



Abschließende Mitteilung

an die
Bundesagentur für Arbeit

über die Prüfung

Einsatz künstlicher Intelligenz und analytischer
Verfahren in der Bundesagentur für Arbeit

Teil 1 – Verfahren zur Kassensicherheit in ALLEGRO

Diese Prüfungsmitteilung enthält das vom Bundesrechnungshof abschließend im Sinne des § 96 Abs. 4 Satz 1 BHO festgestellte Prüfungsergebnis. Sie ist auf der Internetseite des Bundesrechnungshofes veröffentlicht (www.bundesrechnungshof.de).

Gz.: VI 1 - 2020 - 0543 / VII 3 - 2020 - 0079

Bonn, den 10. November 2021

Inhaltsverzeichnis

Abkürzungsverzeichnis	3
0 Zusammenfassung	4
1 Vorbemerkung	7
2 Informationen zu ALLEGRO	9
3 Kontrollsystem für ALLEGRO und Anwendung durch die gE	10
3.1 Sachverhalt	10
3.1.1 Nutzungsgrad des „Zwei-Augen-Prinzips“ in ALLEGRO	10
3.1.2 Anwendung des Kontrollsystems bei der Überprüfung auffälliger Anordnungen durch die gE	12
3.2 Würdigung	16
3.3 Empfehlung	17
3.4 Stellungnahme	17
3.5 Abschließende Würdigung	19
4 Auswahl und Eignung der analytischen Verfahren	20
4.1 Sachverhalt	20
4.2 Würdigung	23
4.3 Empfehlung	25
4.4 Stellungnahme	25
4.5 Abschließende Würdigung	26
5 Wirtschaftlichkeit des Einsatzes analytischer Verfahren	26
5.1 Sachverhalt	26
5.2 Würdigung	28
5.3 Empfehlung	28
5.4 Stellungnahme	28
5.5 Abschließende Würdigung	29

Abkürzungsverzeichnis

BestMaVB-HKR	Bestimmungen über die Mindestanforderungen für den Einsatz automatisierter Verfahren im Haushalts-, Kassen- und Rechnungswesen des Bundes
BMF	Bundesministerium der Finanzen
Bundesagentur	Bundesagentur für Arbeit
EFM	Enterprise Fraud Management
ERP	Enterprise Resource Planning
gE	gemeinsame Einrichtungen
nzStpp	nachträgliche zufallsorientierte Stichproben
PAL	Predictive Analysis Library
VZÄ	Vollzeitäquivalente

0 Zusammenfassung

Der Bundesrechnungshof prüfte bei der Bundesagentur für Arbeit (Bundesagentur) den Einsatz künstlicher Intelligenz und analytischer Verfahren zur Gewährleistung der Kassensicherheit.

Die vorliegende Teilprüfungsmitteilung umfasst die Feststellungen zum Verfahren „ALLEGRO-EFM“. Die Feststellungen zum Projekt „ERP-2AP“ sowie zur strategischen Ausrichtung und Koordination der Entwicklungen zur Kassensicherheit in der Bundesagentur beschreibt der Bundesrechnungshof in einer weiteren Teilprüfungsmitteilung.

Unter Berücksichtigung der Stellungnahme der Bundesagentur stellt er die wesentlichen Prüfungsergebnisse abschließend wie folgt fest:

- 0.1 Mit dem Einsatz des Vier-Augen-Prinzips sollen bei Zahlungen Fehler so weit wie möglich vermieden werden. Ab dem Jahr 2016 konnten in den gemeinsamen Einrichtungen Zahlungen bis 500 Euro auch durch nur eine Person angeordnet werden („Zwei-Augen-Prinzip“).

Der Bundesagentur ist es nicht gelungen, die gemeinsamen Einrichtungen bei ALLEGRO durch technische Hilfsmittel und nachträgliche manuelle Prüfungen spürbar zu entlasten. Das Verfahren zur nachträglichen Kontrolle war aufwändig und als Alternative zur Gewährleistung der Kassensicherheit nicht geeignet. Der mit dem Verfahren verbundene Aufwand trug vermutlich dazu bei, dass die gemeinsamen Einrichtungen das „Zwei-Augen-Prinzip“ nur selten nutzten. Zudem meldeten sie der Bundesagentur zu den nachträglich geprüften Fällen oft unvollständige Daten zurück. Obwohl die Bundesagentur die Schwächen des Verfahrens seit dem Jahr 2017 kannte, entwickelte sie dieses nicht weiter. Letztendlich stellte sie es zum 1. Juli 2020 ein.

Zum Scheitern des Verfahrens hat wesentlich beigetragen, dass die Bundesagentur es vor dem flächendeckenden Einsatz nicht systematisch getestet und die gemeinsamen Einrichtungen nicht ausreichend eingebunden hat. Sie hat damit riskiert, dass deren Beschäftigte zusätzlich belastet wurden. Sie muss künftig darauf achten, dass neue Verfahren anhand von aussagekräftigen Rückmeldungen der gemeinsamen Einrichtungen zielgerichtet weiterentwickelt werden können.

Die Bundesagentur hat eingeräumt, dass das Verfahren optimierungsfähig gewesen sei und zu erhöhten Anforderungen bei den Beschäftigten geführt habe. Einen eindeutigen Zusammenhang zwischen dem Aufwand des Verfahrens und dessen Nutzungsgrad habe sie aber nicht feststellen können. Die Bundesagentur hat der Feststellung des Bundesrechnungshofes widersprochen, dass das Verfahren gescheitert ist. Sie hat jedoch eingeräumt, dass die Rückmeldungen der gemeinsamen Einrichtungen nicht aussagekräftig genug gewesen seien.

Der Bundesrechnungshof bleibt bei seinen Würdigungen und Empfehlungen. Die Bundesagentur benötigt aussagekräftige Rückmeldungen aus den gemeinsamen Einrichtungen, um solche neuartigen Verfahren zielgerichtet weiterentwickeln zu können. Dazu muss sie vor Aufnahme des Wirkbetriebs Tests in ausgewählten gemeinsamen Einrichtungen durchführen und diese enger in die Entwicklung einbinden. (Tz. 3)

- 0.2 Die Bundesagentur nutzte Methoden des maschinellen Lernens, um in den Buchungen des jeweils vergangenen Monats auffällige Anordnungen zu suchen. Diese Anordnungen sollten die gemeinsamen Einrichtungen nachträglich manuell prüfen. Die Wahl ihrer Methoden sowie einiger dazugehöriger Parameter hat sie weder angemessen dokumentiert noch nachvollziehbar begründen können.

Die Bundesagentur beachtete nicht ausreichend, dass ihre Ergebnisse wesentlich von der Wahl der Parameter abhängen. Sie untersuchte diese Abhängigkeit nicht und nutzte bei mehreren Parametern ohne weitere Prüfung die vorgegebenen Standardwerte. Um die auch von ihr erkannten Qualitätsmängel zu beheben, wollte die Bundesagentur ihre Analysen mit Methoden des überwachten maschinellen Lernens ergänzen. Hierzu hätte es allerdings hochwertiger Trainingsdaten bedurft, die bereits frühzeitig hätten erhoben werden müssen. Dies hatte die Bundesagentur versäumt.

Die Bundesagentur setzte maximal drei Vollzeitbeschäftigte für die maschinellen Überprüfungen beim „Zwei-Augen-Prinzip“ in ALLEGRO ein. Sie hätte frühzeitig erkennen müssen, dass sie mit dieser personellen Ausstattung das System nur geringfügig anpassen konnte. Der unzureichende Ressourceneinsatz trug wesentlich zu den methodischen

Schwächen und damit letztendlich zum Scheitern des „Zwei-Augen-Prinzips“ bei.

Der Erfolg künftiger Vorhaben der Bundesagentur wird wesentlich davon abhängen, dass sie Parameter und Ergebnisse neuer Verfahren kritisch überprüft und die Qualität dieser Verfahren auf einer verlässlichen Datengrundlage stetig weiterentwickelt. Hierfür ist eine ausreichende Personalausstattung unerlässlich.

Die Bundesagentur hat mitgeteilt, dass sie die Wahl ihrer Methoden sowie die zugehörigen Parameter ausreichend begründet und dokumentiert habe. Sie hat dem Bundesrechnungshof darin zugestimmt, dass eine angemessene Personalausstattung für vergleichbare Projekte unerlässlich ist.

Die Bundesagentur konnte gegenüber dem Bundesrechnungshof keine ausreichende Dokumentation nachweisen. Er bleibt daher bei seinen Würdigungen und Empfehlungen. (Tz. 4)

- 0.3 Bei Gesamtkosten von rund 7 Mio. Euro für die Entwicklung maschineller Kontrollen in ALLEGRO hätte die Bundesagentur eine Wirtschaftlichkeitsuntersuchung durchführen müssen. Zudem hatte sie versäumt, zu Beginn des neuen Verfahrens eindeutige Ziele festzulegen. Damit konnte sie während des Betriebs und abschließend keine belastbaren Erfolgskontrollen vornehmen.

Die Bundesagentur hat eingeräumt, dass sie für das Verfahren „ALLEGRO-EFM“ keine gesonderte Wirtschaftlichkeitsuntersuchung erstellt habe. Die wesentlichen Teile des Verfahrens sollten in anderen Projekten entwickelt werden. Für diese Projekte lägen Wirtschaftlichkeitsuntersuchungen vor. Das Verfahren „ALLEGRO-EFM“ sei in diesen Wirtschaftlichkeitsuntersuchungen ausreichend berücksichtigt.

Für ein Verfahren mit geschätzten Kosten in Höhe von 7 Mio. Euro hätte die Bundesagentur eine **gesonderte** Wirtschaftlichkeitsuntersuchung durchführen müssen. Die Wirtschaftlichkeit lediglich zusammen mit anderen Projekten zu betrachten, hält der Bundesrechnungshof in diesem Fall für nicht ausreichend. (Tz. 5)

1 Vorbemerkung

Die Bundesagentur für Arbeit (Bundesagentur) nutzt für Verwaltungs- und Bewirtschaftungsprozesse des Finanz- und Personalwesens das System „Enterprise Resource Planning“ (ERP) des Unternehmens SAP SE. Ein wesentlicher Zweck des IT-Verfahrens ERP ist die Zahlbarmachung von Leistungen und Rechnungen bei der Bundesagentur und den gemeinsamen Einrichtungen (gE)¹.

Daneben setzt die Bundesagentur in den gE das IT-Verfahren ALLEGRO ein. In ALLEGRO werden die wesentlichen Leistungsansprüche in der Grundsicherung für Arbeitsuchende festgestellt und angeordnet.² Eine elektronische Schnittstelle übermittelt die Daten unveränderbar von ALLEGRO an ERP.

Beim Einsatz von automatisierten Verfahren sind die Bestimmungen über die Mindestanforderungen für den Einsatz automatisierter Verfahren im Haushalts-, Kassen- und Rechnungswesen des Bundes (BestMaVB-HKR) einzuhalten. Diese dienen dazu, die Kassensicherheit in automatisierten Verfahren zu gewährleisten. Ein Kernelement der Vorgaben ist das Vier-Augen-Prinzip. Hiernach darf eine Person, die an der Feststellung einer Zahlung beteiligt war, nicht an der Anordnung beteiligt sein. Bei geringen haushaltswirtschaftlichen Risiken³ kann die Prüfung durch das zweite Augenpaar auf Stichproben beschränkt werden. Diese Voraussetzung hat die Bundesagentur für die Zahlungen in ALLEGRO und ERP bestätigt.

Wir stellten bei früheren Prüfungen der Bundesagentur fest, dass diese vom Vier-Augen-Prinzip abwich, ohne das in den BestMaVB-HKR beschriebene Stichprobenverfahren durchzuführen. Insbesondere konnte die Bundesagentur die in der Stichprobe ausgewählten Zahlungen nicht, wie vorgeschrieben, bis zum Abschluss der Stichprobenprüfung anhalten⁴.

Die Bundesagentur beantragte daher eine Ausnahmegenehmigung, um von den Vorgaben der BestMaVB-HKR für Stichprobenverfahren abzuweichen und die Kassensicherheit für ERP und ALLEGRO auf „andere Art und Weise“ zu

¹ Jobcenter nach § 44 b Absatz 1 SGB II, in dem die Bundesagentur und die zuständigen Kreise und kreisfreien Städte die Aufgaben gemeinsam wahrnehmen.

² Hierzu gehören u. a. Arbeitslosengeld II und Unterkunftskosten.

³ Gefährdungsanalyse nach VV Nummer 6.3 ZBR BHO.

⁴ Nummer 5.4.2.2 BestMaVB-HKR, Stand: Januar 2017.

gewährleisten.⁵ Sie sah für ALLEGRO automatisierte Prüfungen und für ERP von der Vorschrift abweichende nachträgliche zufallsorientierte Stichproben (nzStpp) vor. Die Bundesagentur kündigte in ihrem Ausnahmeantrag an, dass sie auch für ERP automatisierte Prüfungen als Ersatz für die nzStpp entwickeln wolle. Diese fasste sie unter dem Begriff „Data-Mining“ zusammen.

Dem Ausnahmeantrag zum Einsatz von ERP und ALLEGRO erteilten wir unser Einvernehmen zunächst befristet und später unbefristet.

Der Bereich „Enterprise Fraud-Management“ (EFM) der Bundesagentur hatte im Jahr 2016 begonnen, die Kassensichersicherheit für ALLEGRO automatisiert zu prüfen (ALLEGRO-EFM).

Am 14. November 2019 führten wir ein Informationsgespräch mit der Bundesagentur. Sie stellte ihre Überlegungen und bisherigen Entwicklungen zum Einsatz künstlicher Intelligenz und analytischer Verfahren vor und bestätigte, dass sie erste Module ab dem Jahr 2020 in ERP einsetzen wolle. Sie kündigte einen neuen Antrag zur Ausnahmegenehmigung für ERP an.

Die Vorhaben der Bundesagentur sind für die Kassensicherheit automatisierter IT-Verfahren bedeutend. Wir untersuchten daher das Vorgehen und die Methoden der Bundesagentur sowohl fachlich als auch informationstechnisch.

Mit der Prüfung wollten wir insbesondere feststellen, inwieweit mit den analytischen Verfahren die Kassensicherheit „auf andere Art und Weise“ als bei der Einhaltung des Vier-Augen-Prinzips sichergestellt werden kann. Nachfolgend stellen wir die Ergebnisse unserer Prüfung unter Berücksichtigung der Stellungnahme der Bundesagentur vom 2. Juli 2021 abschließend fest.

Die vorliegende Teilprüfungsmitteilung umfasst die Feststellungen zum Verfahren „ALLEGRO-EFM“. Die Feststellungen zum Projekt „ERP-2AP“ sowie zur strategischen Ausrichtung und Koordination der Entwicklungen zur Kassensicherheit in der Bundesagentur beschreiben wir in einer weiteren Teilprüfungsmitteilung.

⁵ Nummer 1 Absatz 2 BestMaVB-HKR (Stand: Januar 2017) i. V. m. VV Nummer 6.6.2 ZBR BHO.

2 Informationen zu ALLEGRO

Die Bundesagentur nutzt ALLEGRO zur Feststellung und Anordnung von Leistungen nach dem SGB II.

Soweit die Feststellung und Anordnung im „Zwei-Augen-Prinzip“ erfolgt, konnte die Bundesagentur die Vorschriften für das ersatzweise anzuwendende Stichprobenverfahren der BestMaVB-HKR nicht einhalten und bedurfte daher einer Ausnahmegenehmigung.⁶ Ab dem 3. August 2016 konnte eine Person alleine Einmal- und Nachzahlungen in Höhe von bis zu 500 Euro feststellen und anordnen. Die Ausnahmegenehmigung enthielt folgende Auflagen:

- die Berechtigungsprofile anzupassen,
- regelmäßig automatisierte Stichprobenprüfungen durchführen zu lassen,
- sicherzustellen, dass eine systemseitige Rückkehr zum Vier-Augen-Prinzip für alle Anordnungen möglich ist,
- nachzuweisen, wie viele Stellen sie bis wann haushaltswirksam abbaut und
- Einmalzahlungen bis 500 Euro regelmäßig durch die Interne Revision prüfen zu lassen.

Die Bundesagentur gab an, dass das „Zwei-Augen-Prinzip“ den Aufwand für die gE reduziere. Für die vollständige Umsetzung im Vier-Augen-Prinzip in ALLEGRO hatte die Bundesagentur zum 1. Januar 2016 zusätzlich 300 Stellen und 370 Ermächtigungen⁷ erhalten. Diese stellte sie zum 31. Dezember 2016 in Abgang.

Zum 30. Juni 2020 schaffte die Bundesagentur das „Zwei-Augen-Prinzip“ in ALLEGRO wieder ab. Sie begründete dies u. a. mit hohen Fehlerquoten bei der Anweisung von Zahlungen.

