What the continual-learning defences assume

Three recent defences make their assumptions explicit, and each depends on a reference the stream cannot supply.

Table 1. Published measurements on poisoning and adaptation in continual, adaptive and federated learners. Values are as reported by the source named on each row; "reading" is this paper's interpretation.

Source	Setting	Reported value	Reading
Abbasi et al. 2023	CIFAR-100, EWC, eps 0.3	BWT -5.2 clean, -12.2 uniform noise, -29.1 BrainWash	Attack and noise share a bound, not an effect
Guo et al. 2025	Latent task backdoor, CNN	Mean ASR 90.6% (regularisation), 90.7% (replay)	One poisoned task persists
Guo et al. 2025	LwF under SentiNet / I-BAU	ASR drop 0.56% / 2.2% (relative; 0.9 to 0.895 / 0.88)	Adaptive attack evades detectors
Pawlak et al. 2025	Single-task poison detection	80% (BASE), 60% (BAIT) of attacks caught	Works with one clean task as anchor
Wu et al. 2023	Test-time adaptation, ImageNet-C	ASR at least 92.80% with 10 of 200 in batch	Batch statistics are a shared self
Steinhardt et al. 2017	MNIST-1-7, 30% poison	Oracle bound 7%; data-dependent error 40%	Anchored versus self-estimated reference
Koh et al. 2022	Enron spam, 3% poison	Test error 3% to 24% past sanitisers	Sanitisers estimated from poisoned data
Qi et al. 2023	GPT-3.5, benign Alpaca	Harmfulness 5.5% to 31.8%	Benign update, attack-like effect
He et al. 2024	100 selected benign examples	ASR 71% vs 13% random (Alpaca)	Found only against known-harmful set
Zhang et al. 2024	Pretraining poison	0.1% persists for 3 of 4 objectives	Small, persistent minority
FLTrust (Cao et al. 2021)	MNIST-0.5, 100-example root set	Error near no-attack baseline up to 90% malicious (best case; over 40% under adaptive attacks)	Anchor cost is small when fresh
Chen et al. 2023	Android malware, monthly labels	F1 0.99 to 0.76 after 6 months; 50 labels/month reach 89% mean F1	Anchor renewal priced

Pawlak et al. (2025) state the defender's position unusually plainly. In a realistic continual setup, they write, the defender has no clean validation set, because validation data are split from the new task and are poisoned along with it, and has no access to earlier tasks. Monitoring current-task validation accuracy therefore does not reveal the attack. Their detector compares task vectors and sets its threshold from one clean task at the start of the sequence; with the simplest threshold choice it reached 92.3 and 86.2 percent overall accuracy for their two attack types, catching 80 and 60 percent of attacks respectively. They also note that their single-task threat model gives memory-based methods an unrealistic guarantee that stored samples are clean. The anchor is the first task.

Hu and Duan (2026) give a defence with a proof, but the proof covers a variant of the rule they first describe. Their task-to-task verification flags a task when a detection score built from two consecutive updates rises above a rolling-window mean, and rolls the model back two steps. The authors call that rolling-window test "the heuristic detection condition", and it carries no guarantee. Their guarantee - final excess risk of order K squared over T with K poisoned tasks among T - holds only when the defence "replaces its heuristic detection condition with $dt > \theta t$ ", where the threshold θt is set by the noise variance σ^2 of the clean data, known in advance, rather than by the history of the scores; the regulariser it uses also needs σ^2 and a known bound W on the size of the ground-truth parameter. The guarantee further assumes that every task's clean data come from one linear model with a single shared ground-truth parameter, with label poisoning only. The proof thus

holds the benign target still, takes its threshold from outside the stream, and restricts the attacker: the defence is designed for what the authors call infrequent attacks, a finite number of poisoned tasks, and its bound vanishes only while K grows more slowly than the square root of T .

Lin et al. (2026) address backdoors in continual learning with purification by prototype discrepancy, recovery of informative poisoned samples, and robust routing among experts; their experiments fix the poisoning rate at 5 percent of each task. Separation by feature-space discrepancy presumes that poisoned samples are a minority whose features depart from their class's prototype. Umer and Polikar (2023) take a different route, training a stronger, perceptible pattern to overpower an imperceptible backdoor without knowing the target task or class; it is a defence against one attack form, not a detector of departures.

7.3 Test-time adaptation

Test-time adaptation is the purest Reading 2 system: the model updates on each incoming unlabelled batch, so the batch is the self. Wu et al. (2023) showed what follows. Because predictions on benign samples depend on other samples in the same batch, an attacker who contributes a few samples can change predictions for everyone; with 10 malicious samples in a batch of 200 on ImageNet-C, their attack reached at least 92.80 percent success against every adaptation method tested, and near 100 percent with 16 against TeBN, the method that only re-estimates normalisation statistics. Park et al. (2024) replaced batch means with medians in normalisation, a robust-statistics remedy that limits a

minority's influence on the shared estimate.

