Fairness, Transparenz und Verantwortung: Ethische Rahmenbedingungen in der automatisierten Betrugsdetektion

Onlinebetrug ist ein weltweit wachsendes Problem, das auch in Österreich stark zunimmt. 2024 verzeichnete das Bundeskriminalamt 31.768 Fälle von Onlinebetrug, was einen Anstieg von 50% innerhalb von 5 Jahren bedeutet (Bundesministerium für Inneres 2025). Durch Fake-Shops, betrügerische Investmentplattformen oder auch Phishing-Fallen droht nicht nur das unmittelbare Risiko finanzieller Verluste, auch Identitätsdiebstahl und Datenmissbrauch nehmen in der Welt des Online-Betrugs zu. Hinzu kommt, dass generative Künstliche Intelligenz Betrugsfallen zusätzlich professionalisieren und damit das Erkennen von Betrug erheblich erschwert.

Die Watchlist Internet, die größte Präventionsplattform gegen Internetbetrug im deutschsprachigen Raum, erhält jährlich tausende Meldungen von Nutzer:innen – allein 2025 waren es mehr als 15.000. Diese manuelle Informationsbasis wird zunehmend durch technische Toolings ergänzt, etwa durch den Fake-Shop Detector (FSD). Der FSD analysiert Domains automatisiert und bewertet sie anhand eines KI-basierten Risikoscores, der auf Daten aus Webcrawlern, Scraping-Tools und Nutzer:innenmeldungen beruht. So entsteht ein hybrides System, in dem automatisierte Prozesse die Arbeit von Expert:innen entlasten, gleichzeitig aber menschliche Kontrolle erhalten bleibt.

Automatisierte Systeme wie der FSD eröffnen neue Möglichkeiten zur Skalierung von Betrugsprävention. Gleichzeitig werfen sie grundlegende ethische und organisatorische Fragen auf. Vor diesem Hintergrund untersucht die vorliegende Studie am Beispiel der Watchlist Internet, wie ethische Prinzipien in der (teil-)automatisierten Betrugsdetektion gewährleistet werden können. Ziel ist es, Best Practices entlang folgender drei Schwerpunkte zu identifizieren:

1. Die Entwicklung von KI-Ethik-Leitlinien für den Fake-Shop Detector und den Shop-Check.
2. Die Konzeption einer Model Card zur transparenten Dokumentation der Systemarchitektur und Entscheidungslogik (siehe 5. Anhang).

3. Die Evaluation der manuellen Qualitätssicherung (Human in the Loop), die insbesondere arbeitsrechtliche und ethische Fragen bei der Zusammenarbeit mit Data Worker

1. Ethische KI-Leitlinien

Der Fake-Shop Detector (FSD) ist ein KI-gestütztes System zur Identifikation betrügerischer Online-Shops, das seit 2018 im Rahmen mehrerer Forschungsprojekte (MAL2, KOSOH, SINBAD, DETECT, RIO) am Austrian Institute of Technology (AIT), dem Österreichischen Institut für angewandte Telekommunikation (ÖIAT – insbesondere in Zusammenarbeit mit der Initiative Watchlist Internet) und X-NET entwickelt wurde. In Form eines Plugin warnt der Fake-Shop Detector rechtzeitig vor betrügerischen Webshops, direkt im Browser. Zusätzlich ermöglicht der Shop-Check eine Abfrage einzelner Domains ohne Plugin-Installation.

Die technologische Grundlage des FSD bildet ein Machine-Learning-Ansatz, der betrügerische Webshops anhand struktureller Merkmale im Quellcode identifiziert. Dabei erreicht das System Genauigkeitsraten von bis zu 97 % auf den Trainingskorpus, in der aktuellen Implementierung liegt die Genauigkeit von 90,38 % (Beltzung u. a. 2020).

Seit etwa 2020 ist der FSD im Einsatz und wurde von rund 13.000 Nutzer:innen installiert. Gleichzeitig unterstützt er – gekoppelt durch weitere Tools wie Crawler und Scraper – die Redaktion der Watchlist Internet bei der teilautomatisierten Betrugsdetektion. Mit diesen zunehmenden Einsatz des FSD gewinnt die ethische Gestaltung der KI-basierten Betrugsdetektion an Bedeutung.

Die Europäische Kommission hat 2019 durch eine unabhängige Expertengruppe sieben Kerndimensionen vertrauenswürdiger KI definiert. Im Folgenden wird untersucht, inwieweit der Fake-Shop Detector diese ethischen Anforderungen umsetzt und in welchen Bereichen Verbesserungsbedarf notwendig ist.

1.1. Vorrang menschlichen Handels und menschlicher Aufsicht

> *„KI-Systeme sollten Menschen in die Lage versetzen, fundierte Entscheidungen zu treffen und ihre Grundrechte zu fördern. Gleichzeitig müssen für angemessene Aufsichtsmechanismen gesorgt werden, die durch menschliche Ansätze in den Kreisläufen, Menschen in den Kreisläufen*

und „Human-in-the-Loop"-Konzepte sowie durch menschen-in-command-Konzepte erreicht werden können."

Der FSD setzt diese Leitlinie durch mehrere ineinandergreifende Mechanismen um: Die Kommunikation der Systementscheidungen erfolgt sowohl im Browser-Plugin als auch im Shop-Check transparent und verständlich. Nutzerinnen und Nutzer erhalten klar gekennzeichnete Risikobewertungen, die zwischen KI-gestützten Einschätzungen, manuell verifizierten Ergebnissen sowie zusätzlichen Kontextinformationen (Impressumsdaten oder Trustpilot-Reviews) unterscheidet.