⁶ Nummer 1 Absatz 2 BestMaVB-HKR i. V. m. VV Nummer 6.6.2 ZBR BHO.

⁷ Die Bundesagentur weist in ihrem Personalhaushalt neben Planstellen und Stellen sogenannte „Ermächtigungen“ aus. Diese verwendet sie, um Beschäftigte befristet einzustellen, sofern entsprechende Haushaltsmittel vorhanden sind.

3 Kontrollsystem für ALLEGRO und Anwendung durch die gE

Im Folgenden beschreiben wir das Kontrollsystem, das die Bundesagentur bei ALLEGRO eingeführt hatte, um die Kassensicherheit „auf andere Art und Weise“ als in Nummer 5.4.2 BestMaVB-HKR⁸ zu gewährleisten.

Die Bundesagentur prüfte Einmal- und Nachzahlungen bis 500 Euro, die durch eine Person festgestellt und angeordnet wurden.

Die Bundesagentur suchte mit Prüfalgorithmen nach Auffälligkeiten in den Zahlungen. Sie ermittelte zu jeder Zahlung insgesamt 50 verschiedene Kennzahlen, u. a.

- Betragshöhe,
- Zahlungsempfänger,
- Auszahlungshistorie,
- Kontowechsel und
- Doppelzahlungen.

3.1 Sachverhalt

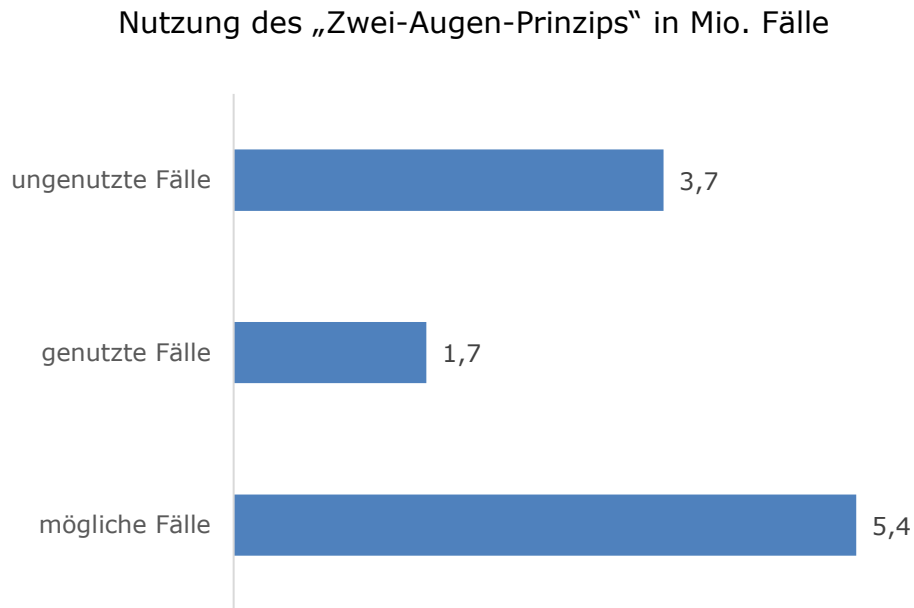
3.1.1 Nutzungsgrad des „Zwei-Augen-Prinzips“ in ALLEGRO

In ALLEGRO wurden in den Jahren 2017 bis 2019 insgesamt 56,5 Millionen Anordnungen ausgeführt. Davon wurden 18,3 Millionen Anordnungen für Einmal- und Nachzahlungen getätigt. 5,4 Millionen der Einmal- und Nachzahlungen beliefen sich auf bis zu 500 Euro und waren damit grundsätzlich geeignet, im „Zwei-Augen-Prinzip“ angeordnet und anschließend analytisch überprüft zu werden. Tatsächlich ordneten die gE nur rund 1,7 Millionen davon im „Zwei-Augen-Prinzip“ an.

Die gE entschieden damit in rund 70 % der Fälle, das neu entwickelte Verfahren des „Zwei-Augen-Prinzips“ nicht zu nutzen (Abbildung 1).

⁸ Stand: Januar 2017.

Abbildung 1

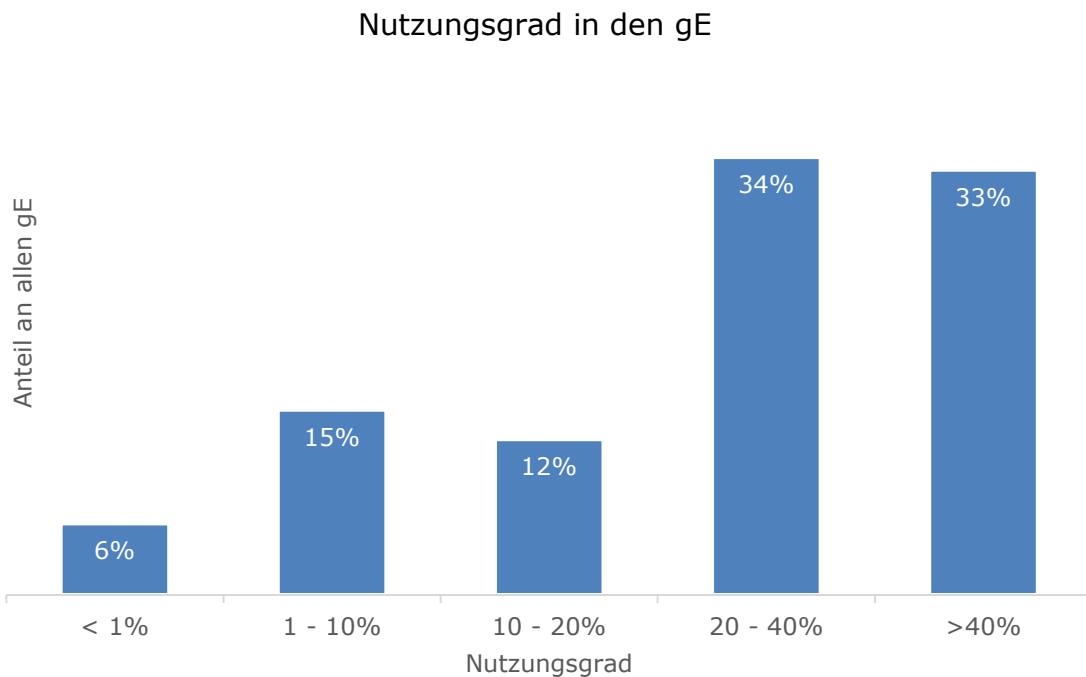


Quelle: Eigene Darstellung des Bundesrechnungshofes, Basis: Tabellen der Bundesagentur zur Nutzung des „Zwei-Augen-Prinzips“ in ALLEGRO.

Die Bundesagentur nannte diese Fälle *„ungenutztes Potential 2AP“*. Sie stellte den geringen Nutzungsgrad erstmals im Jahr 2017 fest.

Die einzelnen gE setzten das „Zwei-Augen-Prinzip“ unterschiedlich oft ein. 6 % der gE verzichteten fast vollständig auf den Einsatz. D. h. sie nutzten für weniger als 1 % der möglichen Anordnungen das „Zwei-Augen-Prinzip“. Zwei Drittel der gE setzten das Verfahren in weniger als 40 % der möglichen Fälle ein (Abbildung 2). Der niedrigste Nutzungsgrad lag bei 0,1 % und der höchste bei 75,2 %.

Abbildung 2



Quelle: Bundesrechnungshof auf Grundlage der Erhebungen

Beispiel:

Die gE A hätte im Jahr 2019 insgesamt 20 581 Zahlungen im „Zwei-Augen-Prinzip“ anordnen können. Tatsächlich ordnete sie auf diesem Wege 105 Zahlungen (0,5 %) an.

Die Bundesagentur appellierte an die gE, Anordnungen im „Zwei Augen Prinzip“ vorzunehmen und damit den Nutzungsgrad zu erhöhen. Der Bereich EFM besprach den geringen Nutzungsgrad in monatlichen Statusrunden mit dem fachlichen Bedarfsträger. Da die Bundesagentur keine Weisungsbefugnis gegenüber den gE hatte, konnte sie die Nutzung des „Zwei-Augen-Prinzips“ nicht verpflichtend durchsetzen. In einem Workshop mit Beschäftigten zweier gE versuchte sie, den geringen Nutzungsgrad zu evaluieren.

Weitere Anstrengungen den Nutzungsgrad zu erhöhen unternahm sie nicht.

3.1.2 Anwendung des Kontrollsystems bei der Überprüfung auffälliger Anordnungen durch die gE

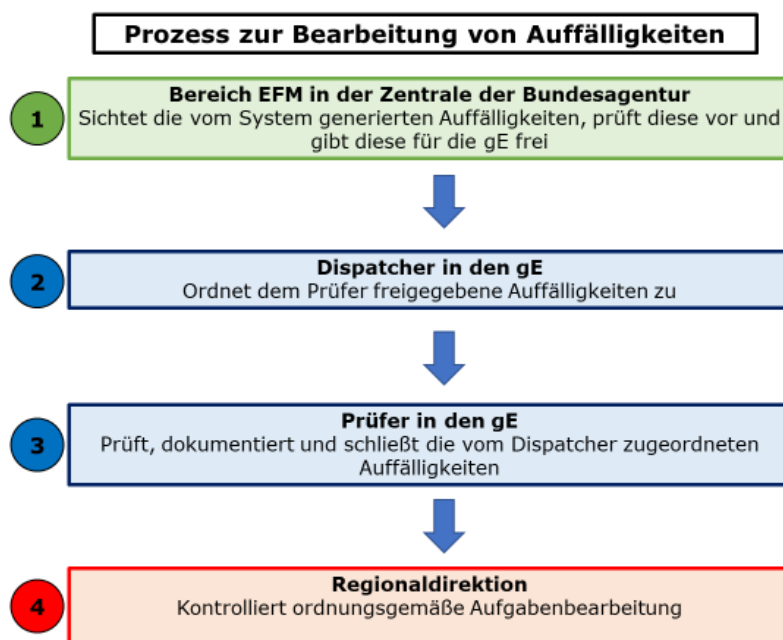
Die im „Zwei-Augen-Prinzip“ als auffällig identifizierten Anordnungen prüfte die Bundesagentur monatlich. Auffällige Anordnungen des jeweils vergangenen Monats ordnete sie den gE zu. Sie stellte den gE die Daten über das Verfahren

„Mitarbeiterportal“ von SAP bereit. Darin konnten die Beschäftigten die ihnen zugeordneten Auffälligkeiten einsehen.

Die Bundesagentur nahm jeden Monat eine Stichprobe der monatlich erzeugten auffälligen Fälle. Diese Stichprobe überprüfte sie vor Abgang an die gE selbst, um grobe Fehler auszuschließen. Die Bundesagentur bezeichnete diesen Prozess als „zentrale Vorprüfung“. Nach dieser Vorprüfung sah sie noch 64 % aller Auffälligkeiten zum Versand vor.

Jede gE benannte sogenannte „Dispatcher“⁹, die in ihrem Mitarbeiterportal auf die gesamten Fälle der jeweiligen gE zugreifen konnten. Die Dispatcher wiesen die Fälle einzelnen Beschäftigten zur abschließenden Überprüfung zu (Abbildung 3).

Abbildung 3



Quelle: Bundesrechnungshof auf Grundlage der Erhebungen

Anschließend kontrollierten die Regionaldirektionen, ob die gE ihre Fälle nach den Vorgaben der Zentrale bearbeitet hatten.

Ein unmittelbarer Aufruf der Auffälligkeiten in ALLEGRO oder anderen Verfahren aus dem Mitarbeiterportal heraus war nicht möglich. Die Beschäftigten mussten jeweils die benötigten Verfahren starten und dort manuell die

⁹ In der Regel handelte es sich dabei um die Beauftragten für den Haushalt (BfH) in der gE.

Fallnummer eingeben, um die auffällige Zahlung aufzurufen. Das Ergebnis konnten sie im Mitarbeiterportal dokumentieren. Das dazugehörige Feld in der dafür vorgesehenen Eingabemaske war ein Freifeld. Die Bundesagentur wertete die in dem Freifeld vorgenommenen Einträge aus. Die häufigste Rückmeldung im Freifeld war „.“ (ein einzelner Punkt).¹⁰ Die dritthäufigste Rückmeldung war „ “ (eine einzelne Leertaste).

In einem weiteren Feld konnten die Beschäftigten den entstandenen oder durch die Prüfung vermiedenen Schaden in Euro angeben. Auch diese Angabe war freiwillig. Die gE meldeten nur in 0,15 % aller Fälle (139 von über 90 000) einen Vermögensschaden. Insgesamt belief sich der Vermögensschaden auf 23 492,40 Euro, mit einem durchschnittlichen Schaden von 169,01 Euro je Fall. In 161 weiteren Fällen hatten die gE zwar einen Fehler mit Vermögensschaden zurückgemeldet, dabei aber einen Schaden von 0 Euro angegeben.

Als Ergebnis ihrer Überprüfung konnten die Beschäftigten folgende Auswahl treffen:

- Bestätigt (vermuteter Fehler ist bestätigt)
- Fehlalarm (geprüft, aber kein Fehler festgestellt)
- Geschlossen ohne Untersuchung (Abschluss ohne Prüfung)

In einer Arbeitshilfe wies die Bundesagentur darauf hin, dass die Sachbearbeiter den Grund „Geschlossen ohne Untersuchung“ nicht auswählen dürfen.

In den Jahren 2017 bis 2019 stellte die Bundesagentur 73 773 auffällige Zahlungen fest und leitete diese an die gE zur Prüfung weiter. In nur 727 Fällen (1 %) bestätigten die gE den Fehler. In 48 417 Fällen meldeten sie Fehlalarm. Die restlichen Fälle schlossen sie entweder ohne Untersuchung ab oder sie hatten der Bundesagentur noch keine Rückmeldung gegeben (Tabelle 1).

¹⁰ Nach Abzug der bei der zentralen Vorprüfung erzeugten Einträge, diese meldeten den Text „zentral, aufgrund Analyse/Prüfung zentral“ zurück.

Tabelle 1

Ergebnisse der Überprüfung durch die gE

Jahr	Fälle insgesamt	Fehler	Fehlalarm	Geschlossen ohne Untersuchung	Noch nicht geprüft
2017	35 685	267	14 038	21 380	0
2018	21 164	256	19 013	1 895	0
2019	16 924	204	15 366	1 270	84
2017 bis 2019	73 773	727	48 417	24 545	84

Quelle: Eigene Auswertung von Erhebungsdaten.

Bei den zentralen Vorprüfungen¹¹ von Januar bis August 2018 prüfte die Bundesagentur insgesamt 2 471 Auffälligkeiten (18 % der von Januar bis August 2018 versandten Auffälligkeiten). Die Ergebnisse ordnete sie in fünf Kategorien ein:

- 1 – Fehler ersichtlich
- 2 – potenziell fehlerhaft
- 3 – komplexer Fall (Prüfung wird empfohlen)
- 4 – unschädlich
- 5 – falscher Treffer

Insgesamt stufte sie 61 % der Auffälligkeiten als „zwingend prüfwürdig“¹² ein. Somit betrachtete der Bereich EFM die von den fachlichen Bedarfsträgern vorgegebene Mindestquote von 60 % Treffergenauigkeit als erfüllt. In 13 % der Fälle sah der Bereich einen „Fehler ersichtlich“, in 39 % einen „Fehlalarm“ (Tabelle 2).

Tabelle 2

Ergebnisse der Überprüfung durch den Bereich EFM der Bundesagentur

Fehlerkategorie	Anteil in %
1 – Fehler ersichtlich	13
2 – potenziell fehlerhaft	21
3 – komplexer Fall (Prüfung wird empfohlen)	27
4 und 5 – unschädlich oder falscher Treffer	39

Quelle: Eigene Auswertung von Erhebungsdaten.

¹¹ Vor der Weiterleitung an die gE.

¹² Als „zwingend prüfwürdig“ gelten Fälle der Kategorien 1–3.

Die gE hatten für denselben Zeitraum nur für 1 % aller Auffälligkeiten zurückgemeldet, dass ein Fehler vorliege. In 94 % der Fälle meldeten sie „Fehlalarm“, 5 % der Fälle beendeten sie, ohne die Auffälligkeiten zu untersuchen. Die Bundesagentur schloss daraus, dass sich die Rückmeldungen der gE nicht dazu eigneten, die Auffälligkeitsanalysen zu verbessern. Ursprünglich hatte die Bundesagentur geplant, die Rückmeldungen der gE als primäre Datenbasis heranzuziehen, um die Analysen zu verbessern. Stattdessen musste sie die Erkenntnisse aus der Vorprüfung nutzen.

Der Bereich EFM hatte empfohlen, die Umsetzung des „Zwei-Augen-Prinzips“ mit dem dafür entwickelten Kontrollsystem und die anschließenden Fallprüfungen vor der Überführung in den Wirkbetrieb zuerst durch einzelne gE testen zu lassen. Die Bundesagentur verzichtete auf diese Tests in den gE.

3.2 Würdigung

Mit dem Einsatz des Vier-Augen-Prinzips sollen Fehler bei Anordnungen so weit wie möglich vermieden werden. Der Bundesagentur ist es nicht gelungen, die gE bei ALLEGRO durch technische Hilfsmittel und nachträgliche manuelle Prüfungen spürbar zu entlasten.

Die Mehrzahl der gE wandte das Verfahren zur Überprüfung auffälliger Anordnungen nicht korrekt an. Stattdessen ordneten sie Zahlungen im Vier-Augen-Prinzip an. Möglicherweise war das Verfahren zur Überprüfung von Auffälligkeiten für die gE zu aufwendig. Den geringen Nutzungsgrad durch die gE hätte die Bundesagentur daher frühzeitig zum Anlass nehmen müssen, das Verfahren zu evaluieren und falls erforderlich zu beenden.

Die wenigen Rückmeldungen der gE konnte die Bundesagentur aufgrund deren geringen Qualität nicht verwenden, um die Analysen zu verbessern. Ein wesentlicher Grund für die mangelnden Rückmeldungen war auch der Zeitpunkt des Versands der Auffälligkeiten. Die monatlich nachträglich erstellten Prüflisten hatten keinen zeitlichen Bezug mehr zu den ursprünglichen Anordnungen. Die Bundesagentur hat es versäumt, die Auffälligkeiten zeitnah zu übersenden.

Hinzu kam der breite Einsatz von Freifeldern in den Rückmeldemasken, sodass die gE diese nicht oder z. B. nur mit einem „.“ befüllten.

Unverständlich ist auch, weshalb die Bundesagentur bei ihrer Prüfung andere Fehlerkategorien nutzte als sie den gE im Verfahren vorgab. In jedem Fall hätte sie die Diskrepanz zwischen der selbst festgestellten hohen Fehlerquote und den niedrigen Fehlerquoten der gE aufklären müssen. Stattdessen ließ sie das Verfahren im Grunde unverändert weiterlaufen.

Warum sich die Empfehlung des Bereichs EFM nicht durchgesetzt hat, die Umsetzung des „Zwei-Augen-Prinzips“ mit dem dafür entwickelten Kontrollsystem und die anschließenden Fallprüfungen vor der Überführung in den flächendeckenden Wirkbetrieb zuerst durch einzelne gE testen zu lassen, ist nicht nachvollziehbar. Auch wenn es sich bei der Entwicklung und Umsetzung des Kontrollsystems nicht um ein Projekt im engeren Sinn handelte, hätte auch im vorliegenden Fall eine „Pilotphase“ unter Einbeziehung ausgewählter gE Erkenntnisse liefern können, um bei der Wirksamkeit des Verfahrens nachzusteuern. Damit hätten die Beschäftigten gezielter in die Entwicklung des Kontrollsystems eingebunden werden können.

3.3 Empfehlung

Wenn die erfolgreiche Einführung eines Verfahrens wesentlich von der Mitarbeit der gE abhängt, muss die Bundesagentur diese enger in die Entwicklung einbinden. Hierzu gehören auch Tests in ausgewählten gE vor der Aufnahme des Wirkbetriebs. In dem vorliegenden Fall unterschätzte sie den Aufwand für die gE bei Anwendung des „Zwei-Augen-Prinzips“. Die Bundesagentur muss die gE dazu anhalten, die Daten zu neuen Verfahren konsequent zurückzumelden. Dabei sollte sie möglichst auf Freifelder verzichten. Sie muss künftig darauf achten, dass neue Verfahren anhand von aussagekräftigen Rückmeldungen der gE zielgerichtet weiterentwickelt werden können.

3.4 Stellungnahme

Die Bundesagentur hat unsere Sichtweise, die gE würden das „Zwei-Augen-Prinzip“ aufgrund des damit verbundenen Aufwands selten nutzen, nicht geteilt. Sie hat eingeräumt, dass der mehrstufige Prozess zur Bearbeitung der ermittelten Auffälligkeiten optimierungsfähig gewesen sei. Warum die Nutzung des „Zwei-Augen-Prinzips“ in ALLEGRO jedoch weit hinter den Erwartungen zurückgeblieben sei, habe sie letztlich nicht objektiv ermitteln können.

Das Fachverfahren ALLEGRO und die Prüfungen zum „Zwei-Augen-Prinzip“ seien fast zeitgleich eingeführt worden. Dies habe bei den Beschäftigten der gE zu erhöhten Anforderungen geführt. Die Einführung des neuen Prüfverfahrens habe die Beschäftigten zusätzlich belastet. Beim Bearbeiten von Auffälligkeiten hätten Beschäftigte in den gE einen Medienbruch erlebt, der die Akzeptanz nicht gefördert habe.

Beschäftigte der gE seien nicht zusätzlich belastet worden. Die gE hätten eigenständig und freiwillig entscheiden können, ob sie das „Zwei-Augen-Prinzip“ anwenden oder nicht.

Die Bundesagentur hat erwidert, dass sie das Verfahren gerne vor der Anwendung getestet hätte. Dies sei jedoch nicht möglich gewesen. Hochwertige Trainingsdaten hätten nicht vorgelegen, da das Fachverfahren ALLEGRO neu eingeführt worden sei. Sie habe daher einen Ansatz gesucht, der ohne Testdaten durchgeführt werden konnte.

Die Bundesagentur hat uns teilweise zugestimmt, dass aufgrund des geringen Nutzungsgrads das Verfahren zu evaluieren oder ggf. zu beenden gewesen sei. Sie habe 2AP-ALLEGRO wegen der von uns genannten Mängel evaluiert und darüber hinaus einen „Health-Check“ durch ein Unternehmen durchführen lassen. Der Health-Check habe keine gravierenden Mängel aufgezeigt. Erkenntnisse aus der Evaluierung seien dennoch in die Zeitplanung des Bereichs EFM für das Jahr 2020 geflossen.

Die Bundesagentur hat unserer Feststellung widersprochen, dass das Verfahren gescheitert ist. Sie sei nicht wegen fehlender systematischer Tests oder der fehlenden Einbindung der gE, sondern wegen zu hoher Fehlerquoten zum Vier-Augen-Prinzip zurückgekehrt.

Gleichwohl hat sie an anderer Stelle argumentiert, dass sie das „Zwei-Augen-Prinzip“ nicht wegen hoher Fehlerquoten abgeschafft habe. Vielmehr sei es notwendig gewesen, zum Vier-Augen-Prinzip zurückzukehren, um Rentenanpassungen in ALLEGRO automatisch verarbeiten zu können.

Die Bundesagentur hat eingeräumt, dass viele gE aufgrund von Freifeldern in den Rückmeldemasken eine ausführliche Beschreibung ihrer Prüfungsergebnisse umgehen konnten. Auch im entsprechenden Feld zu dem Vermögensschaden hätten die gE nur in den wenigsten Fällen den vorgegebenen Wert

von „0,00 Euro“ angepasst. Diese Eingabemasken seien fest implementiert gewesen. Die Bundesagentur hätte diese nur mit zusätzlichem Aufwand anpassen können.

Des Weiteren sehe SAP im Standard nur drei Fehlerkategorien vor. Diese seien nicht ausreichend gewesen, um die Qualität einer Stichprobe bei der Vorprüfung angemessen beurteilen zu können. Deshalb habe die Bundesagentur intern abweichende Fehlerkategorien verwendet.

Die Bundesagentur hat eingeräumt, dass die Rückmeldungen nicht genügend aussagekräftig waren. Sie habe angeboten, die gE zu unterstützen, um die Qualität zu verbessern und ihre Prüfstrategien zu optimieren.

Die Bundesagentur hat uns zugestimmt, dass der relativ große Zeitraum zwischen Anweisung und hierzu ggf. festgestellter Auffälligkeit zu verringern sei.

3.5 Abschließende Würdigung

Die Bundesagentur stimmt uns zu, dass das Verfahren optimierungsfähig gewesen sei. Wir bleiben bei unserer Einschätzung, dass das Verfahren für die gE aufwendig gewesen ist und diese es aus diesem Grund möglicherweise selten genutzt haben. Die Bundesagentur hat selbst eingeräumt, dass u. a. die Einführung von ALLEGRO zu erhöhten Anforderungen bei den gE geführt hatte.

Die Bundesagentur hat das „Zwei-Augen-Prinzip“ ohne vorherige systematische Tests flächendeckend eingeführt. Sie hätte das Verfahren vorher mit einzelnen gE testen und Trainingsdaten generieren können. Damit hätte sie vermutlich hohe Fehlerquoten vermindern können. Die gE wären in der Folge weniger belastet gewesen. Die Begründung der Bundesagentur, warum das „Zwei-Augen-Prinzip“ letztendlich gescheitert ist, ist widersprüchlich. An einer Stelle ihrer Stellungnahme verneint sie, dass sie das „Zwei-Augen-Prinzip“ aufgrund zu hoher Fehlerquoten beendet hat, an anderer Stelle bestätigt sie dies. Wir bleiben daher bei unserer Würdigung und Empfehlung.

Die ungeeigneten Eingabemasken des Verfahrens haben die Qualität der Rückmeldungen vermindert. Daher hätte die Bundesagentur diese anpassen müssen, auch wenn dies zusätzlichen Aufwand verursacht hätte. Die Bundesagentur hat selbst eingeräumt, dass die Rückmeldungen nicht genügend aussagekräftig gewesen sind. Wir bleiben auch in diesem Punkt bei unserer Empfehlung.

4 Auswahl und Eignung der analytischen Verfahren

Bei den im „Zwei-Augen-Prinzip“ angeordneten Fällen handelte es sich u. a. um Einmalbedarfe und Nachzahlungen bis 500 Euro.¹³ Die Bundesagentur verwendete neben deterministischen Methoden¹⁴ auch Methoden des maschinellen Lernens, um Auffälligkeiten in den Zahlungen zu erkennen. Mit den deterministischen Verfahren versuchte die Bundesagentur anhand von eindeutig festgelegten Vorbedingungen, Auffälligkeiten in den Zahlungen zu erkennen. Als analytische Verfahren bezeichnete die Bundesagentur jene Verfahren, die Methoden des maschinellen Lernens verwendeten. Diese versuchten auf Basis statistischer Auswertung bisher getätigter Eingaben, Muster und Gesetzmäßigkeiten in den Daten zu erkennen.

4.1 Sachverhalt

Die Bundesagentur nutzte die „Predictive Analysis Library“ (PAL)¹⁵ des Unternehmens SAP SE, um die Methoden zu implementieren. PAL ist eine Programmbibliothek, die eine Vielzahl von datenanalytischen Funktionen enthält. Ein Nutzer der PAL braucht diese Funktionen nicht selbst aufwendig zu programmieren, sondern kann diese direkt verwenden. Zu PAL existiert ein Benutzerhandbuch, das von SAP erstellt wurde. Dieses stellt u. a. die Bedeutung einer geeigneten Parameterwahl heraus und erklärt, dass unterschiedliche Parameter auch Ergebnisse unterschiedlicher Qualität produzieren können. Es empfiehlt deshalb, bei Parametern unterschiedliche Werte zu testen, um ein gutes Ergebnis zu erreichen.

Jeden Monat berechnete die Bundesagentur für die im Vormonat im „Zwei-Augen-Prinzip“ angeordneten Fälle zunächst 50 vorab definierte Kennzahlen mittels mathematischer Methoden. Diese Kennzahlen wertete sie anschließend in einem zweistufigen Verfahren mit der PAL aus. Im ersten Schritt aggregierte die Bundesagentur die Kennzahlen, um *„die nachgelagerte Anomalieerkennung zu verbessern“*. Im zweiten Schritt folgte ein Verfahren zur Auffälligkeitserkennung (sog. „Anomaly Detection“). Beim Auswerteprozess unterschied die Bundesagentur je Anwendungsfall nach Einmalbedarfen und Nachzahlungen.

¹³ Weitere Zahlungen, die im „Zwei-Augen-Prinzip“ angeordnet wurden, waren nicht Bestandteil der automatisierten Prüfungen.

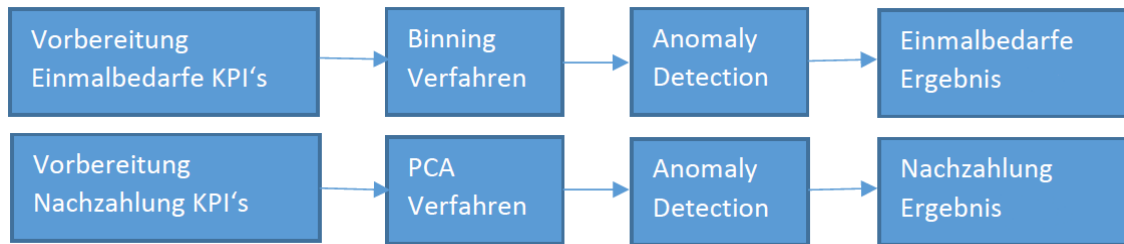
¹⁴ Diese kamen nur für die Erkennung von Doppelzahlungen zum Einsatz.

¹⁵ SAP HANA Predictive Analysis Library.

Für Einmalbedarfe setzte die Bundesagentur das „Binning-Verfahren“ ein, um die Kennzahlen zu aggregieren.¹⁶ Für Nachzahlungen hingegen nutzte sie das Verfahren „PCA“.¹⁷ In Abbildung 4 sind beide Anwendungsfälle schematisch dargestellt.

Abbildung 4

Schematische Darstellung zu den PAL Verfahren



Quelle: Technische Dokumentation 2-AP ALLEGRO auf HANA.

In der technischen Dokumentation begründete die Bundesagentur nicht,

- warum sie zwei verschiedene Verfahren der PAL einsetzte, um die Kennzahlen zu aggregieren und
- warum sie genau diese Verfahren ausgewählt hatte.

Für die Anomaly Detection setzte die Bundesagentur in beiden Anwendungsfällen das gleiche Verfahren ein, das sogenannte „Clustering“¹⁸. Dieses Verfahren des unüberwachten Lernens¹⁹ ermöglichte es, die Anordnungen im „Zwei-Augen-Prinzip“ anhand der berechneten und aggregierten Kennzahlen darzustellen. Das Verfahren erkannte anschließend jeweils in den Anordnungen eines vergangenen Monats automatisch „Häufungen“ (Cluster) ähnlicher Anordnungen. Die Anzahl der Cluster konnte sie dabei entweder fest vorgeben oder durch das Verfahren selbst ermitteln lassen.

¹⁶ Dieses gliederte die berechneten Kennzahlen und sortierte ähnliche Kennzahlen in sogenannte „Bins“.

¹⁷ Dabei kam die mathematische Methode „Hauptkomponentenanalyse“ zum Einsatz. Diese berechnete aus den Kennzahlen einen einzelnen Ergebnisvektor, den die Anomaly Detection dann verarbeiten konnte.

¹⁸ Unter Clusteranalysen (Clustering-Algorithmen, gelegentlich auch: Ballungsanalyse) versteht man Verfahren zur Entdeckung von Ähnlichkeitsstrukturen in (meist relativ großen) Datenbeständen. Die so gefundenen Gruppen von „ähnlichen“ Objekten werden als Cluster bezeichnet, die Gruppenzuordnung als Clustering.

¹⁹ Beim unüberwachten Lernen suchen Algorithmen nach Strukturen und Mustern in den Daten, ohne dass ihnen ein Ziel vorgegeben wurde.

Für Einmalbedarfe und Nachzahlungen setzte die Bundesagentur unterschiedliche Parameter ein. So nutzte sie für Einmalbedarfe eine fest vorgegebene²⁰ Anzahl von vier Clustern. Bei den Nachzahlungen erlaubte sie dem Verfahren hingegen, die Anzahl von Clustern selbst zu bestimmen²¹. Folglich ermittelte das Verfahren für die Quelldaten verschiedener Monate jeweils eine unterschiedliche Anzahl von Clustern. Eine Statistik oder Auswertung erstellte die Bundesagentur hierzu nicht. In der technischen Dokumentation stellte sie nicht dar,

- warum sie bei Einmalbedarfen die Anzahl von vier Clustern für geeignet hielt und
- warum sie bei Nachzahlungen die Anzahl von Clustern nicht festlegte.

Ein weiterer Parameter legte den Soll-Anteil an Auffälligkeiten in den Quelldaten fest. Die Bundesagentur wählte hierbei den Wert 5 %. Nachdem das Verfahren die Cluster ermittelte, stuft es somit jene 5 % der Anordnungen als auffällig ein, die die geringste Ähnlichkeit²² mit den Clustern aufwiesen.

Der Parameter mit der Bezeichnung „INIT_TYPE“²³ bestimmte die Methode, mit welcher die Anfangswerte für die Cluster²⁴ ausgewählt werden. Die Bundesagentur verwendete den voreingestellten Standardwert der PAL. Diese Entscheidung hatte sie nicht begründet.

Bei einer frei wählbaren Anzahl von Clustern kann das Verfahren dynamischer auf äußere Einflüsse reagieren (z. B. Gesetzesänderungen). Das „Clustering“-Verfahren mit einer festen Anzahl von Clustern erkannte im August 2018 ca. 20 % mehr Auffälligkeiten als üblich. Es hatte insbesondere Fälle zur „Schulpauschale“ als auffällig markiert, die fachlich nicht auffällig waren. Infolgedessen entfernte die Bundesagentur diese Fälle vor Versand der Auffälligkeitslisten an die gE manuell.

Im September 2018 stuft der Bereich EFM die Clusteranalyse als noch nicht ausreichend genau ein. Er schlug u. a. vor, den Ansatz um eine Komponente

²⁰ Ist die Anzahl fest vorgegeben, so handelt es sich um den „k-means“-Algorithmus.

²¹ Bei einer frei wählbaren Anzahl an Clustern handelt es sich um den „g-means“-Algorithmus.

²² Demnach waren Anordnungen, die den größten Abstand zu den Clusterzentren aufwiesen, besonders auffällig.

²³ INIT_TYPE.

²⁴ Clusterzentren.

des überwachten maschinellen Lernens²⁵ zu erweitern. Diese sollte lernen, anhand von bereits manuell geprüften und als auffällig oder unauffällig eingestuft Fällen neue Fälle zu klassifizieren. Methoden des überwachten Lernens benötigen im Allgemeinen eine große Anzahl an qualitativ hochwertigen Trainingsdaten, um verlässliche Ergebnisse liefern zu können. Da die Bundesagentur nicht über die dazu erforderlichen Daten verfügte (Tz. 3.1.2), ergänzte sie die Clusteranalyse nicht um einen Ansatz des überwachten Lernens.

Bei Methoden des maschinellen Lernens sind auch im Wirkbetrieb technische Anpassungen und Analysen erforderlich. Die Bundesagentur sah dafür einen einzelnen Beschäftigten mit einem Drittel seiner Arbeitszeit vor. Im Laufe des Wirkbetriebs passte dieser die Parameter der Methoden des maschinellen Lernens nicht an. Der Bereich EFM hatte bereits zu Beginn der Projektphase des „Zwei-Augen-Prinzips“ in ALLEGRO eine mangelnde Stellenausstattung festgestellt und dies den fachlichen Bedarfsträgern mehrfach mitgeteilt. In den Jahren 2016 bis 2020 sah die Bundesagentur für das „Zwei-Augen-Prinzip“ (fachlich und technisch) maximal drei Vollzeitäquivalente (VZÄ) vor, zuletzt noch 1,3 VZÄ.

4.2 Würdigung

Eine technische Dokumentation sollte es sachkundigen Dritten ermöglichen, die Auswahl von Methoden des maschinellen Lernens nachzuvollziehen. Bei einer Clusteranalyse ist die Aufbereitung der Quelldaten ein Schritt, der die Ergebnisse wesentlich beeinflussen kann. Angesichts dessen ist die fehlende Dokumentation hierzu nicht sachgerecht.

Das Ergebnis der Clusteranalyse wird darüber hinaus von der Anzahl der Cluster beeinflusst. Das Verfahren kann bei einer ungünstig gewählten Anzahl an Clustern zu ungewünschten Ergebnissen führen. Trotzdem dokumentierte die Bundesagentur nicht, warum sie sich bei diesem kritischen Parameter für die von ihr gewählten Werte entschieden hatte. Es ist unverständlich, dass sie die Ergebnisse der Clusteranalyse nicht wenigstens in Abhängigkeit von diesem Parameter auswertete. Eine solche Auswertung hätte die Bundesagentur bei der Wahl der Anzahl von Clustern unterstützen können.

²⁵ Beim überwachten maschinellen Lernen werden Algorithmen anhand von vorgegebenen Beispielen trainiert, die bereits korrekte Antworten enthalten. Für die Trainingsdaten ist das gewünschte Ergebnis bekannt und kann dem Algorithmus vorgegeben werden.

Auch beim Parameter „INIT_TYPE“ hat die Bundesagentur es versäumt, ihre Auswahl zu begründen. Es ist unverständlich, warum sie hier den PAL-Standardwert nutzte und nicht unterschiedliche Parameterwerte testete, wie im PAL-Benutzerhandbuch empfohlen.

Insgesamt haben wir das Vorgehen und die Dokumentation der Bundesagentur bei der Wahl der Verfahren und Parameter als unzureichend bewertet.

Da die Bundesagentur den Algorithmus jeweils nur mit den Daten des aktuellen Monats trainierte, kann sie auch nicht ausschließen, dass von Grund auf eigentlich identische oder ähnliche Fälle in einem Monat als Auffälligkeiten ausgewiesen wurden, im nächsten dann aber nicht mehr. Somit konnte das System auch keine sinnvollen Hinweise an die Sachbearbeiter der gE liefern, damit diese eine nachhaltige Qualitätssteigerung erreichen. Beim „Neulernen“ des Algorithmus hätte die Bundesagentur Fälle aus den Vormonaten mit einbeziehen müssen.

Der Bereich EFM stufte die Ergebnisse der Clusteranalyse selbst als noch nicht ausreichend ein und wollte sie daraufhin mit Methoden des überwachten Lernens ergänzen. Diesen erfolgsversprechenden Ansatz konnte die Bundesagentur aufgrund der geringen Qualität der Rückmeldungen aus den gE nicht weiterverfolgen. Methoden des maschinellen Lernens sollten stetig weiterentwickelt und der Datenlage angepasst werden. Dazu hätte die Bundesagentur bereits zu Beginn der maschinellen Unterstützung des „Zwei-Augen-Prinzips“ in ALLEGRO stärker darauf hinwirken müssen, hochwertige Trainingsdaten zu erheben.

Die Bundesagentur hatte das „Zwei-Augen-Prinzip“ u. a. eingeführt, um 300 Stellen und 370 Ermächtigungen einzusparen. Ihre Entscheidung, für die personelle Ausstattung des „Zwei-Augen-Prinzips“ nur maximal drei VZÄ anzusetzen, erscheint uns vor diesem Hintergrund nicht sachgerecht. Die mangelnde Personalausstattung war der Bundesagentur bekannt gewesen, dennoch stockte sie diese nicht angemessen auf. Infolgedessen konnten die Beschäftigten das System nur geringfügig anpassen und optimieren. Dies hat wesentlich zur geringen Güte der Methoden des maschinellen Lernens und damit letztendlich zum Scheitern des „Zwei-Augen-Prinzips“ beigetragen.

4.3 Empfehlung

Wir haben der Bundesagentur empfohlen, ihre Parameterwahl bei Verfahren des maschinellen Lernens künftig zu begründen und dies angemessen zu dokumentieren. Insbesondere sollte sie bei unbrauchbaren Ergebnissen ihre Parameterwerte hinterfragen. Sie sollte dazu unterschiedliche Parameterwerte testen, die entsprechenden Ergebnisse hierzu auswerten und dies nachvollziehbar dokumentieren.

Die Bundesagentur sollte bei weiteren Projekten zum maschinellen Lernen den Zeitbezug der Trainingsdaten berücksichtigen, mit denen der Algorithmus „lernt“ und auch diesen bei zweifelhaften Ergebnissen variieren. Sie sollte Empfehlungen aus dem offiziellen Benutzerhandbuch zur PAL und wissenschaftliche Erkenntnisse zu analytischen Verfahren soweit möglich beachten.

Bei ihren Projekten zum maschinellen Lernen sollte die Bundesagentur mehr Wert auf die Qualität der Trainingsdaten legen. Auch im Wirkbetrieb sollte sie die Datenqualität überprüfen und sicherstellen, dass diese weder die Funktion noch die Weiterentwicklung der Algorithmen einschränkt.

Bei der Personalausstattung für Verfahren des maschinellen Lernens sollte die Bundesagentur berücksichtigen, dass diese einen höheren Aufwand für Anpassungen und Weiterentwicklung im Wirkbetrieb mit sich bringen können als andere IT-Verfahren. Die Bundesagentur sollte dies bei ihren weiteren Projekten beachten. Hinweisen der operativen Ebene zu mangelnder Personalausstattung oder -qualifikation sollte sie nachgehen.

4.4 Stellungnahme

Die Bundesagentur hat mitgeteilt, dass sie die Wahl ihrer Methoden sowie die zugehörigen Parameter ausreichend begründet habe. Sie habe bei der angewandten Methode z. B. unterschiedliche Parameterwerte getestet, um die Anzahl der Cluster zu bestimmen. Die Bundesagentur hat eingeräumt, dass eine technische Dokumentation hierzu fehle.

Die Bundesagentur hat uns zugestimmt, dass eine angemessene Personalausstattung für vergleichbare Projekte unerlässlich ist.

4.5 Abschließende Würdigung

Wir bleiben bei unseren Würdigungen und Empfehlungen. Die Bundesagentur hat die Auswahl ihrer Parameterwerte nicht ausreichend dokumentiert. Z. B. bleibt unklar, weshalb sie bei Einmalzahlungen die Anzahl der Cluster anders bestimmte als bei Nachzahlungen.

Das Verfahren hat unstrittig keine zufriedenstellenden Ergebnisse geliefert. Ungeachtet dessen hätten wir zu der Auswahl aller Parameterwerte eine ausführliche Begründung, ggf. entsprechende Tests und eine dazugehörige Dokumentation erwartet, die Dritte nachvollziehen können. Die Bundesagentur konnte uns diesbezügliche Belege trotz Nachfrage nicht übermitteln.

Wir begrüßen, dass die Bundesagentur uns hinsichtlich der Personalausstattung von ZAP-ALLEGRO zustimmt. Zu unseren weiteren Empfehlungen hat sich die Bundesagentur nicht geäußert.

5 Wirtschaftlichkeit des Einsatzes analytischer Verfahren

5.1 Sachverhalt

Nach § 7 Absatz 2 Satz 1 BHO sind für alle finanzwirksamen Maßnahmen angemessene Wirtschaftlichkeitsuntersuchungen durchzuführen. Zuständig hierfür ist die Organisationseinheit, die für die finanzwirksame Maßnahme verantwortlich ist. Sie muss die Wirtschaftlichkeitsuntersuchung bei der Planung einer neuen Maßnahme im Haushaltsaufstellungsverfahren, bei der Realisierung der Maßnahme sowie zur Erfolgskontrolle während und nach der Umsetzung durchführen.²⁶ Die Ergebnisse muss die maßnahmenverantwortliche Organisationseinheit nachvollziehbar festhalten.

Die Bundesagentur führte für das Verfahren mit Kosten von 7 Mio. Euro (Tabelle 3) keine Wirtschaftlichkeitsuntersuchung durch. Sie hatte auch keine Ziele, beispielsweise sinkende Fehlerquoten bei den Anordnungen, definiert, die sie messen und nachhalten konnte. Eine Erfolgskontrolle sah sie nicht vor.

²⁶ VV Nummer 2 zu § 7 BHO.

Tabelle 3

Kosten „Zwei-Augen-Prinzip“ ALLEGRO

Jahr	Entwicklung	Datenaufbereitung	Hardware	Personal-kosten	Summe
2015	60 876	0	0		60 876
2016	903 924	451 962	500 000	248 700	2 104 586
2017	632 071	316 035	500 000	248 700	1 696 806
2018	496 778	248 389	500 000	248 700	1 493 867
2019	333 851	166 926	333 300	248 700	1 082 777
2020	144 884	72 442	150 000	124 350	491 676
2015 bis 2020					6 930 588

Quelle: Eigene Darstellung des Bundesrechnungshofes, Basis: RCE-EFM – 1430.

In ihrem Antrag zur Ausnahmegenehmigung zum Betrieb von ALLEGRO gab die Bundesagentur an, dass sie mit dem „Zwei-Augen-Prinzip“ bei Einmalzahlungen bis 500 Euro rechnerisch 600 VZÄ einsparen könne. Dies entspräche 39,6 Mio. Euro. Dem standen geplante Kosten von einmalig rund 1,0 Mio. Euro für die Entwicklung und jährlich 0,5 Mio. Euro für die Datenversorgung,-lieferung und -aufbereitung gegenüber. Zusätzlich ging sie von nicht haushaltswirksamen Kosten in Höhe von jährlich rund 0,6 Mio. Euro aus, um auffällige Fälle zentral nachhalten zu können. *„Durch das Erreichen eines höheren Automatisierungsgrades sollte mittelfristig eine deutliche Reduzierung dieser Kosten angestrebt werden.“* Die zusätzlichen Stellen (300) und Ermächtigungen (370) hatte sie zum Ende des Jahres 2016 gestrichen.

Die gE zahlten im Jahr 2019 rund 28 Mrd. Euro Leistungen über ALLEGRO aus. Davon wurden im „Zwei-Augen-Prinzip“ mit EFM nur 67 Mio. Euro (0,25 %) geprüft. Dabei meldeten die gE 204 Fehler. Die Fehler betrafen Zahlungen in Höhe von rund 26 000 Euro.

Seit dem Jahr 2017 war der Bundesagentur bekannt, dass

- die Prüfalgorithmen eine geringe Treffergenauigkeit auswiesen und
- die gE aufgrund des zusätzlichen Aufwandes das „Zwei-Augen-Prinzip“ kaum nutzten.

Die Bundesagentur überdachte ihre im Ausnahmeantrag getroffenen Annahmen nicht.

5.2 Würdigung

Die Bundesagentur hätte für das neue Verfahren eine Wirtschaftlichkeitsuntersuchung durchführen müssen. Da sie zu Beginn des neuen Verfahrens keine eindeutigen Ziele festgelegt hat, konnte sie weder während des Betriebs noch zum Abschluss der Maßnahme eine belastbare Erfolgskontrolle durchführen.

Die Bundesagentur hätte den entstandenen Aufwand in den gE bei ihrer Wirtschaftlichkeitsuntersuchung berücksichtigen müssen. Sie hat es zu keinem Zeitpunkt geschafft, einen wesentlichen Teil der Zahlungen in ALLEGRO auf Auffälligkeiten zu überprüfen. Damit hat das Verfahren nicht die notwendigen Erkenntnisse geliefert, um ihre Analysen zu verbessern. Im Ergebnis hat sie weder die Wirtschaftlichkeit noch die Wirksamkeit des Mitteleinsatzes belegen können. Sie hat nicht angemessen auf die geringe Fehlerrückmeldung der gE reagiert. Auch ohne eine Wirtschaftlichkeitsuntersuchung in der Planungsphase der Maßnahmen hätte sie ihre Annahmen aus dem Antrag zur Ausnahmege-nehmigung überprüfen und das Verfahren entweder neu ausrichten oder rechtzeitig einstellen können.

5.3 Empfehlung

Die Bundesagentur muss bei finanziell wirksamen Verfahren Wirtschaftlichkeitsuntersuchungen durchführen und ihre Annahmen und Ziele angemessen dokumentieren. Sie hat zu Beginn solcher Verfahren klare und messbare Ziele zu definieren. Ihre Annahmen und Ziele muss sie regelmäßig überprüfen.

Bei einer Zielabweichung hat sie über geeignete Schritte zu beraten. Hierzu gehören mögliche Gegenmaßnahmen oder auch die vorzeitige Einstellung eines Verfahrens.

5.4 Stellungnahme

Die Bundesagentur hat eingeräumt, dass sie für das Verfahren „ALLEGRO-EFM“ keine gesonderte Wirtschaftlichkeitsuntersuchung erstellt habe. Die wesentlichen Teile des Verfahrens sollten in anderen Projekten entwickelt werden. Für diese Projekte lägen Wirtschaftlichkeitsuntersuchungen vor. Das Verfahren „ALLEGRO-EFM“ sei in diesen Wirtschaftlichkeitsuntersuchungen ausreichend berücksichtigt.

Weiter hat die Bundesagentur erklärt, sie habe sämtliche notwendigen wirtschaftlichen Betrachtungen durchgeführt und den Aufsichtsgremien berichtet. Zu unseren Empfehlungen hat sich die Bundesagentur nicht geäußert.

5.5 Abschließende Würdigung

Wir bleiben in diesem Punkt bei unseren Würdigungen und Empfehlungen. Für ein Verfahren mit geschätzten Kosten in Höhe von 7 Mio. Euro hätte die Bundesagentur eine **gesonderte** Wirtschaftlichkeitsuntersuchung durchführen müssen. Die Wirtschaftlichkeit lediglich zusammen mit anderen Projekten zu betrachten, halten wir in diesem Fall für nicht ausreichend.

Romers

Rabenschlag

Essers

Fasswald