Bundesrechnungshof

Abschließende Mitteilung

an die
Bundesagentur für Arbeit

über die Prüfung

Einsatz künstlicher Intelligenz und analytischer Verfahren in der Bundesagentur für Arbeit

Teil 2 – Verfahren zur Kassensicherheit in ERP

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

Die Mitteilung des Bundesrechnungshofes ist urheberrechtlich geschützt.

Inhaltsverzeichnis

Abkürzungsverzeichnis 4

0 Zusammenfassung 5

1 Vorbemerkung 10

2 Informationen zum Projekt ERP-2AP 12

3 ERP-2AP – Auswahl und Eignung der Prüfregeln 15

3.1 Sachverhalt 15

3.2 Würdigung 19

3.3 Empfehlung 20

3.4 Stellungnahme 20

3.5 Abschließende Würdigung 21

4 ERP-2AP – Auswahl und Eignung der analytischen Verfahren 22

4.1 Sachverhalt 22

4.2 Würdigung 25

4.3 Empfehlung 26

4.4 Stellungnahme 27

4.5 Abschließende Würdigung 27

5 ERP-2AP – Projektmanagement, Einführung und Wirtschaftlichkeit 28

5.1 Sachverhalt 28

5.2 Würdigung 31

5.3 Empfehlung 32

5.4 Stellungnahme 33

5.5 Abschließende Würdigung 34

6 Strategie und Koordination der Entwicklungen zur Kassensicherheit 35

6.1 Sachverhalt 35

6.2 Würdigung 36

6.3 Empfehlung 36

6.4 Stellungnahme 37

6.5 Abschließende Würdigung 37

7 Übergreifende Bewertung 37

7.1 Stellungnahme 38

7.2 Abschließende Würdigung 38

Anhang – Arbeitspakete im Projekt ERP-2AP

Abkürzungsverzeichnis

AP	Arbeitspaket
BestMaVB-HKR	Bestimmungen über die Mindestanforderungen für den Einsatz automatisierter Verfahren im Haushalts-, Kassen- und Rechnungswesen des Bundes
BMAS	Bundesministerium für Arbeit und Soziales
BMF	Bundesministerium der Finanzen
Bundesagentur	Bundesagentur für Arbeit
EFM	Enterprise Fraud Management
ERP	Enterprise Resource Planning
ERP-2AP	Abgesichertes Zwei-Augen-Prinzip in ERP
gE	gemeinsame Einrichtungen
KI	künstliche Intelligenz
nzStpp	nachträgliche zufallsorientierte Stichproben
PLA	Projektlenkungsausschuss
RD	Regionaldirektion
WiBe	Wirtschaftlichkeitsbetrachtung

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 Projekt „Abgesichertes Zwei-Augen-Prinzip in ERP" (ERP-2AP) sowie zur strategischen Ausrichtung und Koordination der Entwicklungen zur Kassensicherheit in der Bundesagentur. Die Feststellungen zum Verfahren „ALLEGRO-EFM" beschreibt der Bundesrechnungshof in einer weiteren Teilprüfungsmitteilung.

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

0.1 Im Projekt ERP-2AP entwickelte und implementierte die Bundesagentur deterministische und analytische Verfahren, um potenziell fehlerhafte Anordnungen im Verfahren ERP im „Zwei-Augen-Prinzip" zu identifizieren. Deterministische Verfahren zielen auf eindeutig festgelegte Vorbedingungen ab, u. a. gesetzliche Vorgaben oder Weisungen. Analytische Verfahren nutzen Methoden des maschinellen Lernens. Die Bundesagentur wollte mit dem Projekt eine höhere Treffergenauigkeit als bei den bisher genutzten Stichprobenverfahren erreichen und somit die Fehlerquote bei den Anordnungen nachhaltig verringern. (Tz. 2)

0.2 Die Bundesagentur entwickelte auf der Grundlage der im Projekt eingesetzten Verfahren eine Vielzahl von Prüfregeln, mit denen risikobehaftete Prüffälle identifiziert werden. Hierunter war nur eine mit besonders hohem Aufwand entwickelte Prüfregel, die auf einem analytischen Verfahren beruhte. Die Auswahl der Prüfregeln hat die Bundesagentur rein fachlich begründet. 37 % der Prüffälle stammten aus Prüfregeln mit niedriger oder unbestimmter Treffergenauigkeit. Die Prüfregeln sollen mehr als fünfmal so viele Prüffälle wie bisher generieren. Die Bundesagentur läuft dadurch Gefahr, die Beschäftigten zu überlasten, ohne eine hohe Treffergenauigkeit zu gewährleisten. Dies hätte zur Ablehnung des Systems und damit dessen Scheitern führen können.

Das Berichtswesen für die Prüfregeln war unvollständig. Es enthielt keine Kriterien, nach denen das System beispielsweise

Warnmeldungen versendete. Auch erzeugte es keine Handlungsempfehlungen für die fachlichen Bedarfsträger und die Linienorganisation. Darüber hinaus hat es die Bundesagentur bisher versäumt, einen Prozess zur Anpassung der Prüfregeln im Wirkbetrieb aufzusetzen. Sie hat nicht sicherstellen können, dass sie Änderungsbedarfe an den Prüfregeln rechtzeitig erkennen und umsetzen kann.

Die Bundesagentur sollte die Prüfregeln optimieren, unter Beachtung von Wirtschaftlichkeitskriterien einsetzen und auf ein handhabbares Maß begrenzen. Ein aussagefähiges Berichtswesen ist dafür unerlässlich.

Die Bundesagentur hat dargestellt, dass auch die Prüfregeln mit wenigen Prüffällen oder niedriger Treffergenauigkeit fachlich erforderlich seien. Sie hat zugesagt, ihre Prüfregeln und deren Ergebnisse so auszuwählen, dass die Anzahl an Prüffällen die Beschäftigten nicht überfordern werde. Darüber hinaus hat sie zugesichert, die Empfehlungen zum Berichtswesen umzusetzen. (Tz. 3)

0.3 Zur Prüfung auf auffällige Beträge nutzte die Bundesagentur ein Verfahren des überwachten maschinellen Lernens. Dieses sollte auch aus den Rückmeldungen der Beschäftigten lernen. Obwohl die Bundesagentur wusste, dass die Treffergenauigkeit der darauf aufbauenden Prüfregel bei 0 % lag, wollte sie diese in den Wirkbetrieb überführen. Ihr war aus den Erfahrungen mit ALLEGRO-EFM bekannt, wie sehr eine schlechte Treffergenauigkeit die Akzeptanz und den Erfolg neuer Verfahren beeinträchtigt. Dennoch hat sie darauf verzichtet, das Verfahren vor Einführung anhand von Trainingsdaten erneut anzupassen und weiterzuentwickeln. Für diese Aufgabe, die spezielles Fachwissen voraussetzt, hatte sie auch nicht die personellen Voraussetzungen geschaffen.

Der Bundesrechnungshof hat empfohlen, analytische Verfahren vor einer Freigabe für den Wirkbetrieb immer mit einer Güteprüfung zu evaluieren und die Erkenntnisse daraus geeignet zu berücksichtigen. Bei Modellen des überwachten maschinellen Lernens sollte die Bundesagentur die Datenqualität berücksichtigen und diese bei niedriger Datenqualität nicht einsetzen. Sie sollte zudem prüfen, inwieweit sie für

Verfahren des maschinellen Lernens und der künstlichen Intelligenz wissenschaftlich spezialisiertes Personal benötigt oder ihre Beschäftigten geeignet qualifizieren kann.

Die Bundesagentur hat dargestellt, sie habe die Regel zur Prüfung auffälliger Beträge in den Wirkbetrieb überführen wollen, um wenigstens zu einer analytischen Prüfregel weitere Erkenntnisse zu gewinnen. Auch habe sie mittlerweile Stellen mit Datenanalysten/Datenwissenschaftlern besetzen können. Damit wolle sie der Empfehlung des Bundesrechnungshofes zu diesem Punkt nachkommen.

Der Bundesrechnungshof hält an seinen Würdigungen und Empfehlungen fest. Modelle des überwachten maschinellen Lernens sollte die Bundesagentur nur bei ausreichender Datenqualität einsetzen. (Tz. 4)

0.4 Die Bundesagentur erstellte ihre Projektbeendigungskriterien erst knapp zwei Jahre nach Projektgenehmigung. Wegen dieser Verzögerung haben dem Projektlenkungsausschuss für die laufende Bewertung des Projektes entscheidende handlungsleitende Maßstäbe gefehlt. Obwohl die Mehrzahl der Kriterien zur Projektbeendigung im Zeitpunkt ihrer verspäteten Erstellung im Januar 2020 bereits erfüllt waren, hat die Bundesagentur das Projekt fortgeführt. Das Risikomanagement hat erkennbare Projektrisiken nicht oder nicht hoch genug priorisiert. Dies hat zu einem teilweisen Scheitern des Projektes geführt, da ein großer Teil der entwickelten Prüfregeln voraussichtlich nicht in den Wirkbetrieb überführt werden kann.

Die Bundesagentur sollte bei ihren weiteren Projekten alle notwendigen Projektdokumente zeitnah zur Projektgenehmigung erstellen. Die dabei festzulegenden Projektbeendigungskriterien muss sie beachten. Beim Projekt ERP-2AP sollte sie zunächst die Empfehlungen des Bundesrechnungshofes zu den Prüfregeln umsetzen. Anschließend könnte sie die Prüfregeln testweise befristet einführen und auf dieser Grundlage über einen möglichen dauerhaften Einsatz entscheiden.

Die Bundesagentur hat eingeräumt, dass sie die Projektbeendigungskriterien verspätet erstellt hat. Im Projekt sei aber zu keinem Zeitpunkt ein Zustand eingetreten, den sie durch früher vorliegende

Projektbeendigungskriterien hätte vermeiden können. Hinsichtlich der Projektrisiken hat die Bundesagentur erwidert, ihr Risikomanagement habe – entgegen der Feststellungen des Bundesrechnungshofes – die Risiken im Projekt adäquat bewertet. Zu den weiteren Würdigungen und Empfehlungen des Bundesrechnungshofes hat die Bundesagentur sich nicht geäußert.

Der Bundesrechnungshof teilt die Auffassung der Bundesagentur zum Risikomanagement nicht. Er hält weiterhin an seinen Würdigungen und Empfehlungen fest. (Tz. 5)

0.5 Die Bundesagentur hat es bisher versäumt, die Entwicklung analytischer Verfahren zu koordinieren. Insbesondere hat sie verschiedene Verfahren mit dem Schwerpunkt „Kassensicherheit" entwickelt, die sie nicht aufeinander abgestimmt hat. Dies hat auch zu fehlender Akzeptanz bei den Beschäftigten als Anwender geführt.

Die Bundesagentur hätte die Wirtschaftlichkeit der entwickelten Methoden nicht losgelöst vom tatsächlichen Einsatz in den Agenturen und gemeinsamen Einrichtungen betrachten dürfen. Ein nicht oder nur selten genutztes Verfahren kann seine Wirkung nicht entfalten. Damit haben auch die eingesetzten Haushaltsmittel ihr Ziel verfehlt.

Die Bundesagentur muss die Entwicklung ihrer Verfahren stärker koordinieren. Sie muss die Wirtschaftlichkeit der Verfahren unter den konkreten Einsatzbedingungen bewerten und auch nach der Einführung mittels einer Erfolgskontrolle überprüfen.

Die Bundesagentur hat erwidert, sie habe die Entwicklungen zum Einsatz analytischer Verfahren koordiniert. Dazu habe sie u. a. ein Anforderungsmanagement und bereichsübergreifende Statusrunden eingerichtet. Zu unseren Feststellungen zur Wirtschaftlichkeit hat sie sich nicht geäußert.