Su et al. (2025) supply the strongest counterweight in this paper. They argue that earlier attacks on test-time adaptation assumed too much - white-box access, access to benign samples, offline control of attack order - and that under realistic assumptions adaptation methods are more robust than had been believed. They name confidence thresholding, data augmentation, exponential moving averaging and random restoration of parameters as "potential directions for improving the adversarial robustness of TTA methods", and their ablation separates them. Augmentation and the moving average were "very effective defense practices"; random restoration to the source model improved robustness on its own, but added nothing once entropy thresholding, augmentation and the moving average were in place (Su et al., 2025, Table 6). An exponential moving average slows the self, which raises the cost of fast attacks and, by the argument of Section 6.3, leaves slow ones. Restoring parameters to the source model re-anchors the self to a reference the stream did not write, but in this ablation the anchor-like practice added nothing on top of a self-referential one, so it gives the anchor argument only weak support.

7.4 Language models updated after deployment

Zhang et al. (2024) pretrained models from 600 million to 7 billion parameters with poisoned data and found that poisoning 0.1 percent of the pretraining set was enough for three of four attack objectives to persist measurably through supervised fine-tuning and preference optimisation, and that denial-of-service persisted at 0.001 percent. Souly et al. (2025) found that about 250 poisoned documents compromised models from 600 million to 13 billion parameters similarly, despite a more than twentyfold difference in clean data. Mahendrakar (2026c) argues that the near-constant result is narrower than its popular reading, and nothing here depends on the stronger reading. What matters for this paper is that a poisoner of a continually updated model is a small, persistent minority - the regime in which Section 6.3 says a tolerising reference is weakest.

Hubinger et al. (2024) showed that backdoor behaviour can persist through supervised fine-tuning, reinforcement learning and adversarial training, and that adversarial training can teach a model to recognise its trigger better, hiding the behaviour. For the immune proposal this is the recovery half's problem rather than detection's, and Section 10 returns to it.

8. Every located guarantee holds the benign distribution still

8.1 The sorting

Section 6 established that separation of attack from legitimate change is impossible in general. That is compatible with good defences in particular settings, and the literature has several with proofs or certified bounds,

and some with measured detection results only. The useful question is what each one assumes. Table 2 sorts the positive results located for this paper by what they take as given about the benign data and where their reference comes from.

Three patterns account for the rows. Several rows also hold part of their reference where the stream cannot rewrite it: Steinhardt et al. (2017) and FLTrust, taken up in Section 8.2; Pawlak et al. (2025), with a clean first task; Korycki and Krawczyk (2020), with a trained detector and labels; and the proved version of Hu and Duan (2026), with a noise level known in advance. These too assume something fixed - the true centroids, root data close to the training distribution, a clean first task, a known noise level.

The benign distribution is held fixed. Kloft and Laskov (2010) assume that innocuous points have expectation equal to the detector's initial centre; their protection against a capped attacker works because the honest traffic keeps pulling the centre home. Diakonikolas et al. (2016) estimate a single distribution from a corrupted sample. Hu and Duan (2026) prove both of their results for tasks sharing one ground-truth parameter. In each, the self is stationary by assumption, and the proof would not survive a benign target that moves without bound - the case continual learning exists for.

The benign distribution is allowed to move, but by a bounded amount, and the bound becomes the error. Karimireddy et al. (2022) allow heterogeneity up to zeta squared and prove that delta times zeta squared is the best achievable error; they then circumvent the bound only under stronger assumptions such as mild heterogeneity or overparameterisation. Allouah et al. (2023) achieve the optimal rate under heterogeneity, which is the same floor. Translated to the time axis, a learner allowed legitimate drift of size zeta inherits an irreducible error proportional to the attacker's share times that drift. The looser the self, the larger the floor.

The attacker is assumed rare, bursty, late or weak. Korycki and Krawczyk (2020) design for a world in which real drift becomes more common over the stream while adversarial drift tends to be periodic and a small share of instances. Hu and Duan (2026) defend against a number of attacked tasks that grows more slowly than the square root of the horizon, and prove that a linear fraction of shifted attacks cannot be defended. Their second defence, a robust feature defence, does face an attacker who poisons every task, but only one whose per-task budget grows sublinearly in the horizon and whose noise does not shift the target; against that attacker the authors describe it as "provably accelerating the convergence rate". Pawlak et al. (2025) assume one poisoned task and a clean first one. Lin et al. (2026) evaluate at 5 percent per task. These are reasonable priors about many deployments; they are not facts the stream can confirm.

Table 2. Positive results against poisoning - proved guarantees and certified bounds ("Proof"), and empirical detection results with no proof ("Empirical") - sorted by what each assumes about the benign distribution. "Anchor" names the part of the reference that the incoming stream cannot rewrite. Assumptions are as stated by each source; the sorting is this paper's.

