Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Praktische Informatik
Fachgebiet Software Engineering

Vergleich von qualitativen und quantitativen Erhebungsmethoden für Erklärbarkeitsanforderungen am Beispiel einer Personalverwaltungssoftware

Comparison of Qualitative and Quantitative Elicitation Methods for Explainability Requirements Using the Example of Human Resources Management Software

Masterarbeit

im Studiengang Informatik

von

Raymond Ochsner

Prüfer: Prof. Dr. rer. nat. Kurt Schneider
Zweitprüfer: Dr. rer. nat. Jil Ann-Christin Klünder
Betreuer: M.Sc. Martin Obaidi

Hannover, 22.11.2024

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbstständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 22.11.2024

Raymond Ochsner

Zusammenfassung

Die Bedeutung von Erklärbarkeit als zentrale nicht-funktionale Anforderung (NFR) im Software-Engineering wächst mit der steigenden Komplexität moderner Softwaresysteme. Diese Masterarbeit untersucht verschiedene Methoden zur Erhebung von Erklärbarkeitsanforderungen und bewertet deren Effektivität und Effizienz. Im Rahmen dieser Arbeit wurden beim Beratungs- und IT-Dienstleistungsunternehmen adesso SE[1] qualitative und quantitative Erhebungen durchgeführt, darunter zwei Fokusgruppen, 18 Interviews und eine Onlineumfrage mit insgesamt 188 Teilnehmern. Als Anwendungsfall diente eine webbasierte Personalverwaltungssoftware.

Die Ergebnisse zeigen, dass Interviews sowohl bei der Ermittlung einzigartiger Erklärungsbedarfe als auch hinsichtlich des Personalaufwands am erfolgreichsten sind. Umfragen und Fokusgruppen erzielten im Vergleich dazu geringere Werte, wobei Fokusgruppen deutlich schlechter abschnitten. Darüber hinaus wurde eine von Droste et al. [15] entwickelte Taxonomie getestet, die eine systematische Kategorisierung von Erklärungsbedarfen ermöglicht. Die nachträgliche Einführung dieser bei der Durchführung der Erhebungen führte zu einer höheren Anzahl ermittelter Bedarfe.

Diese Arbeit empfiehlt Interviews als die wirksamste Methode zur Erhebung von Erklärbarkeitsanforderungen und betont das Potenzial der Taxonomie zur Unterstützung dieses Prozesses. Zukünftige Arbeiten könnten sich auf die Auswahl geeigneter Erhebungsmethoden, die Weiterentwicklung der Taxonomie, die Untersuchung weiterer Softwaretypen und die Analyse von Einflussfaktoren wie berufliche Rolle oder Nutzungsgewohnheiten konzentrieren. Die Erkenntnisse dieser Arbeit leisten einen Beitrag zur Gestaltung transparenter und benutzerfreundlicher Softwaresysteme.

[1] `https://adesso.de/`

Abstract

Comparison of Qualitative and Quantitative Elicitation Methods for Explainability Requirements Using the Example of Human Resources Management Software

The importance of explainability as a central non-functional requirement in software engineering is growing with the increasing complexity of modern software systems. This master's thesis examines various methods for eliciting explainability requirements and evaluates their effectiveness and efficiency. As part of this thesis, qualitative and quantitative surveys were conducted at the consulting and IT services company adesso SE[2], including two focus groups, 18 interviews and an online survey with a total of 188 participants. Web-based HR management software served as the use case.

The results show that interviews are the most successful in identifying unique explanatory needs as well as in terms of staff effort. Surveys and focus groups scored lower in comparison, with focus groups performing significantly worse. In addition, a study by Droste et al. [15] was tested, which enables a systematic categorization of explanatory needs. The subsequent introduction of this when conducting the surveys led to a higher number of identified needs.

This work recommends interviews as the most effective method for eliciting explainability needs and emphasizes the potential of the taxonomy to support this process. Future work could focus on the selection of suitable survey methods, the further development of the taxonomy, the investigation of further software types and the analysis of influencing factors such as professional role or usage habits. The findings of this work contribute to the design of transparent and user-friendly software systems.

[2] `https://adesso.de/`

Inhaltsverzeichnis

Kapitel 1

Einleitung

Softwaresysteme nehmen eine stetig wachsende Relevanz in unserem alltäglichen Leben ein. Zugleich werden diese zunehmend komplexer. Damit den Nutzern einer Software eine leichte Bedienung mit nachvollziehbarem Systemverhalten geboten wird, muss bei der Implementierung dieser seitens der Entwickler auf die Einbindung von Erklärbarkeit geachtet werden. Erklärbarkeit als NFR wird dabei in den letzten Jahren genauer erforscht und gewinnt immer mehr an Bedeutung [9] [23]. Darüber hinaus trägt sie maßgeblich zur Transparenz eines Softwaresystems bei [12].

Um Erklärbarkeit zu gewährleisten, ist es essenziell, Erklärungsbedarfe der Nutzer zu erheben [15] [39]. Eine Technik, die gleichermaßen für alle Stakeholder und Anforderungen in einem Projekt geeignet ist, gibt es nicht [35]. Erklärbarkeitsanforderungen sind wie andere NFR schwer zu erheben [9]. Welche Erhebungsmethoden sich hinsichtlich Effektivität und Effizienz dafür besonders gut eignen, bedarf jedoch weiterer Untersuchung [6].