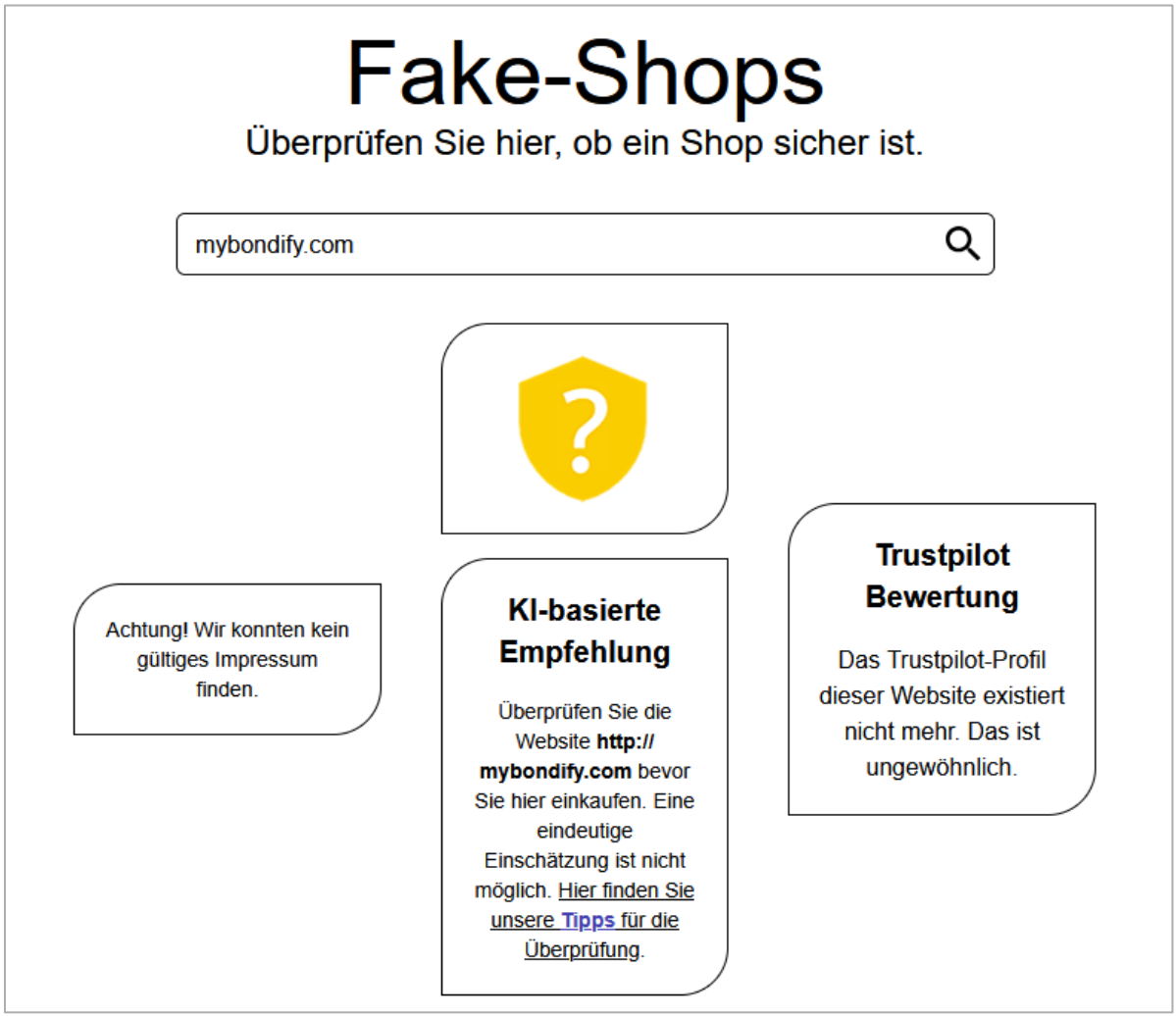

Abbildung 1: Screenshot von https://www.fakeshop.at/shopcheck/

Ein zentrales Element ist zudem die menschliche Qualitätssicherung, die insbesondere durch die Expert:innen der Watchlist Internet erfolgt, in Testszenarien auch durch den gezielten und fairen Einsatz von Data Workern (siehe Kapitel 2 „Manuelle Qualitätssicherung"). Dies gewährleistet eine kontinuierliche menschliche Aufsicht im Analyseprozess und folgt dem Evaluationsverfahren mit Human-in-the-Loop Ansätzen, welche bereits während der

Modellentwicklung etabliert wurde. (Beltzung u. a. 2020). Die Bewertung durch KI-Modelle und menschlichen Expert:innen ermöglicht nicht nur Qualitätssicherung, sondern auch eine kontinuierliche Verbesserung des Systems durch Relearning. Sollte das System eine Fehleinschätzung treffen, wird diese durch das Team schnellstmöglich manuell korrigiert

1.2. Technische Robustheit und Sicherheit

> *„KI-Systeme müssen widerstandsfähig und sicher sein. Sie müssen sicher sein, einen Rückfallplan für den Fall gewährleisten, dass etwas schiefgeht, sowie präzise, zuverlässig und reproduzierbar sein. Nur so kann sichergestellt werden, dass auch unbeabsichtigte Schäden minimiert und verhindert werden können."*

Der FSD erreicht in der aktuellen Implementierung eine Genauigkeit von 90,38%[1]. Dies bildet eine robuste Grundlage für zuverlässige Entscheidungen. Die 2020 dokumentierten Ergebnisse zeigen, dass das aggregierte Modell aus XGBoost und Random Forest für 54% der Fake-Shops Warnungen auf dem höchsten Konfidenzniveau ausgeben konnte. Die detektierten Fake-Shops konnten sofort auf Warnlisten gesetzt werden – ohne eine einzige Fehlklassifizierung (Beltzung u. a. 2020).

Die technische Architektur wurde zudem in mehreren Forschungsprojekten kontinuierlich verbessert. Besonderer Wert wurde dabei auf die Minimierung der False-Positive-Raten gelegt, da falsche Warnungen bei legitimen Shops erheblichen wirtschaftlichen Schaden anrichten können. Das mehrschichtige Warnsystem mit unterschiedlichen Konfidenzleveln ermöglicht eine differenzierte Risikokommunikation und trägt zur Präzision und Zuverlässigkeit des Systems bei.

Für den Fall eines Systemausfalls wurde ein Rückfallplan implementiert, bei den Nutzer:innen den Hinweis „Manuelle Prüfung läuft" erhalten und so Transparenz geschaffen wird. Bei Fehleinschätzungen können Nutzer:innen als auch Betroffene (z. B. Online-Shop-Betreiber:innen, die mit einem hohen Risiko bewertet werden) direkt über das Plugin eine Fehlermeldung übermitteln. Das Team verpflichtet sich, innerhalb von maximal 72 Stunden mit einer Rückmeldung und ggf. mit einer manuellen Korrektur zu reagieren.

1.3. Privatsphäre und Datenqualitätsmanagement

[1] https://www.fakeshop.at/forschung/kuenstliche-intelligenz/

„Neben der Gewährleistung der uneingeschränkten Achtung der Privatsphäre und des Datenschutzes müssen auch angemessene Daten-Governance-Mechanismen sichergestellt werden, die der Qualität und Integrität der Daten Rechnung tragen und einen legitimierten Zugang zu den Daten gewährleisten."

Bei der Entwicklung des FSD wird ein Privacy-by-Design Ansatz verfolgt: Auch wenn der Echtzeitanalyse des FSD zugestimmt wurde, wird das Such- oder Einkaufsverhalten nicht getrackt. Erfasst werden ausschließlich Website-URLs sowie der Bewertungszeitpunkt der jeweiligen URL, sofern diese noch nicht in der Datenbank des FSD verzeichnet ist. Personenbezogene Daten der Plugin-Nutzer:innen werden weder erhoben noch gespeichert.

Ein zusätzlicher Schutz der Privatsphäre erfolgt durch offline und lokal durchgeführte Risikobewertungen. Somit wird selbst im Fall eines Angriffs auf das Backend-Systems die Erstellung eines Profils einzelner User nicht möglich gemacht.

Da mit ebensolchen Angriffen auf den FSD gerechnet werden muss, ist es notwendig in den Log-Files die IP-Adressen zu Serveranfragen zu erfassen – nur so kann gezielt mit Sperren gegen entsprechende Angriffe vorgegangen werden. Diese Log-Files werden weder mit anderen Datenquellen zusammengeführt noch an Dritte weitergegeben.

Ein besonderer Fokus liegt zudem auf der Datenqualität. Die Trainingsdaten für den FSD wurden manuell kuratiert und durchlaufen eine strenge Qualitätssicherung durch das Team der Watchlist Internet. Aktuell kennt der FSD knapp 23.000 betrügerische Onlineshops und rund 26.000 vertrauenswürdige Onlineshops. Der Prozess der Datenerhebung und -annotation basierte von Beginn an auf einem standardisierten Checklisten-Verfahren, womit die Modelle auf qualitativ hochwertigen und verlässlichen Daten trainiert werden, was gleichzeitig die Integrität und Legitimität der gesamten Daten-Governance stärkt.

1.4. Transparenz

„Die Geschäftsmodelle für Daten, Systeme und KI sollten transparent sein. Rückverfolgbarkeitsmechanismen können dazu beitragen. Darüber hinaus sollten KI-Systeme und ihre Entscheidungen in einer Weise erläutert werden, die an die betroffenen Interessenträger angepasst ist. Die Menschen müssen sich dessen bewusst sein, dass sie mit einem KI-System interagieren, und sie müssen über die Fähigkeiten und Grenzen des Systems informiert werden."

Den Nutzer:innen des FSD wird klar kommuniziert, wann es sich um eine KI-basierte Entscheidung handelt und wann eine Bewertung auf menschlichen Einschätzungen (kuratierte White- und Blacklists) basiert. Im Shop-Check werden zusätzlich Kontextinformationen transparent dargestellt (Impressums-Check und Trustpilot-Review-Analysen). Die Systemgrenzen werden zudem in einer Model Card dokumentiert, die technische und nicht-technische Stakeholder über Limitationen informiert.

Eine besondere Herausforderung im Bereich Betrugsprävention und -detektion ist, dass zu detaillierte Transparenz hinsichtlich der verwendeten Tools und Systeme von den Kriminellen missbraucht werden könnte, um ihre betrügerische Angebote zu optimieren bzw. um Strategien zu entwickeln, um der automatisierten Detektion zu entgehen (Fenaux u. a. 2025; Nesvijevskaia u. a. 2021). Der FSD begegnet diesem Spannungsfeld zwischen Transparenzanforderungen und Sicherheitserwägungen mit einem gestuften Kommunikationsansatz: Öffentlich zugängliche Informationen beschränken sich auf allgemeine Warnungen und Hinweise auf verwendete Datenquellen. Behörden und Kooperationspartner erhalten erweiterte Informationen zu technischen Indikatoren.

Hinsichtlich Explainability wurden Methoden wie LIME (Local Interpretable Model-Agnostic Explanations) und SHAP (Shapley Additive Explanations) genutzt, mit denen die wichtigsten Features identifiziert werden konnten, die zu einer bestimmten Risikoeinschätzung führen. Die Analysen zeigen, dass keine einzelnen definitiven Features existieren, die eindeutig auf Betrug hinweisen, sondern dass vielmehr eine Kombination vieler Features mit geringem Gewicht zusammenwirkt. Diese Erkenntnisse sind wiederum im Experten-Dashboard integriert, und dadurch intern abruf- und nachvollziehbar.

1.5. Vielfalt, Nichtdiskriminierung und Fairness

> *„Unfaire Verzerrungen müssen vermieden werden, da sie vielfältige negative Auswirkungen haben könnten, von der Marginalisierung schutzbedürftiger Gruppen bis hin zur Verschärfung von Vorurteilen und Diskriminierungen. Um Vielfalt zu fördern, sollten KI-Systeme für alle zugänglich sein, unabhängig von einer Behinderung, und die einschlägigen Interessenträger während ihres gesamten Lebenszyklus einbeziehen."*

Der FSD begegnet dem Risiko systematischer Verzerrungen durch die bereits beschriebene menschliche Qualitätssicherung. Die manuelle Überprüfung aller Systemausgaben durch die Expert:innen der Wachtlist Internet wirkt als Korrektiv. Eine noch ungeklärte Frage ist jedoch, ob die Trainingsdaten selbst systematische Verzerrungen enthalten, etwa durch die

Unterpräsentation bestimmter Arten betrügerischer Shops sowie der Auswahl der Whitelists. Hier besteht noch Nachholbedarf.

Hinsichtlich Barrierefreiheit wurde eine Kombination aus Maßnahmen umgesetzt. Für sehbeeinträchtigte Nutzer:innen wurde auf die Einhaltung ausreichender Kontrastfarben geachtet. Buttons und grafische Elemente sind visuell klar gestaltet. Für Menschen mit Farbsehschwäche wurde dem Ampelsystem, das die Risikostufe anzeigt, zusätzlich Symbole zugewiesen sowie Text hinzugefügt, die das Risiko auch ohne Farbwahrnehmung erkennbar machen. Nutzer:innen, die aufgrund von Screenreadern keine Plugins verwenden, haben Zugang zum System über den Shop-Check auf der Website fakeshop.at.

Die technische Architektur von Browser-Plugins weist jedoch grundsätzlich Einschränkungen hinsichtlich vollständiger Barrierefreiheit auf. Eine systematische Evaluation nach dem Web Content Accessibility Guidelines (WCAG 2.1) könnte weitere Verbesserungspotenziale identifizieren und sollte in zukünftigen Entwicklungszyklen berücksichtig werden.

1.6. Gesellschaftliches und ökologisches Wohlergehen

> *„KI-Systeme sollten allen Menschen, auch künftigen Generationen, zugutekommen. Daher muss sichergestellt werden, dass sie nachhaltig und umweltfreundlich sind. Darüber hinaus sollten sie der Umwelt, einschließlich anderer Lebewesen, Rechnung tragen, und ihre sozialen und gesellschaftlichen Auswirkungen sollten sorgfältig geprüft werden."*

Der gesellschaftliche Nutzen des FSD zeigt sich in seinem Beitrag zu einem sicheren Internet für alle Bevölkerungsgruppen. Die Warnungen schützen Konsument:innen vor finanziellem, aber auch emotionalem Schaden durch betrügerische Online-Shops. Dies betrifft insbesondere vulnerable Personen – wobei im Betrugskontext von einer situativen Vulnerabilität ausgegangen werden muss. Betroffene können demnach je nach Situation vulnerabel sein (Europäische Kommission 2016).

Ökologische Nachhaltigkeit wird zudem durch Datenminimierung adressiert: Websites, die bereits in der Datenbank enthalten sind, werden nicht erneut analysiert und die Integration externer Datenquellen wie Blacklists und Whitelists ermöglichen eine Verringerung der Analyseleistung. Shops, die auf vertrauenswürdigen Whitelists oder Blacklists gelistet sind, werden direkt als sicher klassifiziert, ohne dass rechenintensive Prozesse notwendig sind.

1.7. Rechenschaftspflicht

„Es sollten Mechanismen geschaffen werden, die die Verantwortlichkeit und Rechenschaftspflicht für KI-Systeme und deren Ergebnisse gewährleisten. Die Überprüfbarkeit, die die Bewertung von Algorithmen, Daten und Entwurfsprozessen ermöglicht, spielt dabei eine Schlüsselrolle, insbesondere bei kritischen Anwendungen. Darüber hinaus sollte sichergestellt werden, dass angemessene Rechtsbehelfe zur Verfügung stehen."

Durch die Definition von klaren Verantwortlichkeiten wird sichergestellt, dass bei Problemen rasch die richtigen Ansprechpartner identifiziert werden können. Das ÖIAT, das AIT und X-Net tragen gemeinsam die Projektverantwortung und teilen sich inhaltliche und technische Verantwortlichkeiten auf, während für die operative Bearbeitung von Nutzer:innenanfragen und für die manuelle Qualitätssicherung die Expert:innen der Watchlist Internet zuständig sind.

Falsche Entscheidungen werden nach einer Meldung innerhalb von maximal 72 Stunden manuell korrigiert. Diese Reaktionszeit ist ein wichtiger Indikator für die Rechenschaftspflicht gegenüber den Nutzer:innen. Das Browser-Plugin verfügt dafür über einen prominent platzierten Button „Fehler melden", der die Meldeschwelle niedrig hält und schnelle Korrekturen ermöglicht.