Source	Result	Benign-side assumption	Anchor
Kloft and Laskov 2010	Proof: attacker gain bounded	Innocuous mean equals initial centre; attacker share capped	Initial centre; false-alarm ceiling
Steinhardt et al. 2017	Proof: certified upper bound on loss	Clean data concentrate; defence uses true centroids	Oracle centroids
Diakonikolas et al. 2016	Proof: dimension-free error	Clean samples from one fixed distribution class	Distribution family
Karimireddy et al. 2022	Proof: optimal rate $O(\delta \log \frac{1}{\delta})$	Honest heterogeneity bounded by ζ	None; error floor remains
Hu and Duan 2026 (task-to-task verification)	Proof: excess risk $O(K^2/T)$, fixed-threshold variant only	One shared ground-truth parameter; known noise variance; $K = o(\sqrt{T})$ poisoned tasks	Threshold from known noise level σ^2 (proved version); none for the rolling-window heuristic, which has no guarantee
Hu and Duan 2026 (robust feature defence)	Proof: faster convergence than an EWC-equivalent regulariser	One shared ground-truth parameter; attacks frequent but bounded per task and non-shifted; stationary protected dimensions	Noise level and attack budget known in advance
Korycki and Krawczyk 2020	Empirical: robust drift detection	Attacks a small, periodic minority; drift grows	Trained detector, labels
Pawlak et al. 2025	Empirical: 60-80% of attacks caught	First task clean	Clean first task
Cao et al. 2021 (FLTrust)	Proof: bounded distance to optimum; empirical: near no-attack error	Root data close to training distribution	100 clean root examples

8.2 What the anchored rows add

Two of the anchored rows of Table 2 stand apart because their reference is held wholly outside the loop, and they show what an anchor buys. Steinhardt et al. (2017) compared an oracle defence that knows the true class centroids with a realistic one that estimates them from the possibly poisoned data. Against the oracle, no attack they considered pushed an SVM's error on MNIST-1-7 or Dogfish above 4 percent even with 30 percent poisoned data, and they certified upper bounds of 7 and 10 percent. Against the data-dependent version, 30 percent poisoned data drove MNIST-1-7 error to 40 percent. Their summary is the thesis of this paper in one sentence: defences relying on the potentially poisoned data can be much weaker than their data-independent counterparts. Koh et al. (2022) then built attacks that evade sanitisers based on nearest neighbours, training loss and singular-value decomposition, raising Enron spam test error from 3 to 24 percent with 3 percent poisoned data.

FLTrust (Cao et al., 2021) makes the opposite point from the federated side. Its authors attribute the weakness of earlier Byzantine-robust methods to the absence of a root of trust, and give the server a small clean dataset that it labels itself, 100 examples by default. Each client update is scored by its direction relative to the server's own update on that root set and rescaled to the same magnitude. On MNIST-0.5 they report error rates close to undefended federated averaging without attack when up to 90 percent of clients are malicious. That is their best case: under their adaptive attacks the MNIST-0.5 figure is "over 40%" of clients, and their abstract gives "40%-60%" across datasets. Even at those figures, a tiny anchor did what no amount of statistics on the clients

alone could.

8.3 The pattern, stated

No located guarantee holds when the benign distribution may move without bound and a persistent attacker holds a minority share. Every proved result either fixes the benign target, bounds its movement and pays the bound as error, or limits the attacker in number, timing or strength, and the two empirical rows rest on assumptions of the same kinds. The anchored defences escape the trilemma only because the stream cannot rewrite their reference - and FLTrust's authors attach the caveat that matters for continual learning: the method works when the root data do not depart too far from the training distribution. A fixed root set in a moving world is Reading 1 again, one level up.

This is an observation about the literature found, not a theorem. A search can miss a result, and Section 13 says what a counterexample would look like.

9. What the danger model asks for, once translated

9.1 Damage, measured on something the update cannot touch

The danger model replaces "is this foreign?" with "is this doing harm?" (Matzinger, 1994). In a learning system, harm is a change in what the model does on inputs that matter, relative to what it should do. Measuring that needs three things the update loop does not provide for itself: inputs that represent what matters, a statement of what the model should do on them, and assurance that neither was written by the party whose update is being judged. Each anchored defence located here supplies

some of these, in one of five forms.

The first form is trusted data with labels. FLTrust's root set (Cao et al., 2021) and the trusted subset that Hendrycks et al. (2018) use to correct for severe label corruption are both small and both assumed clean; FLTrust's is labelled by the server itself. The second is provenance. Baracaldo et al. (2017) filter poison by where each training point came from and what was done to it on the way, as recorded by a provenance system the attacker cannot alter; they target models that learn online or are retrained regularly, and give one version for data that is partly trusted and one for data that is not trusted at all. The third is human labelling at a budget. Chen et al. (2023) keep an Android malware classifier current by having analysts label a monthly sample chosen by active learning; Barbero et al. (2022) set aside the samples a classifier is likely to get wrong and leave them for expert analysis; Dey et al. (2020) bring an analyst's annotations into a continually trained behavioural detector to resist frog-boiling. The fourth is a restoration point. Su et al. (2025) found that random restoration of parameters to the source model improved test-time adaptation's robustness on its own, though not on top of their other practices, Pawlak et al. (2025) roll weights back to the checkpoint before a flagged task, and Hu and Duan (2026) discard the last two updates on detection. The fifth is a known-harmful reference: He et al. (2024) find the benign data that break safety only by measuring similarity to known harmful examples.