1.1 Problemstellung

Eine Technik, die gleichermaßen für alle Stakeholder und Anforderungen in einem Projekt geeignet ist, gibt es nicht [35]. Trotz der zunehmenden Bedeutung der Erklärbarkeit in Software Engineering (SE) und Requirements Engineering (RE) fehlen in beiden Disziplinen spezifische Methoden zur Erhebung, Analyse und Dokumentation von Erklärbarkeitsanforderungen [6]. Während bereits Untersuchungen zur Erhebung solcher Anforderungen für alltägliche Softwaresysteme durchgeführt wurden [15], wurden dabei bisher noch keine verschiedenen Erhebungsmethoden systematisch verglichen. Somit existiert bisher kein Vergleich zwischen verschiedenen qualitativen und quantitativen Erhebungsmethoden für Erklärbarkeitsanforderungen.
Der Schwerpunkt bisheriger Arbeiten liegt vor allem auf der Erhebung von Erklärbarkeitsanforderungen im Kontext von Apps [25] [39]. Da sich

der Bedarf an Erklärungen jedoch je nach verschiedenen Softwaretypen unterscheidet, besteht ein dringender Forschungsbedarf in anderen Anwendungskontexten [15]. Zudem fehlt bislang eine Untersuchung, inwieweit eine konzipierte Taxonomie dabei helfen kann, Erklärungsbedarfe systematisch zu erfassen.

1.2 Lösungsansatz

Im Rahmen dieser Masterarbeit soll die Forschungslücke adressiert werden, indem qualitative und quantitative Erhebungsmethoden verglichen werden. Diese Untersuchungen werden beim Beratungs- und IT-Dienstleistungsunternehmen adesso SE[1] durchgeführt, das im Jahr 2023 über 10.200 Mitarbeitende beschäftigte. Die Erhebungen verfolgen das Ziel, Erklärungsbedarfe zu einer dort genutzten Personalverwaltungssoftware namens „adpeople", die über den Webbrowser zugänglich ist, zu ermitteln. Dadurch wird eine andere Perspektive und ein anderer Softwaretypus betrachtet als in bisherigen Forschungen, die sich auf Apps spezialisieren [25] [39]. Diese Bedarfe können anschließend in Erklärbarkeitsanforderungen überführt werden und dazu dienen, die Software nutzerfreundlicher zu gestalten.
Die Methoden werden im Hinblick auf ihre Effektivität und Effizienz evaluiert, um fundierte Aussagen darüber treffen zu können, welche Methoden unter welchen Bedingungen am besten für die Erklärbarkeitsanforderungserhebung geeignet sind.
Da der Einsatz von Taxonomien als Unterstützung bei Anforderungserhebungen dienen kann [15], wird ebenfalls untersucht, inwiefern dieser Ansatz zu einer Verbesserung der Erhebung von Erklärungsbedarf führt.
Dabei wird eine von Droste et al. [15] vorgestellte Taxonomie zur Kategorisierung von Erklärungsbedarf hinsichtlich ihrer Validität überprüft. Zudem soll berücksichtigt werden, dass quantitative Erhebungen typischerweise eine geringere Qualität im Vergleich zu qualitativen Ansätzen aufweisen. Daher sollen die Erkenntnisse aus den qualitativen Erhebungen genutzt werden, um die Ergebnisse der quantitativen Methode zu ergänzen und zu verbessern.

1.3 Struktur der Arbeit

In Kapitel 2 wird auf die Grundlagen des Requirements Engineerings mitsamt der Anforderungserhebung eingegangen. Zudem wird das Thema Erklärbarkeit von Softwaresystemen und Erklärungsbedarf erläutert. Relevante Evaluationsmetriken werden vorgestellt.

[1] `https://adesso.de/`

Kapitel 3 enthält verwandte Arbeiten im Bereich der Erklärbarkeit von Softwaresystemen und der Anforderungserhebung, die sich unterteilt in Erhebung von Erklärbarkeitsanforderungen, Vergleich von Erhebungsmethoden und der Anforderungserhebung ohne expliziter Erklärbarkeitsanforderungserhebung.
Daraufhin wird in Kapitel 4 das Konzept zum systematischen Vorgehen beschrieben, wobei auch die Forschungsmethodik mitsamt der Forschungsfragen vorgestellt werden.
Kapitel 5 beschreibt die Durchführungen der Erhebungsmethoden dieser Studie.
Folgend werden die Ergebnisse dieser Studie in Kapitel 6 vorgestellt.
Anschließend werden in Kapitel 7 die Forschungsfragen beantwortet und die Ergebnisse evaluiert und diskutiert. Auf mögliche Validity Threats wird ebenfalls eingegangen.
Abschließend werden im letzten Kapitel 8 die Ergebnisse dieser Arbeit zusammengefasst. Ein Ausblick geht auf künftige Anwendungsmöglichkeiten ein.

