English version below

Wenn KI schneller liefert, als geprüft werden kann

Eine Fallstudie über Schatten-KI, den Verifier-Gap und die Frage, warum Nachweisbarkeit in regulierten Branchen nicht verhandelbar ist.

Die Kernaussage vorweg

Mehr Erkenntnis ist nicht mehr geprüfte Erkenntnis. Künstliche Intelligenz macht das Erzeugen billig, das Prüfen bleibt teuer. Wer nur das Erzeugen beschleunigt, schiebt das Risiko nach hinten, dorthin, wo es am teuersten wird. Unsere Fallstudie zeigt an einem realen Beispiel aus dem Finanzsektor, wie aus einer gut gemeinten, produktiven KI-Nutzung ein aufsichtsrechtlicher Ernstfall wurde, und warum die Antwort darauf kein Verbot ist, sondern Verifikation an der richtigen Stelle.

Warum Schatten-KI entsteht

Der Begriff Schatten-IT stammt aus den achtziger Jahren und meinte Systeme, die in Fachbereichen an der zentralen IT vorbei entstanden, weil die offiziellen Wege zu langsam waren. Schatten-KI ist nichts grundsätzlich Neues, sondern dasselbe Muster nur eine Stufe weiter. Neu ist, wie leicht solche Lösungen heute entstehen und wie überzeugend ihre Ergebnisse aussehen.

Dahinter steckt keine Nachlässigkeit, sondern Rationalität. KI-Werkzeuge sind im Fachbereich oft frei verfügbar und liefern in Minuten, wofür das Standardvorgehen inklusive offizieller Prüfung Wochen braucht. Wer unter Termindruck steht und liefern muss, nimmt den schnelleren Weg. Der Fachbereich handelt dabei nicht gegen die Organisation, sondern für seine eigenen Ziele. Das macht Schatten-KI so schwer zu fassen. Sie ist kein Regelverstoß aus Bequemlichkeit, sondern die naheliegende Antwort auf einen echten Engpass.

Der Fall

In einem Unternehmen aus dem Finanzsektor mit komplexer Kreditvergabe baute die gewerbliche Kreditabteilung eine eigene Lösung. Im Browser lief ein großes Sprachmodell,

angebunden über eine Cloud-Schnittstelle, ergänzt um selbst geschriebenen Python-Code. Damit fasste die Abteilung Firmendaten wie Jahresberichte, Bilanzen und Risikoberichte automatisch zusammen und bereitete sie für die Bewertung der Kreditwürdigkeit auf. Für die Sachbearbeiter war das ein enormer Gewinn. Was vorher Stunden konzentrierter Lesearbeit kostete, lag nun in Minuten vor.

Der Impuls war nachvollziehbar, die Lösung funktionierte, und sie wurde genutzt. Über Wochen flossen ihre Auswertungen in echte Kreditentscheidungen ein, ohne dass außerhalb der Abteilung jemand davon wusste.

Später fiel in einer internen Revision auf, dass die offiziell eingereichten Daten und die internen Akten nicht zusammenpassten, vor allem bei den Schuldenständen der bewerteten Unternehmen. Die Ursache lag im Werkzeug. Der KI-Agent hatte bestimmte Tabellenformate falsch ausgewertet und die entstehenden Unstimmigkeiten nicht gemeldet, sondern rechnerisch geglättet. Das Ergebnis sah gerade deshalb schlüssig aus, weil der Fehler unsichtbar gemacht worden war. Das ist das Tückische an flüssig formulierter KI-Ausgabe. Sie wirkt sicher, auch wo sie es nicht ist.

Die Folgen waren erheblich. Kreditrisiken wurden falsch bewertet, aufsichtsrechtliche Sorgfaltspflichten damit verletzt. Die Gefahr von Kreditausfällen war unbemerkt gestiegen und hätte das Unternehmen bei schlechtem Verlauf in eine ernste Schieflage bringen können. Jeder in diesem Zeitraum freigegebene Kredit musste von Hand nachgeprüft werden. Es folgte ein aufwendiges Audit der internen Prozesse, dazu strenge Auflagen der Aufsicht und interne Disziplinarverfahren, denn die vorgeschriebene Modellvalidierung war vollständig umgangen worden.

Der Kern des Problems war nicht die KI. Er lag darin, dass gleich mehrere Kontrollen umgangen wurden:

- das Risikomanagement, also Risk Management und Model Risk Governance
- die IT-Compliance nach den geltenden aufsichtsrechtlichen Vorgaben
- Architektur-Governance (inkl. Einhaltung der Vorgaben und Prinzipien)

Dazu zählen in Deutschland die MaRisk und die durch DORA überlagerten IT-Anforderungen, ergänzt um den EU AI Act, der die KI-gestützte Kreditwürdigkeitsprüfung als Hochrisiko einstuft, dessen Pflichten allerdings erst ab Dezember 2027 verbindlich greifen.

Innovationen sind nicht das Problem, sondern die fehlende Prüfung. Ob eine Organisation neue Lösungen entwickelt, Prozesse verschlankt oder Produkte entwirft, steht nicht in Frage. Es kommt darauf an, dass Prüfinstanzen im Prozess verankert sind, die KI-gestützte Ergebnisse auf den Prüfstand stellen. Das ist keine bürokratische Hürde, sondern ein Sicherheitsnetz, das regulatorische, sicherheits- und datenschutzbezogene Anforderungen zugleich auffängt.

Das Muster dahinter: der Verifier-Gap

Im Kern steckt ein Konflikt zwischen zwei Geschwindigkeiten, das Erzeugen ist schnell geworden, das Prüfen bleibt langsam. Diese Lücke bezeichnen wir als den Verifier-Gap. Auf den ersten Blick wirkt der Fall wie ein Einzelversagen, eine Abteilung, die es zu gut meinte. Tatsächlich zeigt er ein Muster, das mit jedem besseren KI-Werkzeug häufiger wird. Erzeugung ist billig geworden. Ein plausibler Vorschlag, eine formulierte Einschätzung, eine fertige Auswertung entstehen in Sekunden. Die Prüfung, ob das Ergebnis auch stimmt, ist nicht billiger geworden. Sie dauert so lange wie zuvor, manchmal länger.

Zwischen diesen beiden Geschwindigkeiten öffnet sich eine Lücke. Die KI produziert schneller, als die Organisation prüfen kann. Wo früher ein Mensch einen Vorschlag erarbeitet und dabei schon geprüft hat, liefert die KI das fertige Ergebnis, ohne die Prüfung mitzuliefern. Was wie ein Produktivitätsgewinn aussieht, verschiebt nur den Engpass. Er wandert vom Erzeugen zum Prüfen, und dort staut sich, was vorher gleichmäßig verteilt war.

Entscheidend ist, woran sich ein Ergebnis überhaupt prüfen lässt. KI ist stark und verlässlich, wo eine billige, schnelle Kontrolle ihre Vorschläge sofort bestätigen oder widerlegen kann. Beim Erschließen und Strukturieren von Daten ist das so, das Ergebnis lässt sich in Sekunden gegenprüfen. Bei einer Kreditwürdigkeit, einer Tarifierung, einer Architekturentscheidung fehlt diese schnelle Kontrolle. Ob die Aussage stimmt, zeigt sich nicht sofort, sondern erst im Bestand, in der Prüfung, im Schadenfall, manchmal erst nach Jahren. Genau hier erzeugt die Beschleunigung nicht mehr geprüfte Erkenntnis, sondern mehr ungeprüfte.

Warum es in regulierten Branchen kritisch wird

Was der Kreditabteilung passierte, trifft nicht nur eine Branche. Dasselbe Muster entsteht überall, wo KI ein plausibles Ergebnis liefert und die Prüfung dahinter nicht Schritt hält. In der Versicherungswelt zeigt es sich nur an anderer Stelle.

Ein Aktuariat setzt ein KI-Werkzeug ein, um Tarife schneller zu kalkulieren. Das Modell verarbeitet historische Schadendaten, Bestandsmerkmale und Marktvergleiche und liefert in Minuten einen Tarifvorschlag, für den ein Sachbearbeiter früher Tage gebraucht hätte. Der Vorschlag wirkt schlüssig, die Zahlen sind konsistent, das Ergebnis geht in die Kalkulation ein. Was fehlt, ist die aktuarielle Kontrolle, die prüft, ob die Annahmen tragen und ob das Modell nicht bestimmte Risiken systematisch unterschätzt. Fällt eine solche Verzerrung erst, aufgrund von Statistiken, im Bestand auf, ist der Tarif längst verkauft. Dazu kommt eine grundsätzliche Grenze. Historische Daten sagen die Zukunft nicht immer zuverlässig voraus, zumal sich die Rahmenbedingungen schnell ändern, etwa durch Online-Versicherer, verändertes Verhalten jüngerer Generationen, die Alterspyramide oder wirtschaftliche Krisen.

Dasselbe gilt für die Schadenbewertung. Eine KI ordnet eingehende Schäden ein, schätzt Regulierungshöhen, priorisiert Fälle. Schnell, entlastend, plausibel. Aber ob die Einordnung im Einzelfall fachlich stimmt, ob Deckung, Kausalität und Höhe zusammenpassen, kann nur die fachliche Prüfung beurteilen. Keinesfalls ein Modell, das nach Wahrscheinlichkeit sortiert.

Unter VAIT und Solvency II ist diese Prüfung nicht optional. Beide Rahmenwerke verlangen nicht nur ein Ergebnis, sondern den Nachweis, dass seine Angemessenheit kontrolliert wurde. Eine Tarifierung ohne aktuarielle Verifikation und eine Schadenbewertung ohne fachliche Kontrolle erfüllen das nicht. So überzeugend das einzelne Ergebnis auch aussehen mag.

Damit schließt sich das Bild über beide Branchen. In der Bank fehlte die Modellvalidierung, in der Versicherung die aktuarielle und fachliche Verifikation. Der Rahmen ist ein anderer, das Prinzip dasselbe. Nachweisbarkeit ist der Grund, warum diesen Organisationen überhaupt jemand sein Geld anvertraut, und genau sie höhlt eine ungeprüfte KI-Beschleunigung aus.

Der falsche und der richtige Reflex

Der naheliegende Reflex nach so einem Vorfall ist das Verbot. KI-Werkzeuge werden gesperrt, Zugänge geschlossen, die Nutzung untersagt. Das wirkt entschlossen und ist doch der schlechtere Weg. Ein Verbot beseitigt nicht den Grund, aus dem die Kreditabteilung das Werkzeug gebaut hat. Der Druck, schneller zu liefern, bleibt. Die Fähigkeit, sich selbst zu behelfen, bleibt auch. Also wandert die Nutzung dorthin, wo sie niemand mehr sieht, und aus einer bekannten Schatten-KI wird eine verborgene. Das Problem wird nicht kleiner, nur unauffälliger.

Der richtige Reflex setzt nicht am Werkzeug an, sondern an der Prüfung. Nicht jede KI-Nutzung braucht dieselbe Kontrolle. Beim Sortieren von Daten genügt die schnelle Gegenprüfung. Kritisch wird es dort, wo aus einem KI-Ergebnis eine Entscheidung mit Folgen wird, eine Kreditzusage, ein Tarif, eine Schadenregulierung. Hier gehört die Verifikation fest in den Prozess, nicht als nachgelagerte Bremse, sondern als Teil des Wegs, den ein Ergebnis nehmen muss, bevor es zählt.

Das ist keine organisatorische Zutat, sondern eine Frage der Verantwortung. Es braucht eine benannte Stelle, die dafür einsteht, dass ein KI-Ergebnis geprüft wurde, bevor es wirkt. Governance entsteht dann nicht als Schleife nach der Arbeit, sondern als Teil der Arbeit selbst. Die KI liefert den Vorschlag schneller. Ob er zählt, entscheidet weiterhin ein Mensch, der dafür geradesteht.

Drei Fragen, die sich jede Organisation stellen sollte

Diese Fallstudie will nicht verkaufen, sondern klären. Drei Fragen helfen, die eigene Lage einzuschätzen:

1. Wo entsteht in Ihrer Organisation heute KI-Output, der in Entscheidungen einfließt, die langfristig tragfähig sein müssen?
2. Wer prüft diesen Output, bevor er verbindlich zählt?
3. Woran würden Sie merken, dass diese Prüfung nur formal stattfindet und nicht inhaltlich?

Die dritte Frage ist die unbequemste. Eine formal erfüllte, inhaltlich leere Prüfung ist gefährlicher als gar keine, weil sie Sicherheit vortäuscht, ohne sie zu geben. Genau hier setzt incowia an. Der Beschleuniger ist die Maschine, die Bremse bleibt menschlich. Und eine Prüfung gehört an die richtige Stelle der Schleife, dorthin, wo aus einem plausiblen Vorschlag eine Entscheidung mit Folgen wird.

Deutsche Fassung oben

When AI delivers faster than it can be checked

A case study on shadow AI, the verifier gap, and why verifiability is non-negotiable in regulated industries.

The core point up front

More insight is not more verified insight. Artificial intelligence makes generating cheap; checking stays expensive. Whoever accelerates only the generating pushes the risk to the back, to where it becomes most costly. Our case study shows, through a real example from the financial sector, how a well-intended, productive use of AI turned into a regulatory emergency, and why the answer is not a ban but verification in the right place.

Why shadow AI emerges

The term shadow IT dates back to the nineteen-eighties and described systems that arose in business units around central IT, because the official channels were too slow. Shadow AI is nothing fundamentally new, just the same pattern one step further. What is new is how easily such solutions emerge today, and how convincing their results look.

Behind it lies not negligence but rationality. AI tools are often freely available in the business unit and deliver in minutes what the standard procedure, including the official review, needs weeks for. Whoever is under deadline pressure and must deliver takes the faster route. In doing so, the business unit does not act against the organization, but in service of its own goals. That is what makes shadow AI so hard to grasp. It is not a rule violation out of convenience, but the obvious answer to a real bottleneck.

The case

At a company in the financial sector with a complex lending structure, the corporate credit department built its own solution. A large language model ran in the browser, connected through a cloud interface and supplemented by self-written Python code. With it, the department automatically summarized company data such as annual reports, balance sheets, and risk reports, and prepared it for assessing creditworthiness. For the case workers this was an enormous gain. What used to cost hours of concentrated reading now lay ready in minutes.

The impulse was understandable, the solution worked, and it was used. Over weeks, its analysis fed into real credit decisions, without anyone outside the department knowing about it.

Later, an internal audit noticed that the officially submitted data and the internal records did not match, above all in the debt levels of the assessed companies. The cause lay in the tool. The AI agent had misread certain table formats and, rather than reporting the resulting discrepancies, had smoothed them over mathematically. The result looked coherent precisely because the error had been made invisible. That is the treacherous part of fluently phrased AI output. It appears certain, even where it is not.

The consequences were severe. Credit risks were assessed incorrectly, and regulatory due-diligence obligations were thereby violated. The danger of credit defaults had risen unnoticed and, in an unfavorable course, could have brought the company into a serious imbalance. Every credit approved in that period had to be rechecked by hand. An extensive audit of the internal processes followed, along with strict requirements from the supervisory authority and internal disciplinary proceedings, because the prescribed model validation had been bypassed completely.

The core of the problem was not AI. It lay in the fact that several controls were bypassed at once:

- risk management, that is, Risk Management and Model Risk Governance
- IT compliance under the applicable regulatory requirements
- architecture governance, including compliance with the applicable requirements and principles

In Germany these include the MaRisk and the IT requirements now overlaid by DORA, supplemented by the EU AI Act, which classifies AI-supported creditworthiness assessment as high risk, though its obligations do not become binding until December 2027.

Innovation is not the problem; the missing review is. Whether an organization develops new solutions, streamlines processes, or designs products is not in question. What matters is that review instances are anchored in the process, instances that put AI-supported results to the test. This is not a bureaucratic hurdle, but a safety net that catches regulatory, security, and data-protection requirements at once.

The pattern behind it: the verifier gap

At its core lies a conflict between two speeds; generating has become fast, while checking stays slow. We call this gap the verifier gap. At first glance the case looks like an isolated failure, a department that meant too well. In truth it shows a pattern that grows more frequent with every better AI tool. Generating has become cheap. A plausible proposal, a phrased assessment, a finished analysis emerge in seconds. Checking whether the result is also correct has not become cheaper. It takes as long as before, sometimes longer.

Between these two speeds a gap opens. The AI produces faster than the organization can check. Where a person used to work out a proposal and, in doing so, already checked it, the AI delivers the finished result without delivering the check along with it. What looks like

a productivity gain only shifts the bottleneck. It moves from generating to checking, and there piles up what was previously spread evenly.

What is decisive is what a result can be checked against at all. AI is strong and reliable where a cheap, fast control can confirm or refute its proposals immediately. In opening up and structuring data this is the case; the result can be cross-checked in seconds. For a creditworthiness assessment, a pricing calculation, an architecture decision, this fast control is missing. Whether the statement holds shows not immediately, but only in the portfolio, in the audit, in the claim, sometimes only after years. Precisely here, acceleration produces not more verified insight, but more unverified.

Why it becomes critical in regulated industries

What happened to the credit department does not affect only one industry. The same pattern arises wherever AI delivers a plausible result and the checking behind it does not keep pace. In the insurance world it simply shows up in a different place.

An actuarial function deploys an AI tool to calculate tariffs faster. The model processes historical claims data, portfolio characteristics, and market comparisons, and delivers in minutes a tariff proposal that a case worker used to need days for. The proposal looks coherent, the numbers are consistent, the result feeds into the calculation. What is missing is the actuarial control that checks whether the assumptions hold and whether the model does not systematically underestimate certain risks. If such a distortion is noticed only later, through the statistics, in the portfolio, the tariff has long since been sold. On top of this comes a fundamental limit. Historical data does not always predict the future reliably, especially as the conditions change quickly, through online insurers, changing behavior of younger generations, the age pyramid, or economic crises.

The same holds for claims assessment. An AI classifies incoming claims, estimates settlement amounts, and prioritizes cases. Fast, relieving, plausible. But whether the classification is correct in the individual case, whether coverage, causation, and amount fit together, only the expert review can judge. By no means a model that sorts by probability.

Under VAIT and Solvency II this review is not optional. Both frameworks require not merely a result, but proof that its appropriateness was checked. A pricing calculation without actuarial verification and a claims assessment without expert control do not meet this. However convincing the individual result may look.

With this the picture closes across both industries. In the bank the model validation was missing, in the insurer the actuarial and expert verification. The framework is a different one, the principle the same. Provability is the reason anyone entrusts their money to these organizations at all, and it is exactly what an unverified AI acceleration hollows out.

The wrong reflex and the right one

The obvious reflex after such an incident is the ban. AI tools are blocked, access is closed, use is prohibited. This looks decisive and is nonetheless the worse path. A ban does not remove the reason the credit department built the tool. The pressure to deliver faster remains. The ability to improvise for oneself remains too. The use wanders to where no one sees it anymore, and a known shadow AI becomes a hidden one. The problem does not get smaller, only less conspicuous.

The right reflex sets in not at the tool but at the check. Not every use of AI needs the same control. For sorting data, the fast cross-check is enough. It becomes critical where an AI result turns into a decision with consequences, a credit approval, a tariff, a claims settlement. Here the verification belongs firmly in the process, not as a downstream brake but as part of the path a result must take before it counts.

This is not an organizational add-on but a question of responsibility. It takes a named position that stands behind the fact that an AI result was checked before it takes effect. Governance then arises not as a loop after work, but as part of the work itself. AI delivers the proposal faster. Whether it counts is still decided by a human who stands behind it.

Three questions every organization should ask itself

This case study does not want to sell, but to clarify. Three questions help in judging one's own situation:

1. Where in your organization does AI output arise today that feeds into decisions which must hold up over the long term?
2. Who checks this output before it counts as binding?
3. How would you notice that the check takes place only formally and not substantively?

The third question is the most uncomfortable. A formally fulfilled but substantively empty check is more dangerous than none at all, because it feigns security without giving it. This is exactly where incowia comes in. The accelerator is the machine; the brake stays human. And a check belongs in the right place in the loop, where a plausible proposal turns into a decision with consequences.