Each of these is supervision. None is self/non-self discrimination; each is a damage signal grounded outside the learner, which is the role the danger model gives the tissues.

9.2 The decision structure

Algorithm 1 states the two decision rules this paper contrasts, in a form that makes the reference explicit. The first is the tolerising gate of Reading 2; the second adds an anchor and a damage test of the Reading 3 kind.

Lines 1 to 3 are where the identifiability results of Section 6 bite. The reference R is written by the accepted updates, so an attacker who keeps each step below τ moves R in the direction of the attack, exactly as in the finite-horizon centroid analysis of Kloft and Laskov (2010). Line 5 is the danger test. It does not ask whether u is unusual; it asks whether u makes the model worse on A , and A is held where u cannot reach it. Line 6 keeps the self-referential statistic for what it is good at - noticing that something changed - and routes the decision about whether the change is legitimate to a source with authority over that question. Line 8 is the part the literature leaves unpriced.

9.3 How anchored defences fail

The anchored gate is not a solution, and three of its failure modes are already visible in the located work.

Staleness. An anchor drawn once is a frozen self. FLTrust's caveat about root-data distribution (Cao et al., 2021) is one instance; the Android result of Chen et al. (2023), where F1 fell from 0.99 to 0.76 after six months without new labels, shows how fast a reference can go stale in a domain where the data are partly adversarial. An anchor that is not renewed at the rate of legitimate change will block legitimate updates, and the operator who widens $d_{\max}$ in response has rebuilt Rule S.

The renewal channel is a new attack surface. Whatever supplies the renewal - analysts, provenance metadata, a trusted subset of clients - becomes the thing to attack. Chan-Tin et al. (2011) report that frog-boiling defeated a network-coordinate defence that required part of the network to be trusted, so trusting a subset of participants is not by itself an anchor; the anchor must be something the attacker cannot shape over time, and participants who themselves update from the shared state do not qualify.

The damage test inherits the coverage of A. A backdoor is designed to leave behaviour on ordinary inputs unchanged (Tran et al., 2018), so a damage test on an anchor without the trigger passes it. Line 5 catches forgetting attacks such as BrainWash, whose damage is on the anchor's own tasks; it does not by itself catch a trigger it never sees. The danger model narrows the problem to the coverage of the anchor; it does not close it.

9.4 The price, where anyone has stated it

Only one located study puts a number on keeping an anchor current while the data move. Chen et al. (2023) report that on their seven-year Android dataset the best earlier active-learning method needed 400 analyst labels a month to hold a mean F1 of 89 percent, while their method needed 50, an eightfold saving. That is a measured cost of anchor renewal, in one domain, where the drift comes from the evolution of both malware and benign apps. FLTrust's 100 examples give a second, static figure. No continual-learning poisoning paper located here reports what its defence would cost to keep anchored over a long stream, and none reports a false-alarm rate on legitimate shifts of the size its benign tasks contain.

10. The recovery half

An immune-inspired proposal usually pairs detection with repair: once poison is found, remove its effect. The repair half meets a problem that Mahendrakar (2026d) treats at length for unlearning, and the short version is that suppression is easily mistaken for removal.

The backdoor literature shows it directly. Min et al. (2024) find that models purified to a low attack success

Algorithm 1. Self-referential gate versus anchored gate for one update u to a continual learner

```

input : model M, update u, reference R (statistics of accepted past updates/data),
        tolerance tau, anchor A (trusted inputs + targets, held outside the loop),
        damage budget d_max, renewal policy P
-- Rule S: self-referential (Reading 2) -----
1 s <- distance(u, R)                # departure from what has been accepted
2 if s > tau then reject u
3 else apply u to M; R <- update(R, u) # tolerance: accepted change becomes self
  # every accepted u moves R, so a sequence of small departures is never flagged
-- Rule A: anchored (danger-keyed, Reading 3) -----
4 s <- distance(u, R); harm <- loss(M + u, A) - loss(M, A)
5 if harm > d_max then reject u and restore M to last checkpoint passing A
6 else if s > tau then hold u for external review (label, provenance, analyst)
7 else apply u to M; R <- update(R, u)
8 A <- renew(A, P)                  # anchor must track legitimate change; cost is P's budget
  # A is never written from the stream that u comes from

```

rate re-learn the backdoor quickly when fine-tuned on very few poisoned samples, and that the backdoor can be reactivated merely by querying the purified model; they trace the failure to purified models not moving far enough from the backdoored model along the backdoor-connected path. Hubinger et al. (2024) find backdoors that persist through standard safety training. In unlearning, Xu et al. (2025) show that models can appear to forget while minimal fine-tuning restores the original behaviour, and Hu et al. (2025) recover unlearned hazardous knowledge by fine-tuning on loosely related public text.

The counterweight is real. Min et al. (2024) themselves propose Path-Aware Minimization, a tuning defence that pushes the purified model further from the backdoored one along the backdoor-connected path, and report that it "significantly improves post-purification robustness". Xu et al. (2025) also report a case of seemingly irreversible targeted forgetting, and Zhu et al. (2026) formalise complete backdoor unlearning as a continual-learning problem and derive necessary conditions for achieving it. The defensible position is that repair has to be verified against the same kind of external reference as detection: a model "cleaned" by a procedure that only checks its own outputs on its own data has been checked by Reading 2.

There is a plainer form of recovery that the anchored defences already use: roll back. Pawlak et al. (2025), Hu and Duan (2026) and the parameter restoration Su et al. (2025) report all restore an earlier state rather than repair a corrupted one. Rollback needs only a checkpoint that predates the poison and a test that says it does, which is again an anchor.

11. What would settle the open part

The open question is not whether self-referential defences can be fooled; Section 6 settles that. It is how much supervision an anchored defence needs, how often, and whether that amount is affordable for the continual learners now being deployed. Four evaluations would answer it, and none needs new theory.

A stream with both kinds of change and known ground truth. Existing continual-learning poisoning benchmarks inject attacks into otherwise stationary or task-structured streams. A benchmark that also contains legitimate drift of controlled size - including drift that degrades a held-out metric, as benign fine-tuning degraded safety for Qi et al. (2023) - would let a defence be scored on both error types at once: poison accepted and legitimate change refused. The second error is the one current papers do not report.

A rate sweep. For any defence with a tolerance or rolling reference, report attack success as a function of the attack's rate of change at fixed total displacement. Kloft and Laskov (2010) and Chan-Tin et al. (2011) predict that success rises as rate falls for Reading 2 defences; an anchored defence should show no such dependence. The sweep is cheap and would sort published defences into the two classes empirically.

The anchor budget as a reported quantity. Every anchored defence should report the size of its anchor, how often it is renewed, who renews it, and how its accuracy degrades if renewal stops, in the manner of Chen et al. (2023). Without that number, "uses a small trusted set" hides most of the cost.

An attack on the anchor. The renewal channel should be in the threat model. Where renewal comes from human labels chosen by active learning, an attacker can shape which samples are chosen; Korycki and Krawczyk (2020) already note that adversarial drift wastes a scarce labelling budget on partially labelled streams. Reporting robustness when a fraction of renewals is adversarial would show whether the anchor holds.

12. The strongest case against this paper

The priors are usually true. Attacks on deployed systems are typically rare relative to benign traffic, and often arrive in bursts. If so, the assumptions in the third pattern of Section 8 hold in practice, and Hu and Duan (2026), Pawlak et al. (2025) and Korycki and Krawczyk (2020) give working defences. This paper concedes the point for many deployments. The claim is narrower: those defences certify nothing against an attacker who

chooses to be slow and persistent, and Section 6.3 shows that is the cheapest strategy against a tolerising reference.

The risk is overstated. Su et al. (2025) found that for test-time adaptation, realistic attacks do less than earlier papers suggested. That result weakens the urgency of the test-time case and should temper any reading of Table 1 as a measure of real-world risk. It also gives the anchor argument little help: in their ablation, restoration to a source model, an anchor, improved robustness alone but added nothing to the self-referential practices combined (Section 7.3).

The conclusion is close to trivial. "Use supervision" is not a new idea, and FLTrust and Chen et al. show it is cheap. The answer is that the contribution lies in the sorting of Table 2 and in the price that is missing: the continual-learning defences located here report detection rates without the renewal cost or the false-alarm rate on legitimate shift that would make them comparable with an anchored alternative.

The immune framing is a straw man. A second generation of artificial immune systems research turned from frozen self/non-self to danger signals from around 2002, so attacking negative selection may look like attacking a superseded position. It is not superseded: negative selection is still developed, and a 2021 review argues that it outperforms other methods in some settings (Gupta and Dasgupta, 2021). That work does not address references that must move under attack (Section 5.1), and the paper uses the turn to danger signals as evidence: the line of work that took the analogy furthest found it needed a damage signal supplied from outside.

The paper overlaps the author's earlier work. The identifiability argument resembles Mahendrakar (2026a) and (2026b). The objects differ - a stream's classes, a memory store's records, and here a learner's own update reference - and the present paper's evidence on online centroid detection, Byzantine lower bounds, test-time adaptation and immunology's discontinuity theory does not appear in those papers. A reader who finds the shared structure more salient than the difference is not being unreasonable.

13. What is not known

No study located here measures a continual-learning defence's false-alarm rate on legitimate shifts of known size.

No study located here sweeps an attack's rate of change against a continual-learning defence at fixed total displacement.

No continual-learning poisoning defence reports the cost of keeping its anchor current over a long stream; the only priced renewal located is in Android malware detection (Chen et al., 2023).

Whether consolidation rules triggered by loss plateaus, such as that of Aljundi et al. (2019), protect poisoned knowledge once learned is untested.

Whether the slow-change tolerance of the biological immune system described by Pradeu and Vivier (2016) has a quantitative analogue in learning systems - a rate below which an attack is absorbed - has not been measured, although the online centroid analysis of Kloft and Laskov (2010) suggests its form.

Whether any self-referential defence achieves non-trivial guarantees when the benign distribution moves without bound and the attacker holds a persistent minority share is open. A counterexample to Section 8.3 would be a defence with a proof under exactly those conditions and no reference held outside the stream. None was found.

14. Conclusion

The objection that continual learning has no stable self is correct about frozen references and beside the point for most of the defences deployed, which use a tolerised self. A tolerised self fails in a specific way: it absorbs change that is slow enough. That has been shown for retrained anomaly detectors and network coordinate systems, and it is how the biological immune system's own discontinuity theory describes it. For test-time adaptation and rolling-mean detectors in continual learning it is a prediction carried over from those results, not yet tested. Every positive guarantee located for this paper buys its result by fixing the benign distribution, bounding its movement, or limiting the attacker. The defences that escape do so with a reference the stream cannot write, which is what the danger model always required. The immune analogy's useful lesson for continual learning is therefore not self-recognition but grounding: a learner cannot certify its own updates from its own history, and the open work is to price the outside reference it needs and to keep that reference honest while the world moves.

References

- Abbasi, A., Nooralinejad, P., Pirsiavash, H., & Kolouri, S. (2023). BrainWash: A Poisoning Attack to Forget in Continual Learning. arXiv:2311.11995. doi:10.48550/arXiv.2311.11995. [PREPRINT]
- Aickelin, U., & Cayzer, S. (2002). The Danger Theory and Its Application to Artificial Immune Systems. Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS 2002), 141-148. arXiv:0801.3549. doi:10.48550/arXiv.0801.3549.
- Aljundi, R., Kelchtermans, K., & Tuytelaars, T. (2019). Task-Free Continual Learning. CVPR 2019. arXiv:1812.03596. doi:10.48550/arXiv.1812.03596.
- Allouah, Y., Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., & Stephan, J. (2023). Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity. AISTATS 2023. arXiv:2302.01772. doi:10.48550/arXiv.2302.01772.
- Barbero, F., Pendlebury, F., Pierazzi, F., & Cavallaro, L. (2022). Transcending TRANSCEND: Revisiting Malware Classification in the Presence of Concept Drift. 2022 IEEE Symposium on Security and Privacy. doi:10.1109/SP46214.2022.9833659.

- Baracaldo, N., Chen, B., Ludwig, H., & Safavi, J. A. (2017). Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. doi:10.1145/3128572.3140450.
- Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. *Machine Learning*, 81, 121-148. doi:10.1007/s10994-010-5188-5.
- Baruch, M., Baruch, G., & Goldberg, Y. (2019). A Little Is Enough: Circumventing Defenses For Distributed Learning. *NeurIPS 2019*. arXiv:1902.06156. doi:10.48550/arXiv.1902.06156.
- Cao, X., Fang, M., Liu, J., & Gong, N. Z. (2021). FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. *NDSS 2021*. arXiv:2012.13995. doi:10.48550/arXiv.2012.13995.
- Chan-Tin, E., Heorhiadi, V., Hopper, N., & Kim, Y. (2011). The Frog-Boiling Attack: Limitations of Secure Network Coordinate Systems. *ACM Transactions on Information and System Security*, 14(3). doi:10.1145/2043621.2043627.
- Chen, Y., Ding, Z., & Wagner, D. (2023). Continuous Learning for Android Malware Detection. *USENIX Security 2023*. arXiv:2302.04332. doi:10.48550/arXiv.2302.04332.
- Dey, A., Velay, M., Fauvel, J.-P., & Navers, S. (2020). Adversarial vs behavioural-based defensive AI with joint, continual and active learning: automated evaluation of robustness to deception, poisoning and concept drift. *C&ESAR 2019*. arXiv:2001.11821. doi:10.48550/arXiv.2001.11821.
- Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., & Stewart, A. (2016). Robust Estimators in High Dimensions without the Computational Intractability. arXiv:1604.06443. doi:10.48550/arXiv.1604.06443.
- Forrest, S., Perelson, A. S., Allen, L., & Cherukuri, R. (1994). Self-nonself discrimination in a computer. *Proceedings of the 1994 IEEE Computer Society Symposium on Research in Security and Privacy*. doi:10.1109/RISP.1994.296580.
- Forrest, S., Hofmeyr, S. A., Somayaji, A., & Longstaff, T. A. (1996). A sense of self for Unix processes. *Proceedings of the 1996 IEEE Symposium on Security and Privacy*. doi:10.1109/SECPRI.1996.502675.
- Freitas, A. A., & Timmis, J. (2007). Revisiting the Foundations of Artificial Immune Systems for Data Mining. *IEEE Transactions on Evolutionary Computation*, 11(4). doi:10.1109/TEVC.2006.884042.
- Greensmith, J., Whitbrook, A., & Aickelin, U. (2010). Artificial Immune Systems. In *Handbook of Metaheuristics*, 2nd edition. arXiv:1006.4949. doi:10.48550/arXiv.1006.4949.
- Guo, Z., Kumar, A., & Tourani, R. (2025). Persistent Backdoor Attacks in Continual Learning. *USENIX Security 2025*. arXiv:2409.13864. doi:10.48550/arXiv.2409.13864.
- Gupta, K. D., & Dasgupta, D. (2021). Negative Selection Algorithm Research and Applications in the last decade: A Review. arXiv:2105.06109. doi:10.48550/arXiv.2105.06109. [PREPRINT]
- Han, G., Choi, J., Hong, H. G., & Kim, J. (2023). Data Poisoning Attack Aiming the Vulnerability of Continual Learning. *ICIP 2023*. arXiv:2211.15875. doi:10.48550/arXiv.2211.15875.
- He, L., Xia, M., & Henderson, P. (2024). What is in Your Safe Data? Identifying Benign Data that Breaks Safety. arXiv:2404.01099. doi:10.48550/arXiv.2404.01099.
- Hendrycks, D., Mazeika, M., Wilson, D., & Gimpel, K. (2018). Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise. *NeurIPS 2018*. arXiv:1802.05300. doi:10.48550/arXiv.1802.05300.
- Hofmeyr, S. A., & Forrest, S. (2000). Architecture for an Artificial Immune System. *Evolutionary Computation*, 8(4). doi:10.1162/106365600568257.
- Hu, S., Fu, Y., Wu, Z. S., & Smith, V. (2025). Unlearning or Obfuscating? Jogging the Memory of Unlearned LLMs via Benign Relearning. *ICLR 2025*. arXiv:2406.13356. doi:10.48550/arXiv.2406.13356.
- Hu, Y., & Duan, L. (2026). Theory of Continual Learning Against Data Poisoning Attacks. arXiv:2606.29841. doi:10.48550/arXiv.2606.29841. [PREPRINT]
- Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566. doi:10.48550/arXiv.2401.05566. [PREPRINT]
- Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *ICLR 2022*. arXiv:2006.09365. doi:10.48550/arXiv.2006.09365.
- Kim, J., Bentley, P. J., Aickelin, U., Greensmith, J., Tedesco, G., & Twycross, J. (2007). Immune system approaches to intrusion detection - a review. *Natural Computing*, 6(4), 413-466. doi:10.1007/s11047-006-9026-4. arXiv:0804.1266.
- Kloft, M., & Laskov, P. (2010). Security Analysis of Online Centroid Anomaly Detection. arXiv:1003.0078. doi:10.48550/arXiv.1003.0078. [PREPRINT]
- Koh, P. W., Steinhardt, J., & Liang, P. (2022). Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, 111, 1-47. doi:10.1007/s10994-021-06119-y. arXiv:1811.00741.
- Korycki, L., & Krawczyk, B. (2020). Adversarial Concept Drift Detection under Poisoning Attacks for Robust Data Stream Mining. arXiv:2009.09497. doi:10.48550/arXiv.2009.09497. [PREPRINT]
- Li, H., & Ditzler, G. (2023). PACOL: Poisoning Attacks Against Continual Learners. arXiv:2311.10919. doi:10.48550/arXiv.2311.10919. [PREPRINT]
- Lin, K., Ye, F., Liu, Q., Zhou, S., & Yu, J. (2026). Robust Dynamic Expansion for Continual Learning under Backdoor Attacks via Purification and Selective Recovery. arXiv:2609.06346. doi:10.48550/arXiv.2609.06346. [PREPRINT]
- Mahendrakar, P. (2026a). Drift or New Class? Without Labels a Drifted Class and a New One Can Produce the Same Stream, the Two Lines of Work With the Most Explicit Assumptions Each Get an Answer by Freezing the Variable the Other Lets Move, and No Located Benchmark Scores the Attribution. *Life of Research*. doi:10.5281/zenodo.23041727.
- Mahendrakar, P. (2026b). Intent Is Not a Property of the Record: Agent Memory Contamination, the Two Detector Families That Cover Disjoint Failures, and the Base Rate Nobody Has Measured. *Life of Research*. doi:10.5281/zenodo.22088221.
- Mahendrakar, P. (2026c). Poison Below the Base Rate: What the Near-Constant Poisoning Result Establishes About Detection, and Four Things It Does Not. *Life of Research*. doi:10.5281/zenodo.22073424.
- Mahendrakar, P. (2026d). Three Guarantees Under One Word: What Importance-Based Selective Forgetting in Language Models Can and Cannot Promise. *Life of Research*. doi:10.5281/zenodo.22054075.
- Matzinger, P. (1994). Tolerance, Danger, and the Extended Family. *Annual Review of Immunology*, 12, 991-1045. doi:10.1146/annurev.iy.12.040194.005015.
- Matzinger, P. (2002). The Danger Model: A Renewed Sense of Self. *Science*, 296(5566), 301-305. doi:10.1126/science.1071059.
- Min, R., Qin, Z., Zhang, N. L., Shen, L., & Cheng, M. (2024). Uncovering, Explaining, and Mitigating the Superficial Safety of Backdoor Defense. *NeurIPS 2024*. arXiv:2410.09838. doi:10.48550/arXiv.2410.09838.
- Park, H., Hwang, J., Mun, S., Park, S., & Ok, J. (2024). MedBN: Robust Test-Time Adaptation against Malicious Test Samples. *CVPR 2024*. arXiv:2403.19326. doi:10.48550/arXiv.2403.19326.
- Paudice, A., Munoz-Gonzalez, L., Gyorgy, A., & Lupu, E. C. (2018). Detection of Adversarial Training Examples in Poisoning Attacks through Anomaly Detection. arXiv:1802.03041. doi:10.48550/arXiv.1802.03041. [PREPRINT]

- Pawlak, S., Twardowski, B., Trzcinski, T., & van de Weijer, J. (2025). Addressing The Devastating Effects Of Single-Task Data Poisoning In Exemplar-Free Continual Learning. *CoLLAs* 2025. arXiv:2507.04106. doi:10.48550/arXiv.2507.04106.
- Pradeu, T., & Carosella, E. D. (2006). On the definition of a criterion of immunogenicity. *Proceedings of the National Academy of Sciences*, 103(47), 17858-17861. doi:10.1073/pnas.0608683103.
- Pradeu, T., Jaeger, S., & Vivier, E. (2013). The speed of change: towards a discontinuity theory of immunity? *Nature Reviews Immunology*, 13, 764-769. doi:10.1038/nri3521.
- Pradeu, T., & Vivier, E. (2016). The discontinuity theory of immunity. *Science Immunology*, 1(1). doi:10.1126/sciimmunol.aag0479.
- Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693. doi:10.48550/arXiv.2310.03693.
- Rubinstein, B. I. P., Nelson, B., Huang, L., Joseph, A. D., Lau, S., Rao, S., Taft, N., & Tygar, J. D. (2009). ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. doi:10.1145/1644893.1644895.
- Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., Mougan, C., Mavroudis, V., Jones, E., Hicks, C., Carlini, N., & Gal, Y. (2025). Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples. arXiv:2510.07192. doi:10.48550/arXiv.2510.07192. [PREPRINT]
- Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified Defenses for Data Poisoning Attacks. *NIPS* 2017. arXiv:1706.03691. doi:10.48550/arXiv.1706.03691.
- Stibor, T., Mohr, P., Timmis, J., & Eckert, C. (2005). Is negative selection appropriate for anomaly detection? *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation (GECCO 2005)*. doi:10.1145/1068009.1068061.
- Stibor, T., Oates, R., Kendall, G., & Garibaldi, J. M. (2009). Geometrical insights into the dendritic cell algorithm. *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation (GECCO 2009)*. doi:10.1145/1569901.1570072.
- Su, Y., Li, Y., Liu, N., Jia, K., Yang, X., Foo, C.-S., & Xu, X. (2025). On the Adversarial Risk of Test Time Adaptation: An Investigation into Realistic Test-Time Data Poisoning. *ICLR* 2025. arXiv:2410.04682. doi:10.48550/arXiv.2410.04682.
- Tran, B., Li, J., & Madry, A. (2018). Spectral Signatures in Backdoor Attacks. *NeurIPS* 2018. arXiv:1811.00636. doi:10.48550/arXiv.1811.00636.
- Umer, M., & Polikar, R. (2023). Adversary Aware Continual Learning. arXiv:2304.14483. doi:10.48550/arXiv.2304.14483. [PREPRINT]
- Wang, R., Chen, T., Lindsly, S., Stansbury, C., Rehemtulla, A., Rajapakse, I., & Hero, A. (2021). RAILS: A Robust Adversarial Immune-inspired Learning System. arXiv:2107.02840. doi:10.48550/arXiv.2107.02840. [PREPRINT]
- Wang, Y., Jha, S., & Chaudhuri, K. (2020). An Investigation of Data Poisoning Defenses for Online Learning. arXiv:1905.12121. doi:10.48550/arXiv.1905.12121. [PREPRINT]
- Wu, T., Jia, F., Qi, X., Wang, J. T., Sehwag, V., Mahloujifar, S., & Mittal, P. (2023). Uncovering Adversarial Risks of Test-Time Adaptation. arXiv:2301.12576. doi:10.48550/arXiv.2301.12576.
- Xu, X., Yue, X., Liu, Y., Ye, Q., Zheng, H., Hu, P., Du, M., & Hu, H. (2025). Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs. *ICML* 2026. arXiv:2505.16831. doi:10.48550/arXiv.2505.16831.
- Zhang, Y., Rando, J., Evtimov, I., Chi, J., Smith, E. M., Carlini, N., Tramer, F., & Ippolito, D. (2024). Persistent Pre-Training Poisoning of LLMs. arXiv:2410.13722. doi:10.48550/arXiv.2410.13722.
- Zhu, Z., Hu, Y., Liu, Y., Wei, L., Hou, W., Li, B., Li, H., & Luo, W. (2026). Rethinking Backdoor Adversarial Unlearning through the Lens of Catastrophic Forgetting in Continual Learning. *ACM CCS* 2026. arXiv:2606.14078. doi:10.48550/arXiv.2606.14078.