Kapitel 2

Grundlagen

In diesem Kapitel werden die Grundlagen, die für die Arbeit relevant sind, erklärt. Zum einen wird das RE und explizit die Anforderungserhebung konkretisiert. Anschließend wird auf die Erklärbarkeit von Softwaresystemen und auf Erklärungsbedarf eingegangen. Dabei wird ein zum Ermitteln von einzigartigen Erklärungsbedarfen genutztes Codingsystem erläutert. Relevante Evaluationsmetriken werden im letzten Teil dieses Kapitels vorgestellt.

2.1 Requirements Engineering

Für diese Arbeit sind die Grundlagen des RE, einem Teilbereich des Software Engineerings (SE) essenziell, da die Anforderungserhebung ein Teilschritt dieses Prozesses ist. Diese Arbeit orientiert sich an der Definition des International Requirements Engineering Boards (IREB)[34]:

Definition 1 *(Requirements Engineering nach IREB)*
Das RE ist ein systematischer und disziplinierter Ansatz zur Spezifikation und zum Management von Anforderungen mit den folgenden Zielen:

- *die relevanten Anforderungen zu kennen, Konsens unter den Stakeholdern über die Anforderungen herzustellen, die Anforderungen konform zu vorgegebenen Standards zu dokumentieren und die Anforderungen systematisch zu managen,*
- *die Wünsche und Bedürfnisse der Stakeholder zu verstehen und zu dokumentieren, sowie die Anforderungen zu spezifizieren und zu managen, um das Risiko zu minimieren, ein System auszuliefern, das nicht den Wünschen und Bedürfnissen der Stakeholder entspricht.*

Das RE hat eine steigende Signifikanz in der Softwareentwicklung erlangt. Diese wird unter anderem dadurch begründet, dass viele Projekte aufgrund von mangelhaftem RE scheitern [21] [32] [34]. Mangelhaftes RE definiert sich

oftmals durch fehlerhafte Anforderungen. Die nachträgliche Beseitigung von Anforderungsfehlern verursacht hohe Kosten und ist einer der häufigsten Anlässe für Projektabbrüche [32]. Ein weiter Grund ist die implizite Annahme, dass vieles selbstverständlich und nicht erwähnenswert sei [34].
Zur Definition von RE haben Börger et al. [5] ein Referenzmodell ausgearbeitet (siehe Abbildung 2.1), das das RE in zwei Teilbereiche unterteilt: Requirements Analysis und Requirements Management. Ersteres besteht aus den Phasen Erhebung, Interpretation, Verhandlung, Dokumentation und Validierung & Verifikation. Das Requirements Management umfasst demnach die Phasen Änderungsmanagement und Nachverfolgung. Dieses Referenzmodell dient in dieser Arbeit als Grundlage für die Unterteilung des RE, da in vielen anderen Arbeiten der Teilbereich Requirements Analysis als RE aufgefasst wird und diese Unterteilung Unklarheiten zu dem Begriff RE beseitigt [21] [32] [35].

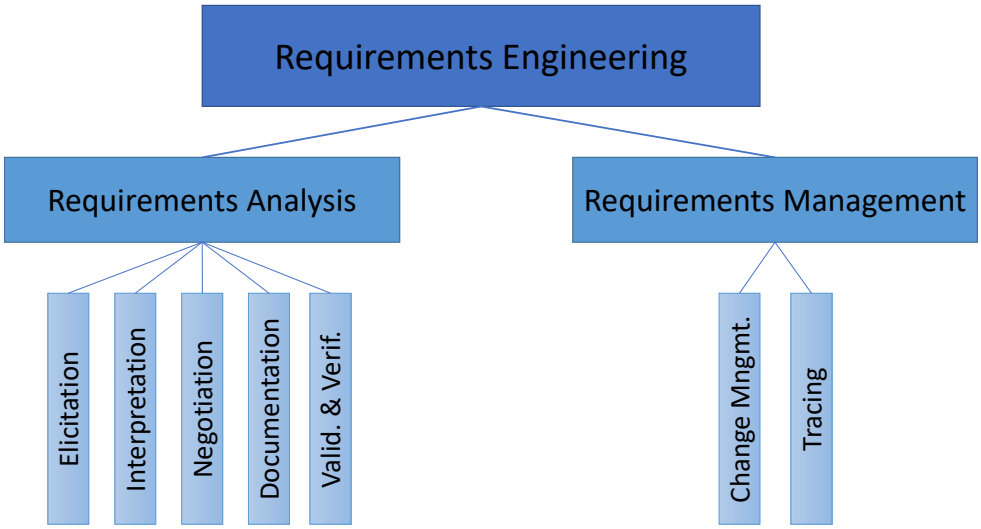

Abbildung 2.1: RE-Referenzmodell nach Börger et al. [5]