Der Bundesrechnungshof stellte bei der Prüfung fest, dass bei ERP-2AP Mängel aufgetreten sind, die bei geeigneter Koordination zu vermeiden gewesen wären. Er bleibt daher bei seinen Würdigungen und Empfehlungen zu diesem Punkt. Er erwartet darüber hinaus, dass die

Bundesagentur nun eine abschließende Erfolgskontrolle für ERP-2AP durchführt. (Tz. 6)

0.6 Die Bundesagentur erprobte den Einsatz von KI und analytischen Verfahren, um perspektivisch die Kassensicherheit „auf andere Art und Weise" als im Vier-Augen-Prinzip gewährleisten zu können. Dabei hat der Bundesrechnungshof fachliche, technische und organisatorische Mängel festgestellt (vgl. auch Teilprüfungsmitteilung 1).

Grundsätzlich hält er den Einsatz von KI und analytischer Verfahren dennoch für geeignet, um Prozesse zu automatisieren und damit die Verwaltung zu entlasten. Die Bundesagentur sollte ihre Vorgehensweisen und Methoden in diesem innovativen und dynamischen Technologieumfeld fortlaufend evaluieren, bedarfsgerecht weiterentwickeln und kontinuierlich verbessern. Die Empfehlungen des Bundesrechnungshofes zeigen hierzu erste Ansätze auf. (Tz. 7)

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 u. a. das IT-Verfahren ALLEGRO ein. In ALLEGRO werden die wesentlichen Leistungsansprüche in der Grundsicherung für Arbeitsuchende festgestellt und angeordnet.[2]

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 Risiken[3] kann die Prüfung durch das zweite Augenpaar auf Stichproben beschränkt werden. Diese Voraussetzung bestätigte die Bundesagentur für die Zahlungen in ALLEGRO und ERP.

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 anhalten[4].

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.

Im Jahr 2018 erfuhren wir, dass die Bundesagentur unter dem Projektnamen „Abgesichertes Zwei-Augen-Prinzip in ERP" (ERP-2AP) für das Verfahren ERP eine technische Lösung entwickelte, die deterministische und analytische Verfahren nutzte. Die deterministischen Verfahren verwendeten Methoden, die anhand von eindeutig festgelegten Vorbedingungen Auffälligkeiten in Zahlungen erkannten.[6] Als analytische Verfahren bezeichnete die BA jene Verfahren, die Methoden des maschinellen Lernens verwendeten.[7] Die Methoden des maschinellen Lernens versuchten auf Basis statistischer Auswertung bisher getätigter Eingaben, Muster und Gesetzmäßigkeiten in Daten zu erkennen.

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 das Vorgehen und die Methoden der Bundesagentur daher 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.

---

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

6 Diese Methoden werden auch als deterministische Methoden bezeichnet.

7 Diese Methoden werden auch als analytische Methoden bezeichnet.

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

2 Informationen zum Projekt ERP-2AP

Die Bundesagentur führte im Jahr 2011 ERP als einheitliches Finanzanwendungsverfahren ein. An ERP sind fast alle IT-Verfahren, in denen Leistungen oder Rechnungen bearbeitet werden, angebunden. Die Bundesagentur unterscheidet in ERP zwischen zwei grundsätzlichen Zahlungsabläufen:

- **Feststellung und Anordnung in vorgeschalteten IT-Verfahren**

  Ein wesentlicher Teil der Leistungen wird in IT-Verfahren bearbeitet und zur Zahlung angeordnet, die ERP vorgeschaltet sind. Eine weitere manuelle Bearbeitung in ERP ist bei diesen Fällen nicht mehr erforderlich. Die Zahlungen werden ab dem Zeitpunkt der Anordnung im vorgeschalteten IT-Verfahren vollautomatisiert verarbeitet und über ERP ausgezahlt. Beispielsweise wird das Arbeitslosengeld I nach dem SGB III im IT-Verfahren COLIBRI bearbeitet und angeordnet. Die erforderlichen Zahlungsdaten (u. a. Betrag, Verwendungszweck, Zahldatum) werden an ERP übergeben und für die Zahlung verwendet (Fallgruppe 1).

- **Feststellung und Anordnung in ERP**

  Zudem gibt es Leistungen, die in einem vorgeschalteten IT-Verfahren teilweise bearbeitet werden. In diesen Fällen werden die Zahlungsdaten ebenfalls an ERP übergeben und den Beschäftigten dort als Vorschlagswerte eingeblendet. Diese Fälle müssen in ERP noch manuell festgestellt und angeordnet werden. So werden insbesondere Leistungen an Anbieter für die berufliche Weiterbildung ausgezahlt. Weiter gibt es Leistungen, die unmittelbar in ERP erfasst und angeordnet werden. Hierzu zählen Rechnungen für Verwaltungsleistungen sowie Nach- und Sonderzahlungen für Leistungen nach dem SGB II und SGB III (Fallgruppe 2).

Da die Bundesagentur das in den BestMaVB-HKR vorgeschriebene Stichprobenverfahren nicht einhält, sieht sie für Zahlungen, die unmittelbar in ERP im „Zwei-Augen-Prinzip“ angeordnet werden (Fallgruppe 2), ein alternatives

Verfahren zur Kassensicherheit vor: In diesen Fällen werden 2,5 % der Zahlungen in der nzStpp überprüft. Die Bundesagentur erstellt die gezogene Stichprobe monatlich und übersendet sie zur Prüfung an die Agenturen für Arbeit (Agenturen) und die gE. Das Bundesministerium der Finanzen (BMF), das Bundesministerium für Arbeit und Soziales (BMAS) und der Bundesrechnungshof erhalten einen Quartalsbericht über die Ergebnisse der Prüfung. Hiernach überprüften die Agenturen und die gE im Jahr 2019 im „Zwei-Augen-Prinzip" 54 904 von insgesamt knapp 2,2 Mio. Zahlungen. In 903 Fällen (1,6 %) stellten sie einen Fehler fest. Demgegenüber stellte die Interne Revision der Bundesagentur fest, dass rund ein Drittel der von ihr geprüften Zahlungen im „Zwei-Augen-Prinzip" fehlerhaft waren.

Aus Sicht der Bundesagentur war das bisherige Verfahren aufwendig und hatte eine geringe Treffergenauigkeit. Sie wollte daher die nzStpp künftig durch automatisierte Überprüfungen ersetzen.

Hierfür entwickelte die Bundesagentur im Projekt ERP-2AP deterministische und analytische Verfahren, um potenziell fehlerhafte Anordnungen im „Zwei-Augen-Prinzip" im Verfahren ERP zu identifizieren. Deterministische Verfahren zielten hierbei auf eindeutig festgelegte Vorbedingungen ab, u. a. gesetzliche Vorgaben oder Weisungen. Die darauf aufbauenden Prüfregeln sollten somit die eindeutige Identifikation von Fehlern erlauben. Analytische Verfahren basierten auf statistischen Auswertungen bisher getätigter Eingaben. Hierbei versuchte die Bundesagentur z. B. mithilfe sogenannter Clusteranalysen[8] Auffälligkeiten in den Anordnungen zu erkennen (vgl. Teilprüfungsmitteilung 1) und setzte diese anschließend in analytischen Prüfregeln um.

Die auffälligen Anordnungen stellte sie den Agenturen und den gE als Prüffälle bereit. Die Beschäftigten sollten diese manuell prüfen und der Bundesagentur dazu Rückmeldungen geben. Anhand der Rückmeldungen wollte die Bundesagentur die Treffergenauigkeit[9] ihrer Prüfregeln feststellen und, falls erforderlich, die Prüfregeln weiterentwickeln (Abbildung 1).

---

[8] 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.

[9] „True positives", Rate an Auffälligkeiten, die vom System erzeugt werden und sich tatsächlich als auffällig herausstellen.

Abbildung 1

Kreisprozess zur Überprüfung der auffälligen Anordnungen



Quelle: Eigene Darstellung des Bundesrechnungshofes auf Basis von Erhebungsdaten.

Das Projekt hatte eine Laufzeit von zwei Jahren und sechs Monaten.[10] Im Projektauftrag gab der Projektlenkungsausschuss (PLA) als Ziel des Projektes aus,

- mit den zu entwickelnden Prüfregeln eine höhere Treffergenauigkeit als bei den bisher genutzten Stichprobenverfahren zu erreichen und
- die Fehlerquote bei Anordnungen in ERP zu verringern.

Die Bundesagentur teilte das Projekt in Arbeitspakete auf. Die Arbeitspakete unterlagen dem agilen Entwicklungsprozess des Projektes und konnten sich daher während der Projektlaufzeit nach Art und Umfang ändern. Tabelle 1 enthält die Zeitplanung des Projektes nach Arbeitspaketen. Die Inhalte der Arbeitspakete beschreiben wir im Anhang zu dieser Teilprüfungsmitteilung (Tz. 8).

[10] Das Projekt begann am 1. Mai 2018 und endete am 31. Oktober 2020.

Tabelle 1

Arbeitspakete und Zeitplanung im Projekt ERP-2AP

Projektphase	**Beginn**	**Ende**	**Bemerkung**
Realisierungsphase	**01.07.2018**	**31.07.2020**	
Arbeitspaket 0 (AP0)	01.07.2018	30.09.2018	Doppelte Anordnungen
Arbeitspaket 1 (AP1)	01.09.2018	30.11.2018	Adressprüfung/Zahlweg
Arbeitspaket 2 (AP2)	15.10.2018	24.01.2019	Zahlungsempfänger
Arbeitspaket 3 (AP3)	15.10.2018	28.02.2019	Fälligkeit/Verwendungszweck
Arbeitspaket 4 (AP4)	24.01.2019	26.09.2019	Finanzposition/Betrag
Arbeitspaket 5 (AP5)	08.07.2019	24.10.2019	Steuerkennzeichen/Abschlagszahlungen/ Buchungsdatum
Arbeitspaket 6 (AP6)	19.08.2019	14.03.2020	Online-Validierung
Arbeitspaket 7 (AP7)	09.10.2019	23.01.2020	Referenzen/Korrekturen
Arbeitspaket 8 (AP8)	09.10.2019	04.02.2020	Machbarkeitsstudie statistische und selbstlernende Betragsprüfung
Arbeitspaket 9 (AP9)	07.01.2020	26.03.2020	Erweiterung Reporting, Ergänzungen Referenzen/Korrekturen/Oberfläche
Arbeitspaket 10 (AP10)	12.03.2020	31.07.2020	Umsetzung verbleibender Anforderungen

Quelle: Eigene Auswertung von Erhebungsdaten.

3 ERP-2AP – Auswahl und Eignung der Prüfregeln

3.1 Sachverhalt

Das Projekt ERP-2AP erstellte und erprobte die identifizierten analytischen und deterministischen Prüfregeln in einem agilen Vorgehen.[11] Mit diesem wollte es alle Beteiligten, z. B. die fachlichen Bedarfsträger, den PLA, den technischen und den fachlichen Teil des Projektes geeignet einbinden. Das Anforderungsmanagement für die Regeln war das „Kernstück" des agilen Vorgehens. Dazu nutzte das Projekt eine zentrale Anforderungsliste in MS Excel. Die Anforderungen übertrug es in Arbeitspakete.

Ein Controlling für sein agiles Vorgehen bei der Auswahl und Erstellung der Prüfregeln hatte das Projekt ERP-2AP nicht aufgesetzt. Es konnte daher auf Ebene der einzelnen Prüfregeln z. B. nicht beantworten,

[11] Der dazugehörige Prozess baute auf dem Cross-Industry Standard Process for Data Mining auf.

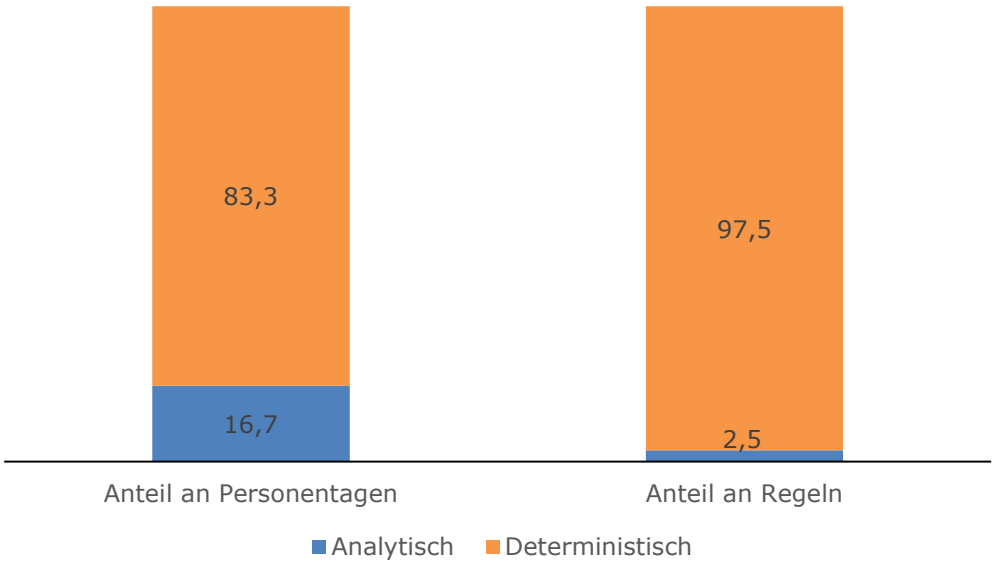
- wie hoch die Entwicklungsaufwände (in Personentagen und Euro) für eine Prüfregel waren und
- welchen Vermögensschaden eine Prüfregel im späteren Wirkbetrieb verhindern würde.

Das Projekt verfolgte die Aufwände lediglich auf Ebene der Arbeitspakete und nicht bezogen auf die darin entwickelten Prüfregeln. Es ermittelte dabei die Aufwände in Personentagen und nicht den Mittelabfluss in Euro. Die Bundesagentur teilte uns mit, dass die Kosten für eine Prüfregel aus den Aufwänden in Personentagen der dazugehörigen Arbeitspakete abgeschätzt werden könnten.

Insgesamt erforderten die Arbeitspakete 3 187 Personentage, davon 531 Personentage (16,7 %), um analytische Prüfregeln zu entwickeln und 2 656 Personentage (83,3 %), um deterministische Prüfregeln zu entwickeln (Abbildung 2).

Abbildung 2

Aufwände und daraus entstandene Prüfregeln in Prozent



Quelle: Eigene Auswertung von Erhebungsdaten.

Aus diesen Aufwänden entstanden insgesamt 40 für den Wirkbetrieb empfohlene Prüfregeln, darunter eine analytische Prüfregel (2,5 %, Abbildung 2).

Um die Wirksamkeit der entwickelten Prüfregeln zu untersuchen und eine Aussage über die spätere Anzahl an Auffälligkeiten[12] im Wirkbetrieb treffen zu können, sah das Projekt je nach Prüfregel

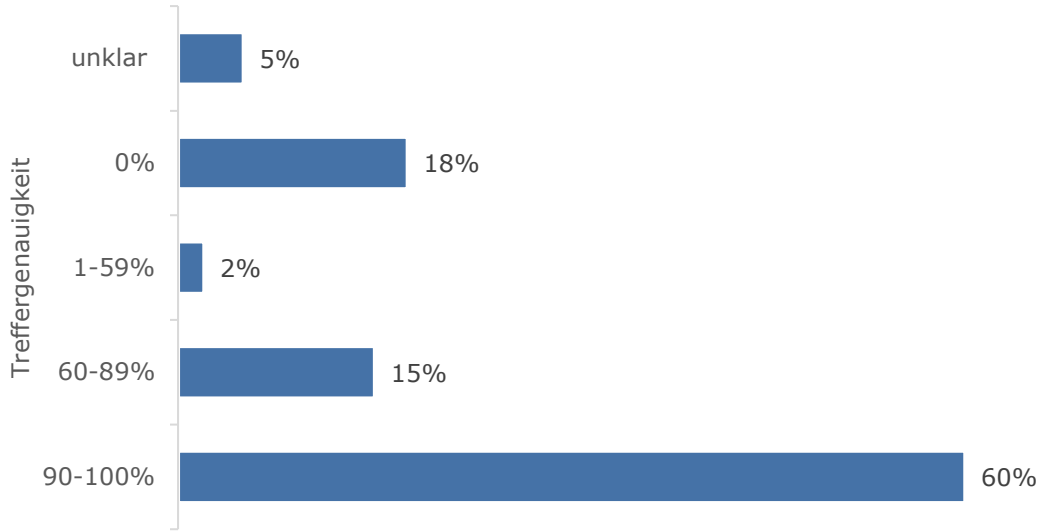
- eine Güteprüfung[13] von auffälligen Fällen durch eine Regionaldirektion (RD),
- Rückmeldungen fachlich zuständiger Organisationseinheiten oder
- interne Güteprüfungen durch Projektbeteiligte

vor. Es ermittelte hieraus die Treffergenauigkeit der Prüfregeln und die Anzahl der Auffälligkeiten für das Jahr 2019 und stellte seine Ergebnisse in einer „Auffälligkeitsmatrix 2019" zusammen.

Unsere Auswertung zeigte, dass 18 % der für den Wirkbetrieb vorgesehenen Prüfregeln eine Treffergenauigkeit von 0 % aufwiesen. Bei weiteren 5 % der Prüfregeln hatte die Bundesagentur die Treffergenauigkeit nicht ermittelt. Sie beabsichtigte dennoch, beide Gruppen von Prüfregeln im Wirkbetrieb einzusetzen (Abbildung 3).

Abbildung 3

Regeln und dazugehörige Treffergenauigkeit in Prozent



Quelle: Eigene Auswertung von Erhebungsdaten.

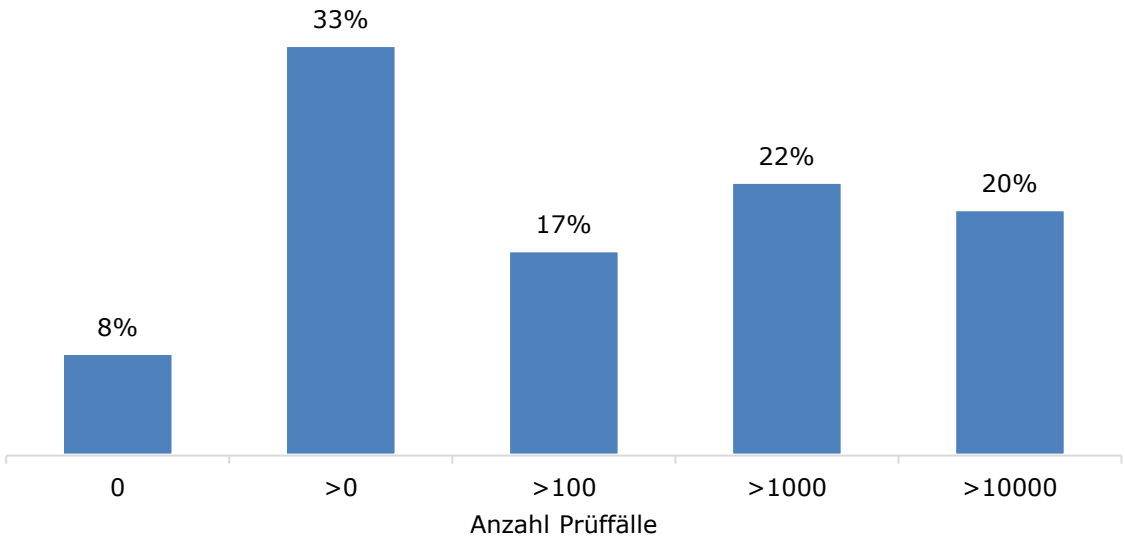
[12] Die Prüfregeln weisen auffällige Fälle aus. Diese werden dann als Prüffall an die Agenturen und gE weitergeleitet, siehe Abbildung 1.

[13] In der Regel testet die Bundesagentur Prüfregeln und Verfahren vor Überführung in den Wirkbetrieb in ausgewählten Agenturen und gE.

Insgesamt rechnete die Bundesagentur für das Jahr 2019 mit 310 804 Prüffällen, wenn die vom Projekt entwickelten Regeln bereits in den Wirkbetrieb überführt worden wären.[14] Im Vergleich dazu wies sie für das gleiche Jahr 54 904 Prüffälle mit der nachträglichen Stichprobe nzStpp aus. Wir stellten fest, dass 8 % der für den Wirkbetrieb vorgesehenen Prüfregeln im Jahr 2019 keine Prüffälle ausgewiesen hätten (Abbildung 4). Die Bundesagentur erläuterte dazu, dass alle Prüfregeln auf fachliche Notwendigkeit geprüft worden seien.

Abbildung 4

Prüfregeln und Anzahl an Prüffällen im Jahr 2019



Quelle: Eigene Auswertung von Erhebungsdaten.

20 % der Prüfregeln erzeugten pro Regel zwischen 10 000 und mehr als 90 000 Prüffälle. Insgesamt stammten 115 428 der 310 804 vom Projekt für das Jahr 2019 prognostizierten Prüffälle – mehr als 37 % – aus Prüfregeln mit 0 % oder bisher unbestimmter Treffergenauigkeit. Die Bundesagentur hatte dazu vermerkt, dass zu vermeiden sei, eine zu hohe Zahl an falsch-positiven Auffälligkeiten zu generieren, da diese letztendlich zu einer erhöhten Arbeitslast in den Dienststellen führen könne.

Für die Prüfregeln erstellte das Projekt ein Berichtswesen. Aus diesem sollten sowohl die Projektbeteiligten als auch die fachlichen Bedarfsträger im späteren

[14] Ohne Prüffälle, die bereits in der Online-Validierung bearbeitet werden.

Wirkbetrieb ablesen können, welche Treffergenauigkeit und welche Anzahl an Prüffällen die einzelnen Regeln erzielten, um ggf. nachsteuern zu können. Kriterien, bei deren Eintreten das System beispielsweise Warnmeldungen versendete (Schwellenwerte), Handlungsempfehlungen daraus für die fachlichen Bedarfsträger oder Hinweise, welche Treffergenauigkeiten fachlich und technisch als „gut" oder „schlecht" bewertet werden, enthielt das Berichtswesen zum Zeitpunkt der Prüfung nicht. Der Prozess einer möglichen Regelanpassung aufgrund niedriger Treffergenauigkeit oder zu niedriger/zu hoher Anzahl an Prüffällen im Wirkbetrieb war zum Zeitpunkt der Prüfung noch nicht geklärt.

3.2 Würdigung

Die Bundesagentur wollte im Projekt ERP-2AP die Kassensicherheit in ERP durch den Einsatz von analytischen Verfahren sicherstellen. Dennoch entwickelte sie im Projekt ERP-2AP überwiegend deterministische und trotz vergleichsweiser sehr hoher Aufwände lediglich eine analytische Prüfregel.

Sie hat versäumt, bereits während der Entwicklung ihrer Prüfregeln zu untersuchen, welche finanziellen Aufwände hierfür entstanden sind. Dies hätte sie gegen die erwarteten Einsparungen durch die Prüfregeln abwägen können, um zu einer datenbasierten Empfehlung für den Wirkbetrieb zu kommen. Stattdessen hat sie dies für einzelne Prüfregeln lediglich fachlich begründet, z. B. ob die Regel aufgrund interner Weisungen erforderlich war. So hat sie auch überflüssige Prüfregeln vorgesehen, die gar keine Prüffälle erzeugen.

Obwohl ihr bereits aus ALLEGRO-EFM das Risiko einer Überlastung der Beschäftigten mit zu vielen Prüffällen bekannt war, hat sie Prüfregeln vorgesehen, die auf ein Jahr gesehen insgesamt mehr als 310 000 Prüffälle erzeugen – über fünfmal so viele wie bei der nzStpp. Ein großer Anteil an Prüffällen hat aus Prüfregeln mit niedrigen oder unbekannten Treffergenauigkeiten gestammt. Falls die Bundesagentur diese Auswahl an Prüfregeln in den Wirkbetrieb überführt, sehen wir ein Risiko, dass dies die Beschäftigten überlasten wird. In der Folge würden sie die automatischen Überprüfungen ablehnen, was zum Scheitern des Verfahrens führen könnte (vgl. auch unsere Erkenntnisse zu ALLEGRO-EFM, Teilprüfungsmitteilung 1).

Das Berichtswesen sollte für fachliche Bedarfsträger und die Linienorganisation verständlich darstellen, ob sich die Prüfregeln im Alltagsgeschäft bewähren

oder nicht. Dies ist erschwert worden, da keine Schwellenwerte und dazugehörige Handlungsempfehlungen enthalten gewesen sind. Diese Informationen wären ggf. für fachliche Bedarfsträger entscheidend gewesen, insbesondere wenn sie nicht mit den technischen Hintergründen vertraut gewesen sind. So hätten diese z. B. leichter einschätzen können, ob einzelne Prüfregeln angepasst oder ggf. abgeschaltet werden sollten. Die Bundesagentur hat bisher auch versäumt, einen Prozess zur Anpassung der Prüfregeln im Wirkbetrieb aufzusetzen. Somit hat sie selbst bei einem vorhandenen Berichtswesen nicht sicherstellen können, dass sie Änderungsbedarfe an den Prüfregeln erkennt und diese anschließend auch umsetzt.

3.3 Empfehlung

Wir haben empfohlen, keine Prüfregeln in den Betrieb zu überführen, deren Treffergenauigkeit unbekannt oder gering ist. Bei Prüfregeln mit unbekannter Treffergenauigkeit sollte die Bundesagentur diese zunächst ermitteln. Sie sollte Prüfregeln, die eine hohe Anzahl an Fällen bei gleichzeitig niedriger Treffergenauigkeit aufweisen entweder optimieren oder nicht mehr für den Wirkbetrieb vorsehen. Prüfregeln mit einer Treffergenauigkeit von 0 % sollten entfallen. Insgesamt würde dies die Anzahl der Prüffälle deutlich reduzieren.

Für das dazugehörige Berichtswesen sollte die Bundesagentur einen Prozess aufsetzen, der es ermöglicht, die Prüfregeln im Wirkbetrieb kontinuierlich zu überwachen und ggf. zu verbessern. Wir haben empfohlen, das Berichtswesen um geeignete Schwellenwerte sowie Handlungsempfehlungen für die fachlichen Bedarfsträger und weiteren Adressaten innerhalb der Bundesagentur zu ergänzen.

3.4 Stellungnahme

In ihrer Stellungnahme hat die Bundesagentur dargestellt, dass sie die Sinnhaftigkeit einzelner Prüfregeln vor deren Einsatz abgewogen habe. Dabei habe sie sowohl den voraussichtlichen Aufwand als auch die Prüfungsergebnisse einbezogen. Durch einzelne Prüfregeln potenziell verhinderte Vermögensschäden ließen sich nicht vernünftig abschätzen. Aufgrund der nicht vorhersehbaren Anzahl an zu verhindernden Fehlern sei dies nicht möglich. Die Bundesagentur habe dies dennoch bei ihrer Wirtschaftlichkeitsbetrachtung zu ERP-2AP berücksichtigt. Hinsichtlich der Prüfregeln mit geringer Treffergenauigkeit hat sie

ausgeführt, dass diese teils bewusst vorbeugend und qualitätssichernd gedacht seien. So könne sie weisungskonformes Anordnen zusätzlich absichern. Demnach seien diese Regeln nicht überflüssig, auch wenn sie keine Prüffälle ausweisen.

Die Bundesagentur hat eingeräumt, dass sie im Projekt ERP-2AP auch Prüfregeln mit niedriger oder 0 % Treffergenauigkeit erstellte. Sie hat erläutert, warum diese dennoch erforderlich seien – z. B. um Dokumentationsmängel im Anordnungsprozess zu adressieren oder nur zulässige „Zahlwege" zuzulassen. Sie habe sich bewusst dafür entschieden, diese Prüfregeln einzusetzen, da sie die Qualität im Anordnungsprozess und die Kassensicherheit erhöhten.

Die Bundesagentur hat zugesagt, die Prüfregeln und deren Ergebnisse so auszuwählen, dass die Anzahl der Prüffälle die Beschäftigten der Agenturen und gE nicht überfordern werde.

Die Bundesagentur hat unsere Feststellungen zum Berichtswesen bestätigt. Sie werde unsere Empfehlungen dazu umsetzen, bevor sie die Prüfregeln in den Wirkbetrieb überführt. Sie werde insbesondere den Einsatz analytischer Prüfregeln mit einem engen Monitoring begleiten.

3.5 Abschließende Würdigung

In ihrer Stellungnahme erläutert die Bundesagentur, warum sie vorab nicht habe abschätzen können, welche potenziellen Vermögensschäden eine Prüfregel im Wirkbetrieb verhindert. Auf unsere Feststellung, sie habe versäumt, bereits während der Entwicklung ihrer Prüfregeln zu untersuchen, welcher finanzieller Aufwand hierfür entstanden ist, ist sie nicht eingegangen.

Die Bundesagentur hat erwidert, dass sie die Prüfregeln mit sehr geringer oder 0 % Treffergenauigkeit benötige, um z. B. „Lücken" in ERP schließen und damit unzulässige „Zahlwege" auszuschließen. Dieser Darstellung folgen wir nicht. Mängel in ERP sollte die Bundesagentur nicht mit Prüfregeln angehen, die zu vermeidbaren Mehraufwänden bei Beschäftigten führen. Stattdessen sollte sie anstreben, die Mängel unmittelbar in ERP zu beheben. Wir bleiben daher in diesem Punkt bei unseren Würdigungen und Empfehlungen.

Die Bundesagentur beabsichtigt, unsere Empfehlungen hinsichtlich der Anzahl an Prüffällen umzusetzen.

Auch unseren Empfehlungen zum Berichtswesen will die Bundesagentur folgen.

4 ERP-2AP – Auswahl und Eignung der analytischen Verfahren

4.1 Sachverhalt

In den Arbeitspaketen 4 und 8 entwickelte das Projekt ERP-2AP u. a. Methoden zur Betragsprüfung, um mithilfe von maschinellem Lernen Anordnungen mit „auffälligen Beträgen" zu finden.[15] Für diese analytische Betragsprüfung hatte das Projekt ERP-2AP auch ein Projektbeendigungskriterium definiert:

Sollte sich Anfang 2020 herausstellen, dass Analytics als optionale Projektkomponente keine belastbaren Ergebnisse liefert, wird Analytics nicht weiterverfolgt.

Zunächst entwickelte das Projekt eine Methode zum „Clustering"[16], um auffällige Beträge zu erkennen.[17] Es erhielt mit dieser Prüfmethode bei internen Tests eine hohe Anzahl von Prüffällen. Daher bat es den fachlichen Bedarfsträger, die zu prüfende Zahl an Fällen festzulegen. Mit einem solchen Mengengerüst wollte das Projekt sicherstellen, dass die Agenturen und die gE nicht mit zu vielen Prüffällen belastet werden. Dabei schlug es vor, jährlich die 30 000 riskantesten Fälle mit dem Fehlertyp „auffälliger Betrag" auszuweisen. Der fachliche Bedarfsträger stimmte dem zu.

Das Projekt ERP-2AP beabsichtigte, die analytischen Verfahren zu erproben und zu verfeinern. Im Oktober 2019 wies die Bundesagentur 303 Anordnungen mit dem Fehlertyp „auffälliger Betrag" zur Prüfung aus. Die RD Baden-Württemberg überprüfte davon 237 Anordnungen anhand der zahlungsbegründenden Unterlagen. Sie bestätigte bei einer der überprüften Anordnungen einen Treffer mit dem Fehlertyp „auffälliger Betrag" als zutreffend. Dies entsprach einer Treffergenauigkeit von 0,4 % (1 von 237). Bei Anordnungen in ERP, die in die zum Zeitpunkt der Prüfung eingesetzte nzStpp fielen,

---

[15] Darüber hinaus erprobte es weitere Methoden mit variablem Erfolg, z. B. um Freifelder wie den Überweisungszweck auf Merkmale wie das Vorhandensein eines Vor- und Nachnamens hin zu untersuchen.

[16] Ähnlich der bei ALLEGRO eingesetzten Methoden, Teilprüfungsmitteilung 1.

[17] Die Methode sollte auffällige Beträge anhand des „Abstands" zu den üblichen Beträgen auf dem jeweiligen Buchungstitel erkennen. Diesen Abstand nutzte das Projekt ERP-2AP auch, um die erzeugten Auffälligkeiten zu bewerten. Es stufte Buchungen als riskanter ein, je höher der Abstand zu den erkannten Clustern war.

ermittelten die Beschäftigten der Agenturen und gE im 1. und 2. Quartal 2019, dass durchschnittlich 0,29 % der Anordnungen einen fehlerhaften Betrag aufwiesen. Bei zwei der 237 Prüffälle entdeckte die RD Baden-Württemberg einen anderen Fehlertyp[18]. Ein finanzieller Schaden entstand bei diesen zwei Fällen nicht.

Aufgrund der Rückmeldungen aus der Güteprüfung sah die Bundesagentur ihr Projektbeendigungskriterium als erfüllt an. Sie verfolgte das „Clustering" nicht mehr weiter. Stattdessen erprobte sie aber drei andere Methoden des überwachten maschinellen Lernens[19],

- „Random Forest Trees",
- Stützvektormaschinen und
- logistische Regression.

Zu dieser Entscheidung hatte sich der PLA nicht geäußert. Die vom Softwarehersteller vorgegebene Standardimplementierung der „Random Forest Trees" bewertete das Projekt als nicht geeignet. Die beiden anderen Methoden kombinierte es zur neuen Aufdeckungsstrategie „Betrag"[20].

Methoden des überwachten maschinellen Lernens benötigen klassifizierte Trainingsdaten, um anhand dieser Schlussfolgerungen ableiten zu können. Diese Trainingsdaten wollte das Projekt aus den Rückmeldungen der Beschäftigten gewinnen. Zu diesen Rückmeldungen hatte es wie bei ALLEGRO-EFM (Teilprüfungsmitteilung 1) festgestellt, dass die Qualität der Rückmeldungen deutlich erhöht werden müsse, um die Performance und Robustheit der damit trainierten Modelle auf ein akzeptables Niveau zu heben. Das Projekt stellte fest, dass die Beschäftigten u. a. wiederholt eigentlich fehlerhafte Anordnungen als richtig markierten. Auch eine ähnliche Befüllung der Rückmeldefelder wie bei ALLEGRO-EFM konnte das Projekt nicht ausschließen. Dennoch plante es, die

---

[18] Bei einer Anordnung fehlte das Ordnungsmerkmal auf den zahlungsbegründenden Unterlagen, bei einer zweiten Anordnung wurde ein falscher Vertragsgegenstand ausgewählt.

[19] 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.

[20] Prüfregel „BE_20_PSCD", diese Prüfregel weist auffällige Beträge nur dann aus, wenn beide Komponenten des überwachten maschinellen Lernens „anschlagen".

Rückmeldungen automatisiert aus der SAP-Oberfläche an das Modell des überwachten maschinellen Lernens zu übergeben.

Die Bundesagentur führte im Jahr 2020 eine Güteprüfung ihrer neuen Aufdeckungsstrategie durch. Dazu wies sie 251 weitere Anordnungen mit dem Fehlertyp „auffälliger Betrag" zur Prüfung aus. Bis Februar 2020 prüfte die RD Baden-Württemberg davon 142 Anordnungen. Bei allen Anordnungen stellte sie fest, dass der Betrag korrekt war.

Im Juli 2020 war die Betragsprüfung mit Methoden des überwachten maschinellen Lernens noch nicht trainiert. Bei der Güteprüfung hatte die RD Baden-Württemberg eine Treffergenauigkeit der Prüfregel von 0 % zurückgemeldet. Unbeeinflusst davon meldete das Projekt ERP-2AP die technische Fertigstellung für diese Methoden und empfahl eine fachliche Freigabe der Prüfregel (Tz. 3.1).

Die Bundesagentur stellte dar, die Prüfregeln seien so ausgestaltet, dass den Beschäftigten bei einem Prüffall der vermutete Fehler zusammen mit einem nachvollziehbaren Hinweis ausgewiesen werde. Der dazu beispielsweise bei der Betragsprüfung vorgesehene Hinweis lautete wie folgt:

Die Kombination des Betrages in Verbindung mit anderen Angaben der Anordnung (z. B. Belegart, Finanzposition) war in der Vergangenheit ein Indiz für einen fehlerhaften Betrag. Bitte überprüfen Sie den Betrag der Anordnung, ebenso wie alle weiteren Angaben.

Bei Annahme des mit dem PLA vereinbarten Mengengerüsts wären somit 30 000 der ausgewiesenen Prüffälle nicht für die Beschäftigten nachvollziehbar.

Die Wirtschaftlichkeitsbetrachtung (WiBe) des Projektes enthielt eine Abschätzung der Anzahl an Prüffällen in den Jahren 2020 bis 2025. Die Bundesagentur ging ab dem Jahr 2021 von einer sinkenden Anzahl aus, da sie aufgrund der Dokumentation der Prüfregeln und der Hinweise zu den Prüffällen Lerneffekte bei den Beschäftigten erwartete. Wie diese Lerneffekte für künftige Anordnungen bei einer für die Beschäftigten nicht nachvollziehbaren, sich stetig ändernden Prüfregel zur Betragsprüfung eintreten könnten, konnte die Bundesagentur uns nicht erläutern.

Im November 2019 stellte die Bundesagentur fest, dass der Wirkbetrieb der entwickelten analytischen Prüfregel noch nicht hinreichend geklärt war. U. a. hatte sie noch nicht abschließend bewertet, ob sie für den Wirkbetrieb eigene Datenanalysten/Datenwissenschaftler[21] gewinnen oder das nötige Fachwissen weiterhin durch Einkauf von Externen sicherstellen möchte. Während der Projektlaufzeit hatte sie dieses Fachwissen durch Vergabe an Externe beschafft.

4.2 Würdigung

Ein grundsätzliches Ziel der Automatisierung ist es, Beschäftigte von Routineaufgaben zu entlasten. Weist die Bundesagentur durch ungeeignete oder nicht getestete Prüfregeln mindestens die gleiche Anzahl an Prüffällen wie bei der nzStpp aus, so verringert sich der Prüfaufwand der Beschäftigten nicht. Zudem hätte das Verfahren des maschinellen Lernens bei Vorgabe eines festen Mengengerüstes jeden Monat die „gewünschte" Anzahl an Fällen gefunden. Damit hätte das Risiko bestanden, dass eigentlich unauffällige Prüffälle in die Fläche gegeben worden wären.

Obwohl das Projektbeendigungskriterium zur Entwicklung der analytischen Betragsprüfung erfüllt war, hat die Bundesagentur dieses Projekt weiterverfolgt. Sie hat versäumt, diese Entscheidung in den Projektunterlagen nachvollziehbar zu dokumentieren.

Bereits bei ALLEGRO-EFM hat die fehlende Evaluation der Modelle im Wirkbetrieb dazu geführt, dass anfänglich zu viele Prüffälle der Kategorie „Falscher Treffer" an die gE übergeben wurden. Die Bundesagentur hat vermutet, dass dies die Akzeptanz für ALLEGRO-EFM nachhaltig negativ beeinflusste. Sie hat ihre neu entwickelte analytische Prüfregel für den Einsatz freigegeben, obwohl ihr deren Treffergenauigkeit von 0 % bekannt war. Da sich bereits die zuvor entwickelte analytische Prüfregel in der Güteprüfung als untauglich erwiesen hatte, erscheint uns ein solches Vorgehen als nicht nachvollziehbar. Darüber hinaus verschlechtert sich diese Prüfregel sogar, wenn das zugrundeliegende Modell wiederholt mit Daten in nachgewiesen schlechter Qualität aus den Rückmeldungen der Beschäftigten trainiert wird. Es besteht das Risiko, dass

[21] Datenanalysten/Datenwissenschaftler gewinnen aus großen Datenmengen Informationen und leiten daraus Handlungsempfehlungen ab. Sie überprüfen ihre Hypothesen mit statistischen Methoden und bereiten diese anschließend für die Leitungsebene als Entscheidungsgrundlage auf. Sie bedienen sich dazu moderner Methoden und Technologien, u. a. dem maschinellen Lernen und der KI.

die Prüfregel dadurch immer ungenauer wird, wodurch die Akzeptanz bei den Beschäftigten weiter sinkt und sich somit die Qualität ihrer Rückmeldungen weiter verschlechtert.

Die Beschäftigten haben bei der analytischen Betragsprüfung nur vage Hinweise erhalten, die auf einen konkreten Fehler hindeuten. Es ist unklar gewesen, wie hier signifikante Lerneffekte eintreten können. Auch dies hat nicht zur Akzeptanz des Verfahrens beigetragen. Daher ist es zweifelhaft, ob die sinkende Zahl von Prüffällen aufgrund der Lerneffekte eintreten wird, die die Bundesagentur in ihrer WiBe angegeben hat.

Methoden des überwachten maschinellen Lernens müssen anhand der Trainingsdaten angepasst und weiterentwickelt werden. Ist dies nicht sichergestellt, besteht ein hohes Risiko des Scheiterns, wie es bei ALLEGRO-EFM eingetreten ist (Teilprüfungsmitteilung 1).

Ohne Datenanalysten/Datenwissenschaftler wird die Bundesagentur das Verfahren ERP-2AP nicht in geeigneter Weise betreuen können. Diese wären z. B. notwendig gewesen, um die analytische Prüfregel im Wirkbetrieb fachlich und technisch zu evaluieren. Falls die Bundesagentur hierzu kein eigenes Personal einsetzt, sehen wir das Risiko, dass sie zunehmend von externen Dienstleistern abhängig wird. Dies gilt umso mehr, wenn sie künftig weitere Verfahren des maschinellen Lernens in den Wirkbetrieb überführen möchte.

4.3 Empfehlung

Die Bundesagentur sollte bei ERP-2AP berücksichtigen, dass sich die Anzahl der von einem Verfahren des maschinellen Lernens ausgewiesenen Prüffälle dynamisch ändern kann. Demnach haben wir empfohlen, keine festen Quoten vorzugeben, ohne die Prüfregeln vorab evaluiert zu haben.

Treten Projektbeendigungskriterien ein, so sollte die Bundesagentur ihr dazugehöriges (Teil-)Projekt beenden. Tut sie das nicht, so muss sie dies entsprechend begründen und ihre Entscheidung dokumentieren.

Wir haben der Bundesagentur empfohlen, analytische Verfahren immer mit einer Güteprüfung zu evaluieren und die Erkenntnisse daraus in geeigneter Weise zu berücksichtigen, bevor sie diese für den Wirkbetrieb bei den gE freigibt. Bei Modellen des überwachten Lernens sollte die Bundesagentur die

Datenqualität berücksichtigen und solche Modelle bei schlechter Datenqualität nicht einsetzen.

Die Bundesagentur sollte ihre Hinweise für die Beschäftigten anpassen und versuchen, ihre Methoden des maschinellen Lernens „erklärbar" darzustellen. Wir haben empfohlen, die WiBe hinsichtlich der Lerneffekte und der Nachvollziehbarkeit des Verfahrens anzupassen.

Die Bundesagentur sollte prüfen, inwieweit sie für Projekte des maschinellen Lernens und der KI Datenanalysten/Datenwissenschaftler benötigt oder eigenes Personal qualifizieren kann, um die erforderlichen Kompetenzen sicherzustellen.

4.4 Stellungnahme

Die Bundesagentur hat zugesagt, die Prüfregeln so zu optimieren, dass die Anzahl der Prüffälle die Beschäftigten der Agenturen und gE künftig nicht überfordert (Tz. 3.4).

Die Bundesagentur hat eingeräumt, dass sie die Prüfregel „auffälliger Betrag" trotz des erfüllten Projektbeendigungskriteriums weiterverfolgt habe. Sie habe die Prüfregel trotz der in der Güteprüfung ermittelten Treffergenauigkeit von 0 % für den Wirkbetrieb empfohlen, um wenigstens zu einer analytischen Prüfregel weitere Erkenntnisse zu gewinnen.

Die Bundesagentur hat darüber hinaus mitgeteilt, sie informiere ihre Beschäftigten über die Wirkungsweise ihrer Methoden des maschinellen Lernens und stelle diese auch „erklärbar" dar. Zu unserer Empfehlung, die WiBe hinsichtlich der Lerneffekte und der Nachvollziehbarkeit des Verfahrens anzupassen, hat sie sich nicht geäußert.

Die Bundesagentur hat erläutert, dass sie zur Entwicklung analytischer Verfahren vermehrt eigene Beschäftigte einsetze. Sie habe mittlerweile Stellen mit Datenanalysten/Datenwissenschaftlern besetzen können und wolle unserer Empfehlung zu diesem Punkt nachkommen.

4.5 Abschließende Würdigung

Die Bundesagentur beabsichtigt, unsere Empfehlung zur Anzahl an Prüffällen umzusetzen.

An unseren Würdigungen und Empfehlungen

- zum (erfüllten) Projektbeendigungskriterium und
- der trotz 0 % Treffergenauigkeit für den Wirkbetrieb empfohlenen analytischen Prüfregel

halten wir fest. Die Bundesagentur hätte diese Prüfregel nicht ohne weitere Tests für den Wirkbetrieb empfehlen sollen. Stattdessen hätte sie die Prüfregeln z. B. überarbeiten und dann erneut zur Güteprüfung ausweisen können.

Wir erkennen an, dass die Bundesagentur ihren Beschäftigten die Wirkungsweise von Methoden des maschinellen Lernens näherbringen möchte. Sie sollte auch ihre Hinweise zur analytischen Prüfregel „auffälliger Betrag" überarbeiten.

Unserer Empfehlung, eigene Beschäftigte als Datenanalysten/Datenwissenschaftler einzusetzen, will die Bundesagentur folgen.

5 ERP-2AP – Projektmanagement, Einführung und Wirtschaftlichkeit

5.1 Sachverhalt

Wir prüften im Jahr 2017 das Projekt „Rollenbasierte Oberflächen in der Bundesagentur".[22] Dabei stellten wir u. a. fest, dass die Bundesagentur ihre Ziele nicht konkret beschrieb sowie Meilensteine und Ausstiegszenarien, trotz deutlicher Zielverfehlung, nicht vorsah. Sie sagte zu, ihr Projektmanagement anzupassen. Insbesondere wollte sie zu Beginn eines Projektes Projektbeendigungskriterien[23] festlegen.

Die Bundesagentur genehmigte das Projekt ERP-2AP am 22. März 2018 mit der Auflage, Projektbeendigungskriterien zu erstellen. Das Projekt erstellte erst im Januar 2020[24] sieben Projektbeendigungskriterien. Am 4. Februar 2020, knapp sechs Monate vor Ende der Realisierungsphase (Tabelle 1), nahm der PLA diese Projektbeendigungskriterien ab. Zu diesem Zeitpunkt

---

[22] Bericht an den Haushaltsausschuss des Deutschen Bundestages nach § 88 BHO über die Prüfung des Projektes „Rollenbasierte Oberflächen (ROBASO)" und weiterer Projekte der Bundesagentur für Arbeit, Gz. VI 1 - 2017 - 1087 / VI 1 - 2016 - 0560 vom 6. Dezember 2018.

[23] Projektbeendigungskriterien sind dazu gedacht, dass ein Projekt bei Eintreten geänderter Bedingungen (z. B. Wegfallen des Grunds für das Projekt) qualifiziert beendet werden kann.

[24] Das Projekt begann am 1. Mai 2018 und endete am 31. Oktober 2020.

waren vier der sieben Projektbeendigungskriterien (ca. 60 %) bereits eingetreten oder nicht mehr relevant. U. a. war die durch das Projekt ursprünglich angestrebte unbefristete Ausnahmegenehmigung für das „Zwei-Augen-Prinzip" in ERP bereits ab dem 1. Juli 2018 erteilt worden. Der PLA hatte das Projekt dennoch fortgeführt und das Projekt nunmehr stärker auf eine gesteigerte Buchungsqualität in ERP ausgerichtet.

Das Risikomanagement des Projektes war dafür zuständig, Projektrisiken zu identifizieren und deren Status mitzuverfolgen. Es stellte bis einschließlich November 2019 insgesamt 21 Projektrisiken fest. Bei 16 Risiken sah es keinen Handlungsbedarf mehr und überführte diese in den Status „geschlossen". Vier Risiken waren weiterhin relevant und damit im Status „in Bearbeitung". Das Risiko *„fehlende Akzeptanz des Systems durch Anwender, Widerstand bei der Einführung*" hatte das Projekt als „wesentlich" bewertet, es hatte den Status „geschlossen". Zu diesem Zeitpunkt war die geringe Nutzungsquote „Zwei-Augen-Prinzip" und niedrige Akzeptanz der Prüfroutinen von ALLEGRO-EFM (Teilprüfungsmitteilung 1) bereits bekannt. Das Projekt ERP-2AP hatte sich regelmäßig auf technischer und fachlicher Ebene mit dem Bereich EFM ausgetauscht. Es hatte das Risiko aufgrund der Rückmeldungen aus seiner Güteprüfung geschlossen, obwohl die Beschäftigten dort nur ausgewählte Prüfregeln erprobten. Zwei weitere, im November 2019 noch offene, miteinander verbundene Risiken waren

- die *„fehlende Entlastung der Fläche*"[25] und
- die fehlende Ausnahmegenehmigung für ERP-2AP in der vom Projekt vorgesehenen Form[26].

Den Einfluss dieser Risiken bewertete das Projekt intern als „moderat" oder „niedrig". Somit war es nicht vorgesehen, diese Risiken dem PLA zu berichten.[27] In der zehnten Sitzung des PLA wurde das Risiko der fehlenden

---

[25] Fehlende Entlastung der Fläche: Trotz Einführung ist derzeit unklar, ob die Stichprobenprüfung künftig entfallen kann. Dies führt zu einer Mehrbelastung der Anwender, statt der geplanten Minderbelastung.

[26] Das Projekt zielte auf eine Ausnahmegenehmigung nach Nummer 6.6.2 VV-ZBR BHO ab, bei der die nzStpp aufgrund der entwickelten regelbasierten und analytischen Überprüfungen entfallen kann.

[27] Das Projekt berechnete intern aus dem Einfluss der Risiken sowie deren Eintrittswahrscheinlichkeit die Kennzahl *„Risikobeurteilung*". Ab einer Risikobeurteilung von „8" oder höher berichtete das Projekt über das dazugehörige Risiko im PLA. Für die beiden erwähnten Risiken lag die Risikobeurteilung bei „4".

Ausnahmegenehmigung dennoch besprochen. Der Ausschuss entschied, an der ursprünglich angedachten Zeitplanung festzuhalten und sah vor, die im Projekt ERP-2AP entwickelten Methoden ab Mitte Juni 2020 einzusetzen.

Die Bundesagentur plante, die im Projekt ERP-2AP entwickelten deterministischen und analytischen Prüfregeln in zwei Stufen einzuführen. Zunächst wollte sie Anteile der in den Arbeitspaketen 0 bis 3 entwickelten, deterministischen Prüfregeln einführen. Diese sollten als „Online-Validierungen" den Beschäftigten in ERP Hinweise geben, während diese ihre Vorgänge bearbeiten. Die Online-Validierungen führte sie im März 2020 ein.

Mitte Juni 2020 wollte die Bundesagentur die weiteren entwickelten Prüfregeln einführen. Dazu hatte das Projekt insgesamt 58 Prüfregeln fertig entwickelt und technisch produktiv gesetzt. Es meldete den fachlichen Bedarfsträgern, dass es seine Prüfregeln produktiv gesetzt habe. Es bewertete diese Prüfregeln fachlich und technisch und stellte dabei fest, dass 40 davon auch für den Wirkbetrieb geeignet wären (Tz. 3.1). Es empfahl daher, diese einzuführen.

Bis zum Abschluss des Projektes hatte die Bundesagentur keine der nachgelagerten Prüfregeln eingeführt. Aufgrund der zusätzlichen Belastung für die Beschäftigten könnten nicht zusätzlich noch die deterministischen und analytischen Prüfregeln des Projektes ERP-2AP eingeführt werden.

Die Bundesagentur legte uns drei Versionen der WiBe des Projektes ERP-2AP vor (Tabelle 2).

Tabelle 2

WiBe-Kennzahlen des Projektes ERP-2AP

Nummer	Datum	WiBe Q	WiBe E	WiBe KN	WiBe KN/R
1	27. April 2018	51	33	-10 755 880	-11 081 308
2	15. Januar 2020	61	73	-7 279 554	-7 609 554
3	28. April 2020	61	73	-7 563 839	-7 563 839

Quelle: Eigene Auswertung von Erhebungsdaten.

In ihrer ursprünglichen WiBe war die Bundesagentur noch von einer Kennzahl zu den qualitativ-strategischen Kriterien (WiBe Q) von 51 von 100 möglichen Punkten ausgegangen. Mit der zweiten Version der WiBe wertete die Bundesagentur diese Kennzahl um 10 Punkte auf. Bei der Kennzahl zu externen

Effekten (WiBe E) nahm sie ursprünglich einen Wert von 33 von ebenfalls 100 möglichen Punkten an. In der zweiten Wirtschaftlichkeitsbetrachtung erhöhte sie den Wert der WiBe E um mehr als das Doppelte auf nunmehr 73 Punkte. Im April 2020 überarbeitete die Bundesagentur die WiBe erneut. Sie hielt dabei an ihren Werten der WiBe Q und Wibe E fest und passte nur die Werte der Kennzahlen zu den Kosten an, da sie u. a. nicht mehr von einem Risikozuschlag von 300 000 Euro ausging.

Die Bundesagentur hatte die dritte Version ihrer WiBe zusammen mit dem Projekthandbuch des Projektes ERP-2AP auch dem BMAS zur Abstimmung vorgelegt. Im Juni 2020 bemängelte das BMAS u. a., dass

- der tatsächliche Mehrwert des Systems gegenüber dem Altsystem nicht ausreichend und nachvollziehbar dargestellt war und
- keine tragfähige Dokumentation aller in Betracht kommenden Alternativen vorhanden war. Die Bundesagentur betrachtete die Kosten (IT- und Personalkosten) sowie Nutzen nicht hinreichend für jede Alternative.

Bei allen in der WiBe und im Projekthandbuch getroffenen Annahmen, insbesondere bei den Kennzahlen „WiBe Q", „WiBe E" und dem monetären Nutzen ging die Bundesagentur davon aus, dass sie die im Projekt ERP-2AP entwickelten Prüfregeln in den Wirkbetrieb überführen kann. Zum Abschluss des Projektes war dies noch nicht der Fall.

5.2 Würdigung

Das Projekt ERP-2AP hat es versäumt, rechtzeitig geeignete Projektbeendigungskriterien zu definieren. Durch die fast zwei Jahre verzögerte Erstellung der Projektbeendigungskriterien haben dem PLA für die laufende Bewertung des Projektes entscheidende handlungsleitende Maßstäbe gefehlt. Obwohl die Mehrzahl der Kriterien zur Projektbeendigung im Zeitpunkt ihrer verspäteten Erstellung bereits erfüllt oder nicht mehr relevant waren, hat die Bundesagentur das Projekt fortgeführt. Es hat die Auflage, unter der die Bundesagentur das Projekt genehmigte, nicht eingehalten. Damit hat die Bundesagentur auch ihre Zusagen zur Verbesserung ihres Projektmanagements nicht eingehalten.

Das Risikomanagement des Projektes hat entscheidende Risiken nicht hoch genug priorisiert oder als geschlossen eingestuft, obwohl Hinweise auf ähnliche Risiken (und deren Eintritt) aus ALLEGRO-EFM bekannt waren. Die

Begründung, das Risiko *„fehlende Akzeptanz des Systems durch Anwender, Widerstand bei der Einführung*" aufgrund der Rückmeldungen durch die Fläche zu schließen, trägt nicht. Dies gilt insbesondere vor dem Hintergrund, dass die Güteprüfung nur auf die Beurteilung einer Auswahl der Prüfregeln durch die RD Baden-Württemberg abzielte. Das Projekt hatte keine ausreichenden Informationen über die Akzeptanz seiner entwickelten Methoden vorliegen.

Die Risiken *„fehlende Entlastung der Fläche*" und der fehlenden Ausnahmegenehmigung hat es zwar mitverfolgt und nicht geschlossen, aber zu niedrig bewertet. Diese Risiken würden eintreten, wenn die Prüfregeln vollumfänglich in den Wirkbetrieb überführt würden. Dies hat zum teilweisen Scheitern des Projektes geführt: ein großer Teil der entwickelten Prüfregeln kann trotz technischer Fertigstellung innerhalb des vorher gesetzten Zeitrahmens und Budgets voraussichtlich nicht in den Wirkbetrieb überführt werden.

Noch zu Beginn des Jahres 2020 hatte der PLA die angepasste WiBe des Projektes abgenommen, in der das Projekt u. a. den qualitativ-strategischen Nutzen sowie die externen Effekte deutlich höher bewertete. Warum der PLA diese Kriterien höher bewertete, obwohl die Schwächen des Systems bekannt waren, haben wir nicht nachvollziehen können. Da die Bundesagentur die nachgelagerten Prüfregeln bisher nicht eingesetzt hat, sind ihre Bewertungen des monetären Nutzens sowie der qualitativ-strategischen Kriterien und der externen Effekte hinfällig. Die Bundesagentur hat es versäumt, ihr Projekt und die dazugehörigen Dokumente so aufzustellen, dass der Nutzen (und die Einsatzmöglichkeit) von ERP-2AP nicht größtenteils von der Abschaffung der nzStpp abhängt. Das Risiko, dass das BMAS und das BMF den dazugehörigen Ausnahmeantrag ablehnen, hat sie konsequent als zu gering bewertet.

5.3 Empfehlung

Die Bundesagentur sollte bei ihren weiteren Projekten ihre Zusagen einhalten und alle nötigen Projektdokumente zeitnah zur Projektgenehmigung erstellen. Sie muss dabei ihre Projektbeendigungskriterien beachten (Tz. 4.1).

Bei weiteren Projekten zum maschinellen Lernen sollte die Bundesagentur mehr Wert auf ihr Veränderungsmanagement legen und die Agenturen und gE noch früher einbinden. So könnte sie Risiken, die im Projekt ERP-2AP eingetreten sind, frühzeitig erkennen und meiden. Die Bundesagentur sollte die

tägliche Belastung der Beschäftigten in den Agenturen und gE mit Routineaufgaben erheben und angesichts dessen ihre weiteren Entwicklungen zur Kassensicherheit steuern – insbesondere hinsichtlich der versandten Anzahl an Prüffällen. Ist ein Verfahren von der Zustimmung nicht im Projekt Beteiligter „Stakeholder" abhängig, so sollte die Bundesagentur diese frühzeitiger informieren und einbinden.

Wir haben die Bundesagentur aufgefordert, die WiBe und das Projekthandbuch von ERP-2AP dem Sachstand anzupassen und mit dem BMAS abzustimmen. Sie sollte zunächst unsere Empfehlungen der voranstehenden Textziffern umsetzen (Tzn. 3.3 und 4.3). Anschließend könnte sie ihre Prüfregeln testweise befristet einführen. Abhängig von den Ergebnissen sollte sie einen Ausnahmeantrag beim BMAS und BMF einreichen, um die nzStpp abzulösen.

5.4 Stellungnahme

Die Bundesagentur hat eingeräumt, dass sie die Projektbeendigungskriterien im Projekt ERP-2AP verspätet erstellt hat. Sie habe bereits vor der formalen Definition der Projektbeendigungskriterien geeignete Zeitpunkte zum möglichen vorzeitigen Projektabbruch ermittelt und diese auch nachgehalten. Zu keinem Zeitpunkt sei ein Zustand eingetreten, den sie durch die Projektbeendigungskriterien hätte vermeiden können. Außerdem müsse ein erfülltes Projektbeendigungskriterium nicht zwingend dazu führen, ein Projekt abzubrechen.

Darüber hinaus hat die Bundesagentur dargestellt, dass sie aus den Mängeln von ALLEGRO-EFM gelernt habe. Sie habe einige dieser Mängel bei ERP-2AP abstellen können. Z. B. sei das Prüfsystem von ERP-2AP wesentlich komfortabler als bei ALLEGRO-EFM. Darüber hinaus seien den Beschäftigten in den Agenturen und gE die nachgelagerten Prüffälle zeitnah bereitgestellt worden.

Die Bundesagentur hat mitgeteilt, dass ihr Risikomanagement entgegen unserer Feststellungen gut funktioniert habe und die Risiken im Projekt adäquat bewertet worden seien. So sei z. B. das Risiko *„fehlende Entlastung der Fläche"* nicht eingetreten, da die Bundesagentur die entwickelten Prüfregeln noch nicht vollständig in den Wirkbetrieb überführt habe. Die Rückmeldungen der Beschäftigten aus den Güteprüfungen seien uneingeschränkt positiv gewesen. Deshalb habe die Bundesagentur auch das Risiko *„fehlende Akzeptanz des*

Systems durch Anwender, Widerstand bei der Einführung" nicht mehr weiterverfolgt.

Zu unseren Feststellungen zur WiBe des Projektes hat sich die Bundesagentur nicht geäußert. Das Risiko der fehlenden Ausnahmegenehmigung sei nur für die nach der Auszahlung stattfindenden Prüfregeln relevant. Es sei demnach kein Projektrisiko im eigentlichen Sinne gewesen. Die Online-Validierungen habe die Bundesagentur auch ohne den Ausnahmeantrag in den Wirkbetrieb überführen können.

5.5 Abschließende Würdigung

Wir stimmen der Bundesagentur zu, dass ein erfülltes Projektbeendigungskriterium nicht zwingend erfordert, ein Projekt zu beenden. Dennoch halten wir es für angezeigt, dass die Bundesagentur ihre Entscheidungen im Projektmanagement nachvollziehbar begründet und diese auch dokumentiert. Dies gilt insbesondere, wenn sie Projekte weiterführt, obwohl deren Projektbeendigungskriterien erfüllt sind. Wir halten in diesem Punkt an unseren Empfehlungen fest (Tzn. 4.3 und 4.5).

Wir begrüßen es, wenn die Bundesagentur aus Fehlern vorangegangener Projekte wie ALLEGRO-EFM lernt. Bei zukünftigen Projekten sollte sie daneben auch unsere Feststellungen, Würdigungen und Empfehlungen zum Projekt ERP-2AP berücksichtigen.

Die Erläuterungen der Bundesagentur zum Risikomanagement können wir nicht nachvollziehen. Die Aussage, dass z. B. das Risiko *„fehlende Entlastung der Fläche*" nur noch nicht eingetreten sei, da die Prüfregeln noch nicht vollständig im Wirkbetrieb seien, entkräftet unsere Würdigungen und Empfehlungen nicht. Sie bestätigt damit, dass dieses Risiko eingetreten wäre, wenn sie ihre Prüfregeln vollständig in den Wirkbetrieb überführt hätte. Dieses Risiko ist eng mit dem Risiko *„fehlende Akzeptanz des Systems durch Anwender, Widerstand bei der Einführung*" verknüpft. Wir halten daher weiterhin an unseren Würdigungen und Empfehlungen zu diesem Punkt fest.

Die Bundesagentur äußerte sich nicht zu unseren Würdigungen und Empfehlungen hinsichtlich der WiBe des Projektes ERP-2AP. Wir erwarten, dass die Bundesagentur unsere Empfehlungen zu diesem Punkt umsetzt. Ungeachtet dessen, dass die Bundesagentur die fehlende Ausnahmegenehmigung des

BMAS und BMF nicht für ein Projektrisiko im eigentlichen Sinne hält, ist es dennoch ein Risiko für den Wirkbetrieb, insbesondere in Hinblick auf die nachgelagerten Prüfregeln. Wir bleiben bei unserer Empfehlung, dass die Bundesagentur diese Prüfregeln zunächst befristet einführen sollte. Abhängig von den Ergebnissen sollte sie einen neuen Ausnahmeantrag beim BMAS und BMF einreichen, um die nzStpp abzulösen.

6 Strategie und Koordination der Entwicklungen zur Kassensicherheit

6.1 Sachverhalt

Die Bundesagentur wollte in ihrem gesamten Geschäftsbereich vermehrt analytische Verfahren zu unterschiedlichen Zwecken einsetzen. Hierzu gehörten neben dem Einsatz in ALLEGRO und ERP auch die Automatisierung in der Arbeitslosenversicherung (AIDAV) sowie die Echtheitsprüfung von Schul- und Studienbescheinigungen beim Kindergeldantrag. Eine übergreifende Koordinierung dieser Vorhaben sah sie bisher nicht vor.

Insbesondere entwickelte sie verschiedene Verfahren, um die Kassensicherheit im Vier-Augen-Prinzip perspektivisch „auf andere Art und Weise" gewährleisten zu können. Mehrere dieser Verfahren setzten dabei auf Prüfregeln – darunter ALLEGRO-EFM und ERP-2AP. Zu Beginn des Projektes ERP-2AP hatte die Bundesagentur bereits Erfahrungen mit ALLEGRO-EFM gemacht. Vor allem wusste sie, dass die geringe Treffergenauigkeit zu einer geringen Nutzung und Rückmeldung durch die gE führte. Trotz dieser Erkenntnis wollte sie auch mit ERP-2AP wieder Prüfregeln in den Wirkbetrieb überführen, die in der Testphase nur eine sehr geringe oder gar keine Treffergenauigkeit aufwiesen.

Der Bereich „Controlling und Finanzen" in der Zentrale der Bundesagentur ist u. a. für die Kassensicherheit der Verfahren der Bundesagentur zuständig, mit denen Haushaltsmittel ausgezahlt werden können. Dem Beauftragten für den Haushalt (BfdH) der Bundesagentur obliegt dabei u. a. die Einhaltung der BestMaVB-HKR. Er trifft im Einvernehmen mit der Bereichsleitung CF 2 strategische Entscheidungen und koordiniert die Entwicklungen zur Kassensicherheit in der Bundesagentur. Der Bereich CF 2 war bei ERP-2AP und bei ALLEGRO-EFM fachlicher Bedarfsträger. Zu seinen Aufgaben gehörte die fachliche Abnahme des Verfahrens in der Projektphase. Während des Betriebes ist er für die fachlich richtige Umsetzung des Verfahrens verantwortlich.

Der Bereich EFM war beim „Zwei-Augen-Prinzip“ in ALLEGRO für die technische Entwicklung und Wartung des Verfahrens zuständig.

Der Bereich CF 2 konnte nicht belegen, welche konkreten fachlichen Anforderungen er für die Entwicklung des Verfahrens an den Bereich EFM gestellt hatte. Der Bereich EFM nahm selbst eine unzureichende inhaltliche Begleitung durch den fachlich verantwortlichen Bereich CF 2 wahr. Aus seiner Sicht war die Rolle, die der Bereich CF 2 wahrnahm, unklar. Zudem war zunächst offen, welche Aufgaben weitere Bereiche hatten, die z. B. für die Umsetzung der Leistungen in den gE verantwortlich waren.

Die Bundesagentur schloss ihre WiBe mit einer Erfolgskontrolle zum Ende des Projektes ab. Zu diesem Zeitpunkt wusste sie nicht, inwieweit die entwickelten Verfahren in der Praxis umgesetzt werden.

6.2 Würdigung

Die Zentrale der Bundesagentur hat es bisher versäumt, ihre Entwicklungen zum Einsatz analytischer Verfahren zu koordinieren. So hat sie verschiedene Verfahren mit dem Schwerpunkt „Kassensicherheit“ entwickelt, die sie nicht aufeinander abgestimmt hat. Ein entscheidender Faktor für die Probleme im Verfahren ALLEGRO ist z. B. die fehlende Akzeptanz bei den Beschäftigten gewesen. Ähnliches ist beim Betrieb von ERP-2AP zu erwarten, wenn die Bundesagentur ihre Prüfregeln nicht optimiert und die Zahl der regelmäßig zu prüfenden Fälle nicht reduzieren kann. Prüfregeln mit niedriger Treffergenauigkeit verursachen für die Beschäftigten in den Agenturen und gE zusätzlichen Aufwand ohne entsprechenden Mehrwert.

Die Bundesagentur hätte die Wirtschaftlichkeit der entwickelten Methoden nicht losgelöst vom tatsächlichen Einsatz in den Agenturen und gE betrachten dürfen. Ein nicht oder nur selten genutztes Verfahren kann seine Wirkung nicht entfalten. Damit haben auch die eingesetzten Haushaltsmittel ihr Ziel verfehlt.

6.3 Empfehlung

Die Bundesagentur muss die Entwicklung von analytischen Verfahren stärker als bisher koordinieren. Dabei hat sie u. a. auch die fachlichen Anforderungen zu berücksichtigen, um die Kassensicherheit „auf andere Art und Weise“ sicherzustellen. Wir haben der Bundesagentur empfohlen, aus ihren

Erfahrungen bei ALLEGRO und ERP-2AP zu lernen. Sie sollte keine Prüfmethoden mehr einsetzen, die viele Prüffälle bei gleichzeitig niedriger Treffergenauigkeit ausweisen.

Die Bundesagentur muss die Wirtschaftlichkeit ihrer Verfahren unter den konkreten Einsatzbedingungen bewerten und auch nach der Einführung mittels einer Erfolgskontrolle überprüfen.

6.4 Stellungnahme

Die Bundesagentur hat mitgeteilt, dass die Verfahren ALLEGRO und ERP völlig unterschiedlich und nicht vergleichbar seien. Sie habe auch nicht versäumt, ihre Entwicklungen zum Einsatz analytischer Verfahren zu koordinieren. Insbesondere habe sie ein Anforderungsmanagement eingerichtet, welches die fachlichen Anforderungen hinsichtlich der vorhandenen Entwicklungsressourcen priorisiere. Darüber hinaus habe sie bereichsübergreifende Statusrunden und Steuerungskreise eingerichtet, um analytische Verfahren zu steuern und zu koordinieren.

6.5 Abschließende Würdigung

Uns ist bewusst, dass und in welchen Punkten sich die Verfahren ALLEGRO und ERP unterscheiden. Dies gilt jedoch nicht im selben Maße für die darauf aufsetzenden analytischen Verfahren ALLEGRO-EFM und ERP-2AP. Wir stellten bei der Prüfung fest, dass das Projekt ERP-2AP einige der Mängel von ALLEGRO-EFM abstellen konnte. Andere wiederum, z. B. hinsichtlich der Anzahl an zu versendenden Prüffällen bei gleichzeitig niedriger Treffergenauigkeit, traten erneut auf. Wir bleiben daher bei unserer Würdigung, dass die Bundesagentur diese Mängel bei geeigneter Koordination hätte vermeiden können.

Zur Wirtschaftlichkeit und unserer Empfehlung einer Erfolgskontrolle ihrer Verfahren hat sich die Bundesagentur nicht geäußert. Wir erwarten, dass die Bundesagentur eine abschließende Erfolgskontrolle von ERP-2AP durchführt.

7 Übergreifende Bewertung

Die Bundesagentur erprobte den Einsatz von KI und analytischen Verfahren, um perspektivisch die Kassensicherheit „auf andere Art und Weise“ als im Vier-Augen-Prinzip gewährleisten zu können. Dabei haben wir fachliche,

technische und organisatorische Mängel festgestellt (vgl. auch Teilprüfungsmitteilung 1).

Grundsätzlich halten wir den Einsatz von KI und analytischer Verfahren dennoch für geeignet, um Prozesse zu automatisieren und damit die Verwaltung zu entlasten.

Wir haben der Bundesagentur empfohlen, an ihren Zielen festzuhalten und weiterhin zu untersuchen, inwieweit der Einsatz von KI und analytischen Verfahren geeignet sein kann, um

- Prüfregeln zu entwickeln, die eine höhere Treffergenauigkeit als die bisher genutzte nzStpp erzielen und
- die Fehlerquote bei Anordnungen in buchführungsrelevanten Verfahren zu verringern.

Ihre Vorgehensweisen und Methoden in diesem innovativen und dynamischen Technologieumfeld hat sie fortlaufend zu evaluieren, bedarfsgerecht weiterzuentwickeln und kontinuierlich zu verbessern. Unsere Empfehlungen zeigen hierzu erste Ansätze auf.

7.1 Stellungnahme

Die Bundesagentur hat unserer übergreifenden Bewertung zugestimmt. Sie hat ergänzt, dass sie auch zukünftig Vermögensschäden vermeiden wolle und dies als Kernelement der Kassensicherheit betrachte. Hier sei es im Projekt ERP-2AP gelungen, die Vorteile der „Maschine" für die Bundesagentur zu erschließen. Sie könne die Beschäftigten der Agenturen und gE so gezielt dabei unterstützen, Fehler zu vermeiden.

7.2 Abschließende Würdigung

Wir schließen die Prüfung an dieser Stelle ab, behalten uns aber vor, die hier adressierten Punkte erneut aufzugreifen.

Romers Rabenschlag Essers Fasswald

Anhang – Arbeitspakete im Projekt ERP-2AP

Im Folgenden beschreiben wir kurz die Inhalte der einzelnen Arbeitspakete (AP) des Projektes ERP-2AP.

AP0: Doppelte Anordnungen

Mit dem AP0 entwickelte die Bundesagentur ausschließlich deterministische Prüfungen. Sie befasste sich im AP0 mit dem Themenbereich der doppelten Anordnungen.

Die deterministischen Prüfungen des AP0 weisen Anordnungen als auffällig aus, wenn bei mindestens zwei Anordnungen die Werte in mehreren zuvor festgelegten Feldern (z. B. Belegart, Bankverbindung, Betrag, Kostenstelle) innerhalb eines definierten Zeitraumes übereinstimmen.

Befindet sich mindestens eine Anordnung noch in der Genehmigungsphase, spricht die Bundesagentur von einer Doppelerfassung, ansonsten von einer Doppelauszahlung.

AP1: Adressprüfung und Zahlweg

Mit dem AP1 entwickelte die Bundesagentur Prüfregeln zu den Themenbereichen Adressprüfung und Zahlweg.

Bei der Überprüfung der **Adressdaten** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgenden Merkmalen als auffällig aus:

- Anordnungen, bei denen keine Adresse zum gewählten Geschäftspartner hinterlegt ist.
- Anordnungen, bei denen keine der hinterlegten Adressen zum Zeitpunkt der Nettofälligkeit der Forderung gültig ist.

Bei der Überprüfung des **Zahlweges** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgenden Merkmalen als auffällig aus:

- Anordnungen, bei denen für den Zahlweg „SEPA-Überweisung" keine gültige Bankverbindung im SEPA-Raum hinterlegt ist.
- Auszahlungsanordnungen außerhalb des SEPA-Raums, bei denen keine gültige Bankverbindung im Ausland hinterlegt ist.

- Anordnungen, bei denen für den Zahlweg „Barzahlung" der Höchstbetrag für Barzahlungen zum gewählten Geschäftspartner überschritten ist.
- Anordnungen, bei denen ein Zahlweg, der für einen bestimmten Sachverhalt reserviert ist (bspw. Kindergeld), für einen anderen Sachverhalt ausgewählt ist.

AP2: Zahlungsempfänger

Mit dem AP2 entwickelte die Bundesagentur Prüfregeln zu dem Themenbereich Zahlungsempfänger.

Die Prüfregeln der Bundesagentur weisen z. B. Anordnungen mit folgenden Merkmalen als auffällig auf:

- Anordnungen, bei denen ein Geschäftspartner mit der Nummer "6" beginnend genutzt wurde.[1]
- Annahmeanordnungen, bei denen bei einem Unterhaltssachverhalt der Verwendungszweck leer ist oder der Verwendungszweck keine plausiblen Eingaben (Vorschrift ist Vorname, Nachname, Geburtsdatum) aufweist.

AP3: Verwendungszweck und Fälligkeit

Mit dem AP3 entwickelte die Bundesagentur Prüfregeln zu den Themenbereichen Verwendungszweck und Fälligkeit.

Bei der Überprüfung des **Verwendungszwecks** weisen die Prüfregeln der Bundesagentur Anordnungen mit folgendem Merkmal als auffällig auf:

- Anordnungen, bei denen der Verwendungszweck leer ist.

Bei der Überprüfung der **Fälligkeit** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgenden Merkmalen als auffällig auf:

- Anordnungen, bei denen die Differenz zwischen Belegdatum und Nettofälligkeit den üblichen Zeitraum übersteigt.
- Anordnungen, die umgebucht wurden und die Fälligkeit in der umgebuchten Anordnung und der ursprünglichen Anordnung nicht übereinstimmen.

---

[1] Geschäftspartnerstammdaten, die aus FINAS-FE migriert wurden, dürfen nicht für das Erteilen neuer Anordnungen genutzt werden. Die migrierten Geschäftspartner sind an der mit der Ziffer 6 beginnenden Geschäftspartnernummer erkennbar.

- Annahmeanordnungen, bei denen die Fälligkeit mit dem Erfassungsdatum übereinstimmt.

AP4: Finanzposition, Betrag, Zahlweg

Mit dem AP4 entwickelte die Bundesagentur Prüfregeln zu den Themenbereichen Finanzposition, Betrag und Zahlweg.

Bei der Überprüfung der **Finanzposition** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgenden Merkmalen als auffällig auf:

- Anordnungen, bei denen bspw. der Verwendungszweck darauf schließen lässt, dass die Anordnung nicht der korrekten Finanzposition zugeordnet wurde.
- Anordnungen, bei denen eine Finanzposition nicht der korrekten Finanzstelle für IT-Ausgaben zugeordnet wurde.

Bei der Überprüfung des **Betrages** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgendem Merkmal als auffällig auf:

- Anordnungen, deren Beträge sich von den Betragsbereichen gleichartiger Anordnungen im Betrachtungszeitraum unterscheiden.

Üblichen Betragsbereiche werden bspw. mittels analytischer Clusterverfahren ermittelt. Gleichartige Anordnungen werden über Finanzpositionen, -stellen oder Geschäftspartner identifiziert.

Bei der Überprüfung des **Zahlweges** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgenden Merkmalen als auffällig auf:

- Anordnungen, bei denen an einem Tag für denselben Geschäftspartner Auszahlungen angelegt werden
- Anordnungen, bei denen definierter Betrag überschritten wird.
- Auszahlungsanordnungen, bei denen die IBAN der Bundesagentur als Empfänger hinterlegt ist.

AP5: Abschlagszahlungen, Buchungsdatum und Steuerkennzeichen

Mit dem AP5 entwickelte die Bundesagentur Prüfregeln zu den Themenbereichen Abschlagszahlungen, Buchungsdatum und Steuerkennzeichen.

Bei der Überprüfung der **Abschlagszahlungen** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgendem Merkmal als auffällig auf:

- Abschlagszahlungen, bei denen der Eintrag im Feld "Referenz" nach dem Anwenderhandbuch ERP nicht zulässig ist.

Bei der Überprüfung des **Buchungsdatums** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgendem Merkmal als auffällig auf:

- Anordnungen, bei denen das Buchungsdatum nicht dem Datum der Erfassung der Anordnung entspricht.

Bei der Überprüfung des **Steuerkennzeichens** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgendem Merkmal als auffällig auf:

- Anordnungen, bei denen das gewählte Steuerkennzeichen in Verbindung mit dem Geschäftsbereich nicht zulässig ist
- Anordnungen, bei denen das gewählte Steuerkennzeichen in Verbindung mit dem Anordnungstyp nicht zulässig ist

AP6: Online-Umsetzung

Bei den überwiegend deterministischen Prüfregeln der AP0 – AP5 handelte es sich um nachgelagerte Prüfungen, die nach der systemseitigen Eingabe angewendet werden.

Mit dem AP6 entwickelte die Bundesagentur die sogenannte Online-Umsetzung. Die Prüfregeln der Online-Umsetzung werden direkt, ohne zeitlichen Versatz, im Verlauf des ERP-Buchungsvorgangs durchgeführt. Hierbei werden dem Anwender bereits vor dem Abschluss einer Anordnung Hinweise zu möglichen Fehleingaben mitgeteilt.

Bei einem Hinweis wird hierbei zwischen einer Warnmeldung, einer Fehlermeldung und einem Informationstext unterschieden:

- Warnmeldungen weisen den Anwender darauf hin, dass eine oder mehrere der getätigten Eingaben unüblich und daher zu prüfen sind.
- Fehlermeldungen weisen auf eine identifizierte Fehleingabe hin. Diese muss behoben werden, um eine Anordnung abschließen zu können.

- Informationstexte geben dem Genehmiger einer Anordnung bekannt, dass für die freizugebende Anordnung bereits eine identische Anordnung im System existiert.

Beispiel für eine Fehlermeldung (hier: Prüfung auf leere Verwendungszwecke)

Getätigte Eingabe:

Der Bearbeiter legt eine Annahme- oder eine Auszahlungsanordnung an, ohne eine Eingabe im Feld „Verwendungszweck" zu tätigen.

Fehlermeldung:

„Für die anzuweisende Anordnung ist kein Eintrag im Verwendungszweck. Bitte stellen Sie sicher, dass der Verwendungszweck einen plausiblen Eintrag aufweist."

AP7: Referenzen & Korrekturen

Mit dem AP7 entwickelte die Bundesagentur Prüfregeln zu dem Themenbereich Referenzen & Korrekturen.

Bei der Überprüfung der **Referenzen & Korrekturen** weisen die Prüfregeln der Bundesagentur z. B. Anordnungen mit folgendem Merkmal als auffällig auf:

- Annahmeabsetzungen, bei denen die Eintragungen nicht mit den Feldinhalten der ursprünglichen Annahmeanordnung übereinstimmen
- Umbuchungsanordnungen, zu denen kein entsprechender Ursprungsbeleg vorliegt

AP8: Statistische Betragsprüfung

Mit dem AP8 untersuchte die Bundesagentur die Güte der analytischen Prüfregeln zu „auffälligen Beträgen".

AP9: Erweiterung Reporting, Ergänzungen Referenzen/Korrekturen/Oberfläche

Mit dem AP9 ergänzte die Bundesagentur das Reporting um weitere Funktionen (u. a. zusätzliche Filterkriterien) und passte die Oberflächen in SAP an.

AP10: Umsetzung verbleibender Anforderungen

Mit dem AP10 entwickelte die Bundesagentur u. a. ein statistisches Modell mit Methoden des überwachten Lernens für die Überprüfung von Beträgen.

Darüber hinaus besserte die Bundesagentur auch bei Verfahren aus dem deterministischen Teil des Projektes nach (z. B. Anpassung Doppelzahlung Vorverfahren, Anpassung Anwendungsoberfläche).