Zudem wurden 2020 die Methodik sowie transparente Evaluationsergebnisse publiziert, um die Überprüfbarkeit des Systems zu gewährleisten. Diese Publikation sowie weitere Informationen, wie unter anderem die für den FSD verwendeten Quellen, sind auf fakeshop.at zugänglich. Die öffentlich zugängliche REST-API basiert auf einer OpenAPI-Spezifikation, die Transparenz über die technischen Schnittstellen schafft.

1.8. Fazit

Die systematische Bewertung des Fake-Shop Detectors entlang der von der Europäischen Kommission definierten ethischen KI Leitlinien zeigt, dass das System in den meisten Dimensionen ausreichende Ansätze zur Umsetzung einer vertrauenswürdigen KI implementiert hat. Besondere Stärken liegen in der Integration menschlicher Aufsicht, der technischen Robustheit mit Fehlerraten unter zehn Prozent sowie dem konsequenten Datenschutz.

Verbesserungspotenzial besteht insbesondere im Bereich der Transparenz, da das Spannungsfeld zwischen öffentlicher Erklärbarkeit und Sicherheitserwägungen weiterhin eine Herausforderung darstellt. Die Barrierefreiheit sollte durch systematische WCAG-2.1-Evaluierung überprüft und optimiert werden. Bei der Vermeidung systematischer

Verzerrungen könnten erweitere Bias-Analysen der Trainingsdaten selbst zusätzliche Erkenntnisse liefern.

Die kontinuierliche Weiterentwicklung im Rahmen unterschiedlicher Forschungsprojekte zeigt das Bestreben, nicht nur die technische Leistungsfähigkeit, sondern auch die ethische Gestaltung des FSD zu verbessern. So zielte die Integration von Gamification-Elementen im Projekt DETECT darauf ab, die Community stärker einzubinden und damit die menschliche Aufsicht auf eine noch breitere Basis zu stellen. Das Projekt RIO adressierte wiederum gezielt die Resilienz im Online-Handel und untersuchte erstmals den Einsatz von Data Worker im Kontext der automatisierten Betrugsdetektion.

Der Fake-Shop Detector demonstriert so, dass KI-gestützte Systeme zur Betrugsdetektion bei sorgfältiger ethischer Gestaltung einen substanziellen Beitrag zum Konsumentenschutz und der Betrugsdetektion leisten können, ohne dabei fundamentale Rechte oder Sicherheitsinteressen zu gefährden.

2. *Manuelle Qualitätssicherung*

Um Fehlentscheidungen in der automatisierten Betrugsdetektion zu vermeiden, ist eine kontinuierliche Qualitätssicherung dieser Entscheidungen erforderlich. In der Praxis erfolgt diese Einbindung menschlicher Expertise über das Prinzip des Human-in-the-Loop (HITL). Dabei werden Menschen in automatisierte und KI gesteuerte Entscheidungsprozesse eingebunden, um deren Qualität, Fairness und Nachvollziehbarkeit zu sichern. Entsprechend werden auch die maschinellen Bewertungen des Fake-Shop Detector durch Expert:innen der Watchlist Internet überprüft, kommentiert und korrigiert. Diese Rückmeldungen fließen wiederum in die Trainingsdaten und verbessern die Modelle kontinuierlich.

Mit wachsender Datenmenge und den vielfältigen Quellen (Crawler, Scraper, Meldungen von Nutzer:innen) stößt dieser Ansatz jedoch an personelle und organisatorische Grenzen. Um die Qualitätssicherung bei wachsender Datenmenge bewerkstelligen zu können, wurde daher in der Vergangenheit auf zusätzliche externe Expertise durch Crowdsourcing-Dienste zurückgegriffen. Dabei können einzelne Prüfaufgaben – etwa die Beurteilung verdächtiger Domains nach definierten Kriterien – an eine große Zahl von Online-Arbeitskräften, sogenannten Data Worker[2] – ausgelagert werden. Diese Microtasks bilden einen zentralen

[2] Oftmals wird für diese Art der Tätigkeit auch der Begriff *Clickworker* verwendet, der in wissenschaftlichen Debatten teils kritisch gesehen wird, da er die Arbeit auf einfache Klick-Tätigkeiten reduziert und die Komplexität der Arbeit unterschätzt und abwertet. Aus diesem Grund verwenden wir im Folgenden den neutraleren Begriff *Data Worker*.

Bestandteil vieler KI-Trainings- und Qualitätssicherungsprozesse, ermöglichen Skalierbarkeit und schnelle Bearbeitung, bringen jedoch neue arbeitsrechtliche und ethische Herausforderungen mit sich.

2.1. Einbindung von Data Worker in der Betrugsdetektion

Um die Einbindung von Data Worker in die Qualitätssicherung des Fake-Shop Detectors zu erproben, wurden spezifische Workflows entwickelt, die eine Qualitätssicherung durch Crowdarbeit abbilden. Die Workflows ergaben sich aus Testdurchläufen, die im Vorgängerprojekt „RIO" durchgeführt wurden. Ziel war es dabei zu evaluieren, welche Aufgaben innerhalb der Betrugsbewertung von Websites durch Data Worker sinnvoll übernommen werden können. Im aktuellen Forschungsprojekt soll darüber hinaus der Einsatz von Data Worker explizit im Hinblick auf ethische Fragestellungen untersucht werden, der in diesem Bericht im Fokus steht.

Exkurs: Taskvergabe im Kontext Betrugsdetektion von Data Worker

In mehreren Testdurchläufen wurden 400 Domains überprüft, wobei pro Bewertung eine Vergütung von 5 bis 15 Cent vorgesehen war. Zur Qualitätssicherung wurden jeweils zwei bis drei unabhängige Data Worker pro Domain eingesetzt. Die Erprobung machte deutlich, wo die Stärken und Limitationen liegen: Standardisierte Aufgaben, wie etwa die Erkennung von Warenkorb-Symbolen, ließen sich effizient auslagern. Bei komplexeren Einschätzungen – beispielsweise der Bewertung eines Impressums – traten jedoch wiederholt Verständnisschwierigkeiten auf. Häufig blieb unklar, was als Impressum (bzw. imprint) zu werten ist, insbesondere bei nicht deutschsprachigen Websites. Gleichzeitig zeigte sich, dass unklare Aufgabenstellungen oder widersprüchliche Beispiele nicht nur die Ergebnisqualität beeinträchtigen, sondern auch zu längeren Bearbeitungszeiten führen. Diese zusätzlichen, unbezahlten „unsichtbaren Zeiten" verlagern den Aufwand einseitig auf die Data Worker und werfen damit zentrale ethische Fragen auf, die im Folgenden behandelt werden.

2.2. Crowdsourcing und Gig-Economy: Einordnung und Definition

Mit der Verbreitung digitaler Plattformen hat sich neben klassischen Beschäftigungsformen eine Gig-Economy etabliert. Ein zentraler Teilbereich ist Crowdwork: Komplexe Tätigkeiten werden in punktuelle, standardisierte Mikroaufgaben und Aufträge („Gigs“) zerlegt, die sich isoliert und ortsunabhängig erledigen lassen. Dabei lagern Auftraggeber (z. B. Unternehmen, Forschungsinstitute oder Organisationen) diese Tasks auf Plattformen aus, über deren Online-Marktplätze sie vermittelt und von einer großen globalen „Crowd“ bearbeitet werden. Aus Sicht der Auftraggeber:innen ermöglicht dieses Modell Just-in-time-Bereitstellung von Arbeitskraft, Skalierbarkeit und Kostensenkung; aus Sicht der Arbeitenden verspricht es zeitliche Flexibilität und niedrigschwelligen Zugang zu bezahlten Aufgaben. Gleichzeitig verschiebt diese Organisationsform Risiken und Transaktionskosten systematisch auf die Crowd und schafft neue Abhängigkeiten über digitale Reputations- und Bewertungssysteme. Die registrierten Data Worker führen kleinere, im Vorhinein festgelegte Tätigkeiten aus wie etwa Bildbeschriftungen, Datenannotation oder Datenbereinigung, meist im Cent- oder einstelligen Eurobereich (Lutz 2017). Formal sind sie freie Dienstleister:innen, tatsächlich aber in ein enges Regime algorithmischer Steuerung eingebunden.

2.3. Problemaufriss

Im Kontext der Industrialisierung entstanden regulierte Arbeitsverhältnisse - unbefristet, sozialversichert, kollektivvertraglich geregelt - als Antwort auf die asymmetrischen Machtverhältnisse zwischen Arbeitenden und Arbeitgeber:innen. Dieses Modell wurde mit der fortschreitenden Flexibilisierung und Deregulierung der Arbeitsmärkte ab den 1980er Jahren zunehmend durch atypische Beschäftigungsformen (Werkverträge, Leiharbeit, Freelancing) abgelöst. Damit wurden Risiken wie Krankheit oder Arbeitslosigkeit von der kollektiven Absicherung auf die individuelle Verantwortung der Arbeitenden selbst übertragen. Mit der Digitalisierung und Plattformökonomie verschärft sich diese Entwicklung. Ursula Huws (2001) beschreibt in ihrer Analyse des „Cybertariats“ die Herausbildung eines neuen digitalen Proletariats, das in hochgradig fragmentierten, entgrenzten und oft unsichtbaren Arbeitsprozessen tätig ist (Krichmayr 2018).

2.3.1. Scheinselbstständigkeit und Abhängigkeit

Digitale Plattformarbeit bewegt sich in einem arbeitsrechtlichen Graubereich: Data Worker sind weder klassische Arbeitnehmer:innen noch selbstständige Unternehmer:innen. Risak

(2017) zeigt, dass insbesondere das Dreiecksverhältnis zwischen Plattform, Auftraggeber:in und Data Worker strukturell dazu beiträgt, Verantwortung zu verschieben und arbeitsrechtlichen Schutz zu umgehen. Plattformen wie Clickworker, Mechanical Turk (Amazon) oder Appen deklarieren ihre Nutzer:innen als selbstständig, obwohl die konkreten Arbeitsbedingungen vielfach Merkmale abhängiger Beschäftigung aufweisen. Lutz (2017) arbeitet heraus, dass Abhängigkeit und Kontrolle über Plattformarchitekturen, Bewertungssysteme und algorithmische Sichtbarkeit hergestellt wird. Die Plattform behält sich einseitige Änderungen der AGB vor, entscheidet über Freischaltung, Qualifizierung und Zugang zu Aufgaben. Die eigentlichen Auftraggeber:innen bleiben für die Arbeitenden häufig anonym.

Obwohl die Tätigkeit formal als „frei" und flexibel gilt, bestehen faktisch kaum Gestaltungsspielräume. Es gibt keine reale Verhandlungsmacht über Vertragsbedingungen, keine organisierte Interessensvertretung und eine starke Bindung an vorgegebene Interfaces und Bewertungslogiken. Risak verweist in diesem Zusammenhang auf das funktionale Arbeitgeber:innenkonzept nach Prassl (2018), wonach Verantwortung dort anzusiedeln ist, wo tatsächliche Steuerung ausgeübt wird – in vielen Fällen also bei den Plattformen selbst.

2.3.2. Schlechte Vergütung

Ein zentrales strukturelles Problem digitaler Plattformarbeit ist die systematisch niedrige Vergütung. Bezahlungen erfolgen meist auf Stücklohnbasis und liegen häufig unter lokalen Mindestlöhnen – insbesondere dann, wenn unbezahlte Arbeitszeiten mitberücksichtigt werden. Dazu zählen Einschulungen, das Suchen und Auswählen von Aufgaben, das Erlernen neuer Interfaces sowie Korrekturen oder nicht akzeptierte Arbeiten. Die daraus resultierenden Mehrzeiten werden vollständig auf die Arbeitenden externalisiert.

Hinzu kommt eine gezielte globale Arbeitsverlagerung. Viele Plattformen rekrutieren gezielt in Niedriglohnländern oder wirtschaftlich prekären Regionen, vor allem im Bereich Data Work, um die KI-Entwicklung kostengünstig voranzutreiben. Gleichzeitig können Nutzer:innen faktisch „ausgehungert" werden, indem Plattformen ihnen keine Aufträge mehr zuteilen. Für Arbeitende entsteht so ein permanenter Wettbewerbsdruck (Hao 2025).

Recherchen zeigen etwa am Beispiel der katastrophalen ökonomische Krise Venezuelas, dass Plattformarbeit häufig als letzte Einkommensquelle genutzt wird, wenngleich unter Bedingungen extremer Unsicherheit und geringer Bezahlung. Mit ansteigender Anzahl der Data Worker aus Venezuela senkten Plattformen die Zahlungen pro Auftrag, sperrten Konten oder

stellten Bonus-Programme ein. Was in Venezuela begann, prägte die Erwartungen der KI-Branche hinsichtlich der Kosten solcher Dienstleistungen und schuf eine Strategie, um die von den Kunden erwarteten Preise zu erreichen (Hao und Hernández 2022). Diese globale Ungleichverteilung verstärkt bestehende Machtasymmetrien und untergräbt faire Arbeitsstandards.

2.3.3. Fehlende Absicherung und Unterstützung bei hoher Belastung

Neben niedriger Bezahlung fehlt Data Worker nahezu jede Form sozialer Absicherung. Es bestehen keine Ansprüche auf Kranken-, Pensions- oder Arbeitslosenversicherung, keine Entgeltfortzahlung im Krankheitsfall und kein Kündigungsschutz. Temporäre Shutdowns vonseiten der Plattformen in afrikanischen Ländern, etwa in Kenia, haben gezeigt, wie abrupt Einkommensquellen wegfallen können, ohne dass Betroffene Anspruch auf Unterstützung oder Entschädigung haben (Okinyi 2024). Plattformarbeit ist damit hochgradig krisenanfällig und verschiebt wirtschaftliche Risiken vollständig auf Individuen.

Diese fehlende Absicherung verstärkt nicht nur materielle Unsicherheit, sondern wirkt sich auch auf psychische Gesundheit, Planbarkeit und Lebensführung aus. Zusätzlich ist die Arbeit, vor allem Labeling und Content Moderation für KI Training, oft psychisch extrem belastend. Data Worker sehen sich mit Gewaltdarstellungen, pornographischen Inhalte oder ähnlichen konfrontiert. In einem Bericht von Gebrekidan (2024) geht hervor, wie die tägliche Auseinandersetzung mit solch erschütternden Inhalte und Bildern über längere Zeiträume zu Sucht oder anderem schädlichen Verhalten führt. Hinzu kommt, dass es entweder gar keine oder nur marginale Supervision bzw. psychologische Unterstützung gibt, um mit solchen Inhalten umzugehen.

Ethisch besonders problematisch ist die Verantwortungsdiffusion: Weder Plattform noch Auftraggeber:innen fühlen sich zuständig für faire Bezahlung, Datenschutz oder Arbeitsschutz. Diese Entkopplung von Handlung und Verantwortung steht im direkten Widerspruch zu den Leitlinien für vertrauenswürdige KI, wie sie die Europäische Kommission formuliert hat, in denen *Fairness, Transparenz und Rechenschaftspflicht* zentrale Prinzipien darstellen (Europäische Kommission 2019). Im Kontext automatisierter Betrugsdetektion stellt sich daher die Frage, wie menschliche Qualitätssicherung – also das „Human in the Loop“ – unter Bedingungen gestaltet werden kann, die den Schutz der Arbeitenden sichern, da auch in diesem Kontext von problematischen bis schädigenden Inhalten auszugehen ist.

2.4. Lösungen

Die gesellschaftliche und wissenschaftliche Debatte zu Fairness und Arbeitsrechten in der Plattformarbeit konzentriert sich bislang überwiegend auf die Plattformen selbst. Studien, Regulierungsansätze und Interessenvertretungen adressieren vorrangig deren Pflichten gegenüber den Data Workern – etwa in Bezug auf Entlohnung, Transparenz oder algorithmisches Management (Fairwork 2025; Richtlinie (EU) 2024/2831 des Europäischen Parlaments und des Rates vom 23. Oktober 2024 zur Verbesserung der Arbeitsbedingungen in der Plattformarbeit 2024). Diese wichtigen Schritte sollten ergänzt werden, indem auch die Rolle der Auftraggeber:innen kritisch beleuchtet wird.

Orientierung an bestehenden Standards: Fairwork und EU-Regulierung

Interessenvertretungen wie *Fairwork Austria* bzw. die internationale *Fairwork Foundation* etnwickelten Kriterien, die als Grundlage dienen können. Diese orientieren sich an fünf zentrale Prinzipien (Fairwork 2023):

- **Fair Pay:** Data Worker sollen ein existenzsicherndes Einkommen erzielen können, das alle Arbeitszeiten und Nebenkosten einschließt, und pünktlich bezahlt werden.
- **Fair Conditions**: Plattformen müssen Risiken wie gesundheitliche oder psychologische Belastungen aktiv minimieren und Schutzmechanismen einführen.
- **Fair Contracts:** Vertragsbedingungen sollen klar, zugänglich und an das nationale Recht angepasst sein; unfaire Haftungsausschlüsse oder intransparente Änderungen sind zu vermeiden.
- **Fair Management:** Entscheidungen über Sperrungen, Bewertungen oder Auftragsvergabe müssen nachvollziehbar sein; Algorithmen dürfen nicht diskriminierend wirken.
- **Fair Representation:** Data Worker müssen Mitsprache- und Organisationsmöglichkeiten haben, unabhängig von ihrem Beschäftigungsstatus.

Diese Prinzipien werden teilweise auch in der EU-Richtlinie zur Verbesserung der Arbeitsbedingungen in der Plattformarbeit (2024) aufgegriffen. Sie geht davon aus, dass Plattformarbeiter:innen als Arbeitnehmer:innen gelten, wenn die Plattform faktisch wie ein Arbeitgeber agiert. Die Plattform kann dies allerdings im Einzelfall widerlegen.

Verantwortung der Auftraggeber:innen

Auch wenn die Auftraggeber:innen formal nicht als Arbeitgeber:innen im Sinne der EU-Richtlinie gelten und diese rechtlichen Regelungen primär auf die Plattformen abzielen, können sie dennoch Standards aus diesen Rahmenwerken übernehmen. Einige der Forderungen betreffen eindeutig die Plattformen, da Auftraggeber:innen nur begrenzten Einfluss auf deren Architektur und Steuerung haben, etwa in Bereichen wie faires Management, Transparenz oder die Repräsentation der Crowd. Andere Aspekte liegen hingegen sehr wohl im Handlungsspielraum der Auftraggeber:innen, sodass sie durch eigene Maßnahmen zur Umsetzung fairer Standards beitragen können, zum Beispiel durch klare Aufgabenbeschreibungen, angemessene Bezahlung oder Qualitätssicherungsmechanismen.

- **Faire Vergütung:** Entlohnung sollte nicht nur die sichtbare Bearbeitungszeit, sondern auch Vorbereitungs-, Lade- und Suchzeiten berücksichtigen. Es bestehen Forderungen, das Gehalt am Arbeitgeberland zu orientieren, und nicht am Herkunftsland der Arbeitenden. Zudem sollte je nach Tätigkeit nicht der Mindestlohn, sondern branchenübliche Gehälter als Orientierung dienen. Bei belastenden Aufgaben (wie Konfrontation mit psychisch belastenden Inhalten) sollten Zulagen wie Gefahren- oder Erschwerniszulagen in Erwägung gezogen werden.
- **Transparente und standardisierte Aufgabenblätter:**
  Klare, konsistente Briefings reduzieren Fehlklassifikationen und sparen sowohl Data Worker als auch Auftraggeber:innen Zeit. Standardisierte Aufgabenformate mit Beispielantworten und visuellen Referenzen wirken sich nachweislich positiv auf die Qualität aus.
- **Feedback- und Schulungssysteme:**
  Statt rein binärer Bewertungen können Auftraggeber:innen auf iterative Tests oder Mini-Schulungen setzen, die gleichzeitig als Qualifizierung für neue Aufgaben dienen – etwa durch interaktive Lernmodule oder Zertifizierungen. Diese Maßnahmen fördern nicht nur Fairness, sondern auch langfristige Qualitätssicherung.
- **Health & Safety:**
  Data Worker, die in der Betrugsdetektion arbeiten, können mit problematischen oder belastenden Inhalten (z. B. pornografischen oder gewaltvollen Websites) konfrontiert sein. Auftraggeber:innen sollten deshalb sicherstellen, dass solche Risiken minimiert werden – etwa durch Vorfilterung, Warnhinweise, Freiwilligkeit bei riskanten Inhalten und gegebenenfalls Risikozulagen oder Supervision.
- **Transparente Kommunikation und Einspruchsmöglichkeiten:**
  Auch wenn Plattformen den direkten Kontakt erschweren, können

Auftraggeber:innen Mechanismen schaffen, über die Data Worker Feedback geben oder Fehler melden können – z. B. über eine Kontaktmail oder einen Kommentarbereich im Auftragstext.

Verantwortung der Plattformen

Die Plattformen fungieren als zentrale Kontrollinstanzen: Sie bestimmen Zugang, Bewertung, Vergütung und Ausschluss. Damit übernehmen sie faktisch Arbeitgeberfunktionen (Lutz 2017; Risak 2016), ohne an arbeitsrechtliche Schutzpflichten gebunden zu sein. Ihre Verantwortung umfasst daher mindestens drei Kernbereiche:

- **Transparente Algorithmen und Auftragslogik:** Offenlegung der Kriterien, nach denen Aufgaben verteilt und Leistungen bewertet werden.
- **Faire Vertragsgestaltung:** Verzicht auf einseitige AGB-Änderungen, willkürliche Sperrungen oder verdeckte Provisionsmodelle.
- **Soziale Sicherheit:** Einführung von Mindestentgelten, Versicherungsschutz und Möglichkeiten kollektiver Interessenvertretung.

Auch wenn einzelne Plattformen hier erste Schritte setzen, bleibt die Implementierung oft freiwillig. Eine verbindliche Regulierung – etwa über EU-weite Transparenzpflichten oder Zertifizierungsverfahren – wäre notwendig, um strukturelle Machtasymmetrien zu verringern.

Ethische Nutzung und Alternativen

Die zentrale Frage bleibt, ob Plattformarbeit unter gegenwärtigen Bedingungen überhaupt ethisch vertretbar ist. In vielen Fällen stehen Auftraggeber:innen vor einem Dilemma: Sie benötigen flexible, skalierbare Human-in-the-Loop-Strukturen, wollen aber nicht zu prekären Arbeitsverhältnissen beitragen.

Eine verantwortbare Nutzung ist nur dann möglich, wenn folgende Mindestbedingungen erfüllt werden:

- Transparenz über Zweck und Verwertung der Arbeit.
- Faire Bezahlung gemäß nationalen Standards.
- Minimierung psychischer Belastung und Freiwilligkeit bei sensiblen Inhalten.
- Rückmeldungskanäle und nachvollziehbare Bewertungsprozesse.

Langfristig sollte jedoch über Alternativen nachgedacht werden, die strukturell fairere Bedingungen ermöglichen. Dazu gehören etwa:

- Kooperation mit Fairwork-zertifizierten Plattformen, die verbindliche Standards erfüllen.
- Aufbau interner Microtasking-Pools innerhalb von Institutionen, z. B. mit Studierenden oder geschulten Honorarkräften, unter klar definierten Verträgen.
- Open-Crowd-Frameworks im Non-Profit-Bereich, bei denen Aufgaben freiwillig oder honoriert, aber transparent und mit Lern- oder Beteiligungsmehrwert gestaltet werden.

Für den Fake-Shop Detector könnte mittelfristig ein hybrides Modell sinnvoll sein: die Kombination von qualifizierten Freiwilligen (z. B. aus Konsument:innenschutz-Communities) und geschulten Teilzeitkräften, unterstützt durch KI-basierte Tools. So ließe sich das Ziel der Skalierung mit einem stärker sozial verankerten Verständnis von Verantwortung verbinden.

3. Conclusio

Onlinebetrug stellt weiterhin eine wachsende Bedrohung für Konsument:innen dar, die durch täuschend echt gestaltete Websites und den Einsatz von KI-Technologien verstärkt wird. Die vorliegende Untersuchung zeigt, dass teil-automatisierte Systeme wie der Fake-Shop Detector die Betrugsprävention erheblich skalieren können, indem sie Daten aus unterschiedlichen Quellen analysieren und menschliche Expertise gezielt ergänzen.

Die Evaluation entlang der ethischen Leitlinien der Europäischen Kommission macht deutlich: Technische Leistungsfähigkeit allein reicht nicht aus, um ein vertrauenswürdiges System zu gewährleisten. Nur die Kombination aus robusten Algorithmen, transparenter Kommunikation, kontinuierlicher menschlicher Aufsicht und datenschutzkonformen Prozessen kann sowohl die Effizienz steigern als auch ethische Anforderungen erfüllen.

Im Kontext von Human-in-the-Loop Ansätzen zeigt sich zudem, dass aktuell die Arbeitsbedingungen von Data Workern u.a. hinsichtlich fehlender Absicherungen bei gleichzeitiger Abhängigkeit und schlechter Bezahlung nur unzureichend sind. Plattformen müssen hier stärker in die Verantwortung gezogen werden, um Mindeststandards einzuhalten, aber auch Auftraggeber:innen können durch klare Aufgabenstellungen, angemessene Vergütungen, Schulungen sowie transparente Feedbackmechanismes aktiv dazu beitragen, dass HITL-Prozesse fairer gestaltet werden.

Die Analyse zeigt, dass ethische Verantwortung in der automatisierten Betrugsdetektion nicht bei algorithmischer Fairness endet, sondern auch jene Menschen einschließt, die die Systeme trainieren und überprüfen. Faire Arbeitsbedingungen sind dabei eine Voraussetzung für die Verlässlichkeit der Ergebnisse: Nur wer die Menschen hinter den Daten schützt und ihre Arbeitsbedingungen fair gestaltet, kann die Zuverlässigkeit, Nachvollziehbarkeit und gesellschaftliche Legitimation automatisierter Betrugsdetektionssysteme langfristig sicherstellen.

4. Quellen

Beltzung, Louise, Andrew Lindley, Olivia Dinica, Nadin Hermann, und Raphaela Lindner. 2020. „Real-Time Detection of Fake-Shops through Machine Learning“. *2020 IEEE International Conference on Big Data (Big Data)*, Dezember, 2254–63. https://doi.org/10.1109/BigData50022.2020.9378204.

Bundesministerium für Inneres. 2025. „Cybercrimereport 2024“. https://www.bmi.gv.at/magazin/2025_11_12/01_Cybercrime_Report.aspx.

Europäische Kommission. 2016. „Understanding Consumer Vulnerability in the EU's Key Markets - European Commission“. https://commission.europa.eu/publications/understanding-consumer-vulnerability-eus-key-markets_en.

Fairwork. 2025. *Fairwork Cloudwork Ratings 2025. Advancing Standards in Digital Labour and AI Supply Chain Governance*. Fairwork Cloudwork Ratings. Fairwork. https://fair.work/wp-content/uploads/sites/17/2025/05/Fairwork-Cloudwork-Report-2025-FINAL.pdf.

Fenaux, Lucas, Christopher Srinivasa, und Florian Kerschbaum. 2025. „On the Trade-Off Between Transparency and Security in Adversarial Machine Learning“. arXiv:2511.11842. Preprint, arXiv, November 14. https://doi.org/10.48550/arXiv.2511.11842.

Gebrekidan, Fasica Berhane. 2024. *Content Moderation: The harrowing, traumatizing job that left many African data workers with mental health issues and drug dependency.* Data Workers' Inquiry. https://data-workers.org/fasica/.

Hao, Karen. 2025. *Empire of AI - Dreams and Nightmares in Sam Altman's OpenAI*. Penguin Press.

Hao, Karen, und Andrea Paola Hernández. 2022. „How the AI Industry Profits from Catastrophe“. *MIT Technology Review*, AI Colonialism, April 20. https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels/.

Huws, Ursula. 2001. „The Making of a Cybertariat: Virtual Work in a Real World“. *Socialist Register*, Working Classes, Global Realities, Bd. 37.

Krichmayr, Karin. 2018. „Arbeitsforscherin: ‚Die Digitalisierung kreiert ein neues Prekariat' - Welt - derStandard.de › Wissen und Gesellschaft“. *Der Standard*, Mai 1. https://www.derstandard.de/story/2000078786980/arbeitsforscherin-die-digitalisierung-kreiert-ein-neues-prekariat.

Lutz, Doris. 2017. „Virtuelles Crowdwork: Clickworker“. In *Arbeit in der Gig-Economy: Rechtsfragen neuer Arbeitsformen in der Crowd und Cloud*, herausgegeben von Doris Lutz und Martin Gruber-Risak. Varia. ÖGB Verlag.

Nesvijevskaia, Anna, Sophie Ouillade, Pauline Guilmin, und Jean-Daniel Zucker. 2021. „The Accuracy versus Interpretability Trade-off in Fraud Detection Model“. *Data & Policy* 3 (Januar): e12. https://doi.org/10.1017/dap.2021.3.

Okinyi, Mophat. 2024. „Impact of Remotasks' Closure on Kenyan Workers". *Data Workers' Inquiry*. https://data-workers.org/mophat/.

Richtlinie (EU) 2024/2831 des Europäischen Parlaments und des Rates vom 23. Oktober 2024 zur Verbesserung der Arbeitsbedingungen in der Plattformarbeit, 2024/2831 2024/2831 (2024).

Risak, Martin. 2017. „(Arbeits-)Rechtliche Aspekte der Gig-Economy". In *Arbeit in der Gig-Economy: Rechtsfragen neuer Arbeitsformen in der Crowd und Cloud*, herausgegeben von Doris Lutz und Martin Gruber-Risak. Varia. ÖGB Verlag.

5. Anhang

Fake-Shop Detector Model Card

Model Cards sind eine Möglichkeit, um eine einfache und strukturierte Übersicht über die Funktionsweisen sowie Limitierungen von KI-Modellen zu erstellen. Ziel der entwickelten FSD Model Card ist es neben der Funktionsweise des FSD, auch die Limitierungen sowie empfohlene Anwendungsarten für Stakeholder verständlich zu machen. Die Model Card kann auch hier als pdf-Datei heruntergeladen werden:

https://www.netidee.at/sites/default/files/2026-02/FSD_ModelCards.pdf

Was macht der Fake-Shop Detector?

Der Fake-Shop Detector ist ein KI-gestütztes Warnsystem, das Nutzer:innen beim Online-Shopping hilft und vor betrügerischen Webshops warnt.

Für wen wurde der Fake-Shop Detector entwickelt?

Hauptzielgruppe sind Personen aus dem deutschsprachigen Raum, die online einkaufen, sowie Konsumentenschutz-Organisationen, die das Tool zur Überprüfung von Online-Shops nutzen.

Konkret wurde das System entwickelt für:

- Menschen, die beim Online-Shopping sicher sein wollen.
- Alle Altersgruppen und technische Erfahrungslevel.
- Besonders vulnerable Personen mit erhöhtem Betrugsrisiko.

Empfohlene Verwendung

- Als Warnhinweis vor verdächtigen Online-Shops direkt beim Surfen (Browser-Plugin).
- Zur Überprüfung unbekannter Webshops vor dem Kauf (Browser-Plugin und Shop-Check).

Wie funktioniert der Fake-Shop Detector?

Das System analysiert den Quellcode von Webshops und extrahiert über 22.000 strukturelle Merkmale aus HTML, CSS und JavaScript. Fake-Shop-Betreiber hinterlassen dabei unbewusst wiederkehrende Muster im Code – sogenannte digitale Fingerabdrücke –, die von der KI erkannt werden.

Welche Merkmake werden analysiert?

Feature-Extraktion

- Tokenisierter HTML/CSS/JavaScript-Text
- DOM-Baumstruktur und Tag-Attribut-Muster
- Verwendete Bibliotheken und Frameworks
- Code-Kommentare
- Strukturelle Patterns und fehlende Elemente

Vektorisierung

Die Vektorisierung erfolgt mittels TF-IDF (Term Frequency – Inverse Document Frequency), wobei häufige, unspezifische Elemente gering und seltene, charakteristische Merkmale hoch gewichtet werden.

Die drei KI-Modelle

Modell	Typ	Accuracy	Precision	Recall	F1
XGBoost Höchste Einzelgenauigkeit, erkennt komplexe Muster	Gradient Boosted Trees	97%	97%	97%	97%
Random Forest Reduziert False Positives, stabil bei legitimen Shops	Ensemble Decision Trees	95%	97%	94%	96%
Neural Network Findet abstrakte, nicht linieare Zusammenhänge	Multi-Layer Perceptron	94%	99%	90%	94%

Wie genau ist das System?

Eine gleichgewichtete Kombination der drei KI-Modell XGBoost, Fandom Forest und Neural Network führt zu **90,38% Genauigkeit** im realen Einsatz.

Risikostufen und ihre Zuverlässigkeit

Risikostufe	Fehlerrate	Bedeutung für Sie
Hoch (90-100%)	0-1%	Sehr hohe Wahrscheinlichkeit eines Fake-Shops. Stark vom Kauf abgeraten.
Mittel (70-89%)	5-10%	Verdächtige Merkmale erkannt. Zusätzliche Prüfung empfohlen.
Niedrig (50-69%)	11%	Einige Auffälligkeiten. Weitere Recherche ratsam.

Mehr Informationen:

Beltzung, L., Lindley, A., Dinica, O., Hermann, N., & Lindner, R. (2020). „Real-Time Detection of Fake-Shops through Machine Learning." In Big Data and Cyber-Physical Systems in Industrial and Cyber-Physical Systems Applications.

Welche Limitationen hat der Fake-Shop Detector?

Was der Fake-Shop Detector nicht kann:

- 100% Sicherheit garantieren
- Neue, völlig unbekannte Betrugsmuster sofort erkennen
- Shops außerhalb des deutschsprachigen Raums bewerten

Mögliche Fehlerquellen:

- Sehr neue Betrugsmaschen, die das System noch nicht kennt
- Technisch sehr aufwändige Fake-Shops, die legitime Shops perfekt imitieren
- Legitime Shops, die technisch ähnlich aufgebaut sind wie bekannte Fake-Shops

Datenschutz & Privatsphäre

Was speichern wir:

- URL der überprüften Website
- Zeitpunkt der Überprüfung
- Technische Merkmale der Website

Was speichern wir nicht:

- Identität der Nutzer:innen
- Kaufabsichten der Nutzer:innen
- Zahlungsdaten der Nutzer:innen
- Surfverhalten der Nutzer:innen

Wie funktioniert der Datenschutz:

- Risikobewertungen erfolgen lokal auf Ihrem Gerät (Cache-First-Prinzip)
- Nur unbekannte Websites werden an unseren Server gesendet
- Log-Files werden nicht mit Dritten geteilt

Qualitätssicherung & Datenquellen

Menschliche Kontrolle

- Manuelle Qualitätssicherung wird kontinuierlich von Expert:innen der Watchlist Internet durchgeführt
- Integration von kuratierten Blacklists (unbekannte Fake-Shops) und Whitelists

Datenquellen

- Warnliste der Watchlist Internet
- Zertifizierte Online-Shops des Österreichischen E-Commerce Gütezeichen, des Schweizer Gütezeichens, der Trustmark Austria, des EHI Siegels und des Trusted Shops Gütesiegels
- Liste österreichischer Buchhandlungen
- Händlerliste des Elektrofachhändlers Expert
- Liste der Händler auf Online-Shop-Austria.at
- Verzeichnis österreichischer und deutscher Versandapotheken
- Liste der Händler auf Kaufsregional.at
- Falter Onlineshop-Fibel
- Liste der Händler des Onlineshop-Verzeichnis

Verantwortlichkeiten, Kontakt & Feedback

Entwicklung & Betrieb

AIT: Technische Entwicklung, ML-Modelle
ÖIAT: Froschung & Projektkoordination
X-Net Services: Technische Infrastruktur
Gefördert durch: netidee

Kontakt

Website: fakeshop.at
E-Mail: kontakt@fakeshop.at

Fehler melden:

- Direkt im Plugin auf „Fehler melden“ klicken
- Mail an kontakt@fakeshop.at