2.1.1 Anforderungserhebung

Diese Arbeit befasst sich vorwiegend mit der Erhebung von Erklärungsbedarfen, die in Anforderungen umformuliert werden können. Deswegen sind die Grundlagen der Anforderungserhebung relevant. Die Anforderungserhebung ist nach Börger et al. [5] die erste Phase des Requirements Analysis, wie in Abbildung 2.1 einsehbar. Sie bildet daher die Grundlage für gutes RE.
Fehlerfreie und vollständige Anforderungen sind von hoher Wichtigkeit für das RE und somit für die Qualität und den Erfolg eines Projektes [34]. Aufgrund dessen sollte bereits in dieser Phase darauf geachtet werden, dass Fehler vermieden werden, da sonst erhebliche Zusatzkosten durch

nötige Nachbesserungen entstehen können [32]. Rupp et al. [35] bezeichnen die Ermittlung von Anforderungen als „ein Handwerk, das viel Geschick erfordert". Umso wichtiger ist es, ein richtiges Vorgehen bei der Erhebung dieser einzusetzen [42]. Vor allem die frühe Einbeziehung von nichtfunktionalen Anforderungen (NFR) ist entscheidend, da sie die Auswahl von Technologien, die Zuweisung von Hardware und die Einhaltung von Sicherheitsstandards beeinflusst [42]. Anwar et al. [3] identifizierten vier Hauptfaktoren: Technikmerkmale, Eigenschaften der Stakeholder, Quellen der Anforderungen und die Projektumgebung, die allesamt Einflüsse auf die Wahl der richtigen Erhebungsmethoden nehmen. Die am häufigsten genutzten Erhebungsmethoden sind dabei Interviews, Workshops, Fokusgruppen und Umfragen [2] [3] [11] [31].
Es gibt keine Technik, die gleichermaßen für alle Stakeholder und Anforderungen in einem Projekt geeignet ist [35]. Im Verlauf eines Projekts müssen vielfältige Anforderungen unter unterschiedlichen Rahmenbedingungen und mit verschiedenen Beteiligten erfasst werden. Daher ist es in der Regel nicht ausreichend, nur eine einzelne Ermittlungstechnik einzusetzen [35].

2.2 Erklärbarkeit von Softwaresystemen

Im Zeitalter von Künstlicher Intelligenz (KI) werden Softwaresysteme immer undurchsichtiger und komplexer, weswegen es für den Nutzer zunehmend schwieriger wird, die Mechaniken und Entscheidungen der Software zu verstehen [9]. Aber auch in alltäglichen Softwaresystemen fehlen Nutzern Erklärungen für bestimmte Funktionen oder Verhaltensweisen [15]. Um dem entgegenzuwirken, kann es dabei helfen, Software erklärbarer zu machen.
Nach vereinfachter Definition von Chazette et al. [9] gilt ein System in Bezug auf einen Aspekt des System als erklärbar, wenn der Adressat im Kontext von einer Entität, die als Erklärer gilt, einen Informationskorpus erhält, der dem Adressaten verhilft, einen Aspekt des Systems in dem Kontext zu verstehen. Erklärbarkeit wird dabei als eine NFR angesehen [12] [23]. Somit ist es ein Teilaspekt des RE.
Erklärbarkeit wird zunehmend als bevorzugte Lösung angesehen, um die Transparenz eines Softwaresystems zu erhöhen [23]. Das Einbeziehen von Erklärbarkeit kann somit dabei helfen, dem Nutzer zu vermitteln, wieso ein bestimmtes Ergebnis entstanden und ihnen dabei helfen, bessere Entscheidungen zu treffen [12]. Außerdem hilft es dabei, Qualitätsaspekte wie Ethik und Fairness eines Softwaresystems zu erreichen [12]. NFR sind jedoch schwer zu erheben, zu verhandeln und zu validieren, somit gilt dies wahrscheinlich auch für Anforderungen an die Erklärbarkeit [9].
Dabei sollte darauf geachtet werden, dass die Einbindung von Erklärungen auch negative Effekte haben kann, da sie andere NFR wie Benutzerfreundlichkeit stark beeinflussen können [12].

Somit müssen Anforderungsanalysten Kosten und Vorteile bei der Präsentierung von Erklärungen abwägen [12].

2.2.1 Erklärungsbedarf

Treten bei einem Nutzer Fragen bezüglich der Erklärbarkeit auf, so spricht man von Erklärungsbedarf [39]. Droste et al. [15] entwickelten eine Taxonomie, um Erklärungsbedarfe zu kategorisieren. Abbildung 2.2 zeigt eine ins Deutsche übersetzte Version mit der Erweiterung durch die Kategorien *Business* und *Nicht vorhandenes Feature*.

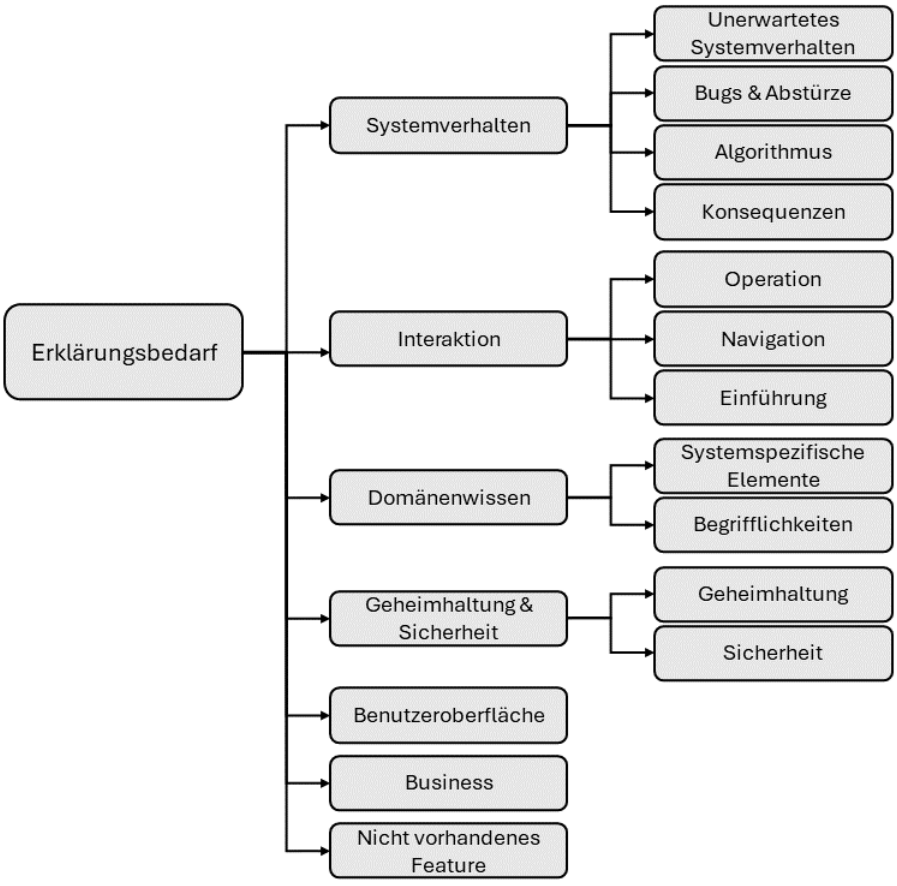

Abbildung 2.2: Übersetzte und erweiterte Taxonomie für Erklärungsbedarfe nach Droste et al. [15]

Einige kategorisierte Erklärungsbedarfe aus dem Datensatz [30] von Dominik Kupczyks Masterarbeit [25] dienen für die folgende Liste an Beispielen:
Systemverhalten:
Erklärungsbedarfe bezüglich des Verhaltens des Systems.

- Unerwartetes Systemverhalten: Für den Nutzer unerklärliches Verhalten.
  Beispiel: „Wieso kann ich mich nicht einloggen?“
- Bugs & Abstürze: Fehlermeldungen oder Abstürze.
  Beispiel: „Warum stürzt die Anwendung ab?“
- Algorithmus: Fragen zu den Ergebnissen der Funktionen.
  Beispiel: „Warum werden mir falsche Informationen vorgeschlagen?“
- Konsequenzen: Fragen zu Auswirkungen verschiedener Aktionen.
  Beispiel: „Was passiert, sobald ich meine Daten hier eingegeben habe?“

Interaktion:
Erklärungsbedarfe, die sich auf die Interaktion mit der Software beziehen.

- Operation: Unklarheiten zu Funktionen.
  Beispiel: „Wie kann ich mein Passwort zurücksetzen?“
- Navigation: Unklarheiten, die bei der Benutzung auftreten.
  Beispiel: „Wie komme ich zu den Einstellungen?“
- Einführung: Wenn nach einem Tutorial gefragt wird.
  Beispiel: „Ich brauche eine Anleitung, wie ich zu den Einstellungen komme.“

Domänenwissen:
Fragen zu indirekten Systemaspekten.

- Systemspezifische Elemente: Fragen, die an die spezifische Software gerichtet sind.
  Beispiel: „Wieso benötige ich Zugriff auf Dateien, die ich nicht aktiv anschaue?“
- Begrifflichkeiten: Unklare Abkürzungen oder Fachbegriffe.
  Beispiel: „Was bedeutet COP?“

Geheimhaltung und Sicherheit:
Fragen zu Datensicherheit und Vertraulichkeit.

- Geheimhaltung: Wie mit den gesammelten Daten verfahren wird.
  Beispiel: „Werden meine Daten weitergeleitet?“
- Sicherheit: Sicherheitsbedenken.
  Beispiel: „Können meine Daten gehackt werden?“

Benutzeroberfläche: Fragen zum Design.
Beispiel: „Warum ist die Suchleiste unten?“

Business: Fragen bezüglich des unternehmensspezifischen Businessmodells. Beispiel: „Warum ist die App teurer geworden?“

Zusätzlich wird in dieser Arbeit die Kategorie „nicht vorhandenes Feature“ betrachtet.

Nicht vorhandenes Feature: Wenn ein bestimmtes Feature fehlt.
Beispiel: „Wann kommt das Feature, um sich Bewertungen anzeigen zu lassen?“

2.2.2 Coding von Erklärungsbedarf

Ein spezifischeres Codingsystem wurde von Kliti Nikollau [29] entwickelt. Dabei werden mehrere Aspekte betrachtet, wodurch eine detaillierte Kategorisierung ermöglicht wird. Das Codingsystem ist dabei speziell an mobile Apps gerichtet. Die verschiedenen Kriterien können aus der Masterarbeit von Kliti Nikollau [29] übersetzt entnommen werden:

Das Kriterium Zeitkontext

- **Zukunft**
  Definition: Kennzeichnet Erklärungsbedarfe zu zukünftigen Plänen oder Entwicklungen.
- **Update**
  Definition: Kennzeichnet Fragen zu Problemen im Zusammenhang mit Updates.
- **Zeitunabhängig**
  Definition: Kennzeichnet Bewertungen ohne spezifischen Zeitbezug.

Das Kriterium Unerwartetes Systemverhalten

- **Leistungsverschlechterung**
  Definition: Bewertungen, die eine Verschlechterung der Leistung erwähnen.
- **Datenverlust**
  Definition: Bewertungen, die Probleme mit dem Verlust von Benutzerdaten ansprechen.
- **Falsche Ausgabe**
  Definition: Bewertungen über fehlerhafte Ausgaben der App.
- **Unerwartetes Interaktionsresultat**
  Definition: Bewertungen zu unerwarteten Ausgabenproblemen.

- **Inkonsistentes Verhalten**
  Definition: Bewertungen über Probleme mit inkonsistentem Verhalten der App.

- **Sonstiges**
  Definition: Fälle, die in keine der anderen Kategorien passen.

Das Kriterium Feedback zu Funktionen

- **Funktion geändert**
  Definition: Bewertungen zu Änderungen an bestehenden Funktionen.

- **Funktion entfernt**
  Definition: Bewertungen, die das Entfernen von Funktionen erwähnen.

- **Funktionswunsch**
  Definition: Benutzeranfragen für neue Funktionen.

- **Unklarheit über Funktion**
  Definition: Unsicherheiten der Benutzer über bestehende Funktionen.

- **Funktion vorhanden?**
  Definition: Unsicherheit, ob eine Funktion existiert oder auffindbar ist.

- **Tutorial benötigt**
  Definition: Bewertungen, in denen Benutzer Hilfe bei der App-Navigation oder Konzepten benötigen.

- **Keine Funktion erwähnt**
  Definition: Bewertungen, die keine Funktion ausdrücklich erwähnen.

Dabei wird zwischen Systemaspekte und nicht Systemaspekte unterschieden.
Das Kriterium Systemaspekte

Account

- **Login**
  Definition: Fasst Bewertungen zusammen, die Fragen zum Login-Prozess betreffen.

- **Profildaten**
  Definition: Beinhaltet Fragen zur Verwaltung von Benutzerdaten, wie das Ändern persönlicher Informationen oder das Zurücksetzen von Passwörtern.

Inhalte

- **Werbung**
  Definition: Bezieht sich auf Unklarheiten in Bezug auf die in der App angezeigte Werbung.

- **Feed**
  Definition: Fasst Fragen zum Inhalt zusammen, der in den Feeds der Benutzer angezeigt wird, insbesondere in sozialen Medien.

Entwicklungsrationale

- *Definition:* Fasst Bewertungen zusammen, in denen spezifische Entscheidungen im Entwicklungsprozess der App hinterfragt werden.

Sicherheit

- **Berechtigungen**
  Definition: Bezieht sich auf Bewertungen, die unklare Anforderungen an App-Berechtigungen ansprechen.

- **Verifizierungen**
  Definition: Fasst Bewertungen zu Problemen mit den Verifizierungsprozessen der App zusammen.

Basisfunktionen

- **Benachrichtigungen**
  Definition: Bezieht sich auf Unklarheiten über die Benachrichtigungen der App, deren Einstellungen oder Verhalten.

- **Datenaustausch und Synchronisation**
  Definition: Beinhaltet Bewertungen zu Problemen beim Teilen oder Synchronisieren von Daten zwischen Geräten oder Plattformen.

Das Kriterium Benutzeroberfläche (UI)

- *Definition:* Behandelt Bedenken der Nutzer hinsichtlich des Designs, Layouts und der Navigationselemente der App.

Das Kriterium Barrierefreiheit

- *Definition:* Fasst Bewertungen zusammen, die sich auf die Benutzerfreundlichkeit der App für Menschen mit Behinderungen oder speziellen Anforderungen an die Barrierefreiheit beziehen.

Domänenspezifität

- **App-spezifische Funktionen**
  Definition: Bezieht sich auf Bewertungen zu einzigartigen Funktionen, die spezifisch für die App sind.

- **Definition erforderlich**
  Definition: Beinhaltet Bewertungen, in denen Benutzer Definitionen für spezifische Begriffe verlangen, die sie in der App vorfinden.

Meta

- *Definition:* Fasst Bewertungen zusammen, bei denen der Erklärungsbedarf unklar oder nicht direkt ausgedrückt ist.

Das Kriterium Nicht-Systemaspekte

Hardware

- **Geräteunterstützung**
  Definition: Bezieht sich auf Bewertungen zur Kompatibilität der App mit bestimmten Geräten oder Betriebssystemen.

- **Konnektivität**
  Definition: Beinhaltet Bewertungen zu Verbindungsproblemen, wie z. B. Bluetooth-, WLAN- oder Mobilfunkproblemen.

Geschäftsaspekte

- **Finanzielles Konzept**
  Definition: Beinhaltet Bewertungen, die Verwirrung über Preismodelle, Abonnementpläne oder Gebühren ausdrücken.

- **Bestellungen**
  Definition: Behandelt Probleme beim Aufgeben oder Verfolgen von Bestellungen.

- **Zahlungen**
  Definition: Bezieht sich auf Verwirrung oder Probleme bei Zahlungsmethoden in der App.

- **Rückerstattungen**
  Definition: Fasst Bewertungen zu unklaren oder problematischen Rückerstattungsprozessen zusammen.

Kundensupport

- **Keine Antwort**
  Definition: Beinhaltet Bewertungen, in denen Benutzer sich über ausbleibende Antworten vom Kundensupport beschweren.

- **Unbrauchbare Antwort**
  Definition: Behandelt Bewertungen, in denen Benutzer die Antworten des Kundensupports als ineffektiv empfinden.

Da diese spezifisch an mobile Apps gerichtet sind, werden für diese Arbeit einige Anpassungen vorgenommen. So ist das Kriterium Inhalte unter Systemaspekte irrelevant, da weder Werbung noch Feed in der Personalverwaltungssoftware existiert.
Geschäftsaspekte werden ebenfalls nicht betrachtet, da die Nutzer Mitarbeiter eines Unternehmens sind, die für die Nutzung der Software keine Zahlungen tätigen.
Das Kriterium Kundensupport ist hier ebenfalls nicht relevant.
Als Subkategorien werden für diese spezifische Personalverwaltungssoftware angepasste Subkategorien verwendet. Als Beispiel dient hier „quick actions". Dies ist eine Funktion der in dieser Arbeit betrachteten Personalverwaltungssoftware, die es dem Nutzer ermöglicht, auf die wichtigsten Funktionen von der Startseite aus zuzugreifen.
Werden innerhalb eines Erklärungsbedarfs zwei bestimmte Funktionen in einen Zusammenhang gebracht, so werden diese mit einem Bindestrich definiert: „quick actions - links". Diese spezifischen Subkategorien sind vollständig im Anhang A.2 gelistet.
Je nach Detailgrad der Erklärungsbedarfe werden weitere Subkategorien benötigt, die diese bis zu einer vierten Stufe kategorisieren.
Die verschiedenen Kriterien können zu einer Zeichenkette konkateniert werden. Daraufhin wird das Vorkommen dieser einzigartigen Zeichenkette gezählt, wodurch man bestimmen kann, welche Erklärungsbedarfe den gleichen Bedarf benennen, obwohl sie in einer anderen Satzstruktur geäußert wurden. Dafür dienen folgende Beispiele:

„Wofür ist das report center da?"

„Was ist das Report Centre?"

Nach dem Coding mittels des Codingsystems würden diese beiden Aussagen in allen Aspekten die gleichen Werte vergeben bekommen. Die zusammengesetzte Zeichenkette resultiert sich aus den folgenden Werten, wobei nur eine Subkategorie vonnöten war:

1. **Time aspect**: Time independent
2. **Unexpected System behaviour**: No unexpected system behaviour mentioned
3. **Software Feature(s)**: Unclarity about feature(s)
4. **System aspect or non-system aspect**: System aspect
5. **Specific system aspect**: Development Rationale
6. **Subcategory 1**: Report Centre

Das Zählen der Anzahl der einzigartigen Zeichenketten definiert demnach die Anzahl der einzigartigen Bedarfe einer gesamten Liste an Erklärungsbedarfen.

2.3 Relevante Evaluationsmetriken

Die Auswertungen der erstellten Erklärungsbedarfslisten erfordern bestimmte Metriken wie die Interrater-Reliabilität, die mittels Fleiss' Kappa oder Cohen's Kappa bestimmt werden kann.

2.3.1 Interrater-Reliabilität

Zur Ermittlung der Übereinstimmungsrate zwischen Ratern können statistische Kennwerte wie Cohen's Kappa [13] für zwei Rater oder Fleiss' Kappa [18] für zwei oder mehr Rater verwendet werden. Landis und Koch [27] entwickelten eine Einteilung für Kappa-Werte, um die Stärke der Übereinstimmung zu klassifizieren. Die übersetzte Version dieser Kategorien ist in Tabelle 2.1 dargestellt.
Die Kappa-Werte können zwischen -1 und 1 variieren.

Kappa-Wert	**Maß der Übereinstimmung**
<0,00	Schwach
0,00-0,20	Leicht
0,21-0,40	Einigermaßen
0,41-0,60	Moderat
0,61-0,80	Erheblich
0,81-1,00	Fast perfekt

Tabelle 2.1: Übersetzte Kategorisierungstabelle nach Landis und Koch [27]

2.3.1.1 Cohen's Kappa

Zur Berechnung von Cohen's Kappa [13], das für die Ermittlung der Übereinstimmung zwischen zwei Ratern genutzt wird, wird die folgende Gleichung angewendet:

$$k = \frac{p_0 - p_e}{1 - p_e}$$

Dabei steht p_0 für den Anteil der tatsächlichen Übereinstimmungen und p_e für den erwarteten Anteil zufälliger Übereinstimmung [13].

2.3.1.2 Fleiss' Kappa

Fleiss' Kappa wird zur Ermittlung der Übereinstimmungsrate bei zwei oder mehreren Ratern verwendet. Die Berechnung erfolgt ähnlich wie bei Cohen's Kappa, jedoch weichen die Berechnungen für p_0 und p_e ab [18]:

$$p_0 = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right)$$

$$p_e = \sum_{j=1}^{k} p_j^2$$

Dabei bezeichnet N die Anzahl der Fälle und n die Anzahl der Rater.

Kapitel 3

Verwandte Arbeiten

In diesem Kapitel werden wissenschaftliche Arbeiten und Publikationen genannt, die sich entweder mit der Erklärbarkeit von Softwaresystemen oder der Analyse oder Durchführung von qualitativen und quantitativen Studien befassen.

3.1 Erklärbarkeit von Softwaresystemen

Die Erklärbarkeit von Softwaresystemen ist ein in den letzten Jahren immer populärer gewordenes Forschungsfeld [9] [11] [12] [23] [25].
Köhl et al. [23] analysierten die Möglichkeit und die dabei aufkommenden Herausforderungen, Erklärbarkeit in Softwaresystemen als eine NFR zu gestalten. Dabei wird betont, dass unterschiedliche Zielgruppen unterschiedliche Arten von Erklärungen benötigen, denn beispielsweise haben Ingenieure, Endnutzer und Juristen unterschiedliche Erwartungen und Bedürfnisse hinsichtlich der Erklärbarkeit von Systemen. Aufgrund dessen müssen Erklärungen kontext- und zielgruppenabhängig sein. Zur systematischen Erhebung und Spezifikation von Erklärbarkeitsanforderungen entwickelten die Autoren einen Katalog an Fragen, die als Grundlage für die Formulierung spezifischer Erklärbarkeitsanforderungen dienen. Da Erklärbarkeitsanforderungen als nicht-funktionale Anforderungen angesehen werden, bedeutet dies, dass sie nicht als absolute Vorgaben, sondern als Abwägungen im Entwicklungsprozess berücksichtigt werden sollten. Die Autoren betonen, dass weitere Forschung notwendig ist, um herauszufinden, wie Anforderungen und Softwareengineering-Techniken angewendet werden können, um erklärbare Systeme zu entwickeln [23].
Chazette und Schneider [12] publizierten 2020 ein Paper, in dem sie die Integration von Erklärbarkeit als NFR, um Transparenz von Software zu fördern, untersuchten. Durch die Durchführung einer Umfrage mit 107 Endnutzern wurde ermittelt, wie diese die aktuelle Transparenz von Software-Systemen wahrnehmen und welche Vorteile und Nachteile sie in

eingebetteten Erklärungen sehen. Die Ergebnisse zeigen, dass Erklärbarkeit sowohl dafür wichtig ist, um Transparenz und Vertrauen in eine Software zu erzielen, sondern auch Zusammenhänge zu anderen NFR besitzt. Diese Zusammenhänge können auch negativ behaftet sein, da Erklärungen beispielsweise einen starken Einfluss auf die Benutzerfreundlichkeit haben. Um die negativen Effekte zu minimieren, schlagen sie nutzerzentrierte Designtechniken (UCD) vor, die helfen sollen, Anforderungen zu erheben und zu gestalten. Anforderungsspezialisten müssen die Kosten und Vorteile der Bereitstellung von Erklärungen an die Nutzer abwägen [12].
2021 untersuchten Chazette et al. [9] die Erklärbarkeit von Softwaresystemen. Sie erkannten Erklärbarkeit als eine aufkommende NFR, die maßgeblich zur Qualität eines Systems beiträgt. Um die Integration von Erklärbarkeit in den Softwareentwicklungsprozess zu unterstützen, entwickelten die Autoren eine umfassende Definition, ein Modell sowie einen Wissenskatalog. Letzterer umfasst 57 Qualitätsaspekte, die durch die Erklärbarkeit beeinflusst werden und bietet somit eine wertvolle Ressource für Anforderungsingenieure, um die Interaktionen zwischen verschiedenen NFR besser zu verstehen und zu steuern. Demnach trägt die Arbeit von Chazette et al. [9] wesentlich dazu bei, das Verständnis und die Anwendung von Erklärbarkeit in der Softwareentwicklung zu fördern.
Unterbusch et al. [39] haben die Erreichung von Erklärbarkeit von Softwaresystemen untersucht. Die bisherige Forschung konzentriert sich vor allem auf Methoden zur Erstellung von Erklärungen für unterschiedliche algorithmische Ansätze konzentriert (z. B. maschinelles Lernen oder selbstadaptive Systeme). Wenig untersucht ist jedoch, welche Situationen und Verhaltensweisen erklärt werden sollten und wie man Anforderungen an die Erklärbarkeit erfassen kann. Aufgrund dessen analysierten die Autoren 1.730 App-Bewertungen aus acht verschiedenen Apps und entwickelten eine Taxonomie für Erklärungsbedarfe. Zusätzlich wurden verschiedene Ansätze getestet, um solche Bedürfnisse automatisch in App-Rezensionen zu erkennen. Der beste Klassifikator der Studie erreichte bei 486 neuen Bewertungen aus vier weiteren Apps eine gewichtete F-Score von 86 %. Damit leistet die Arbeit einen wichtigen Beitrag zum Verständnis, was Nutzer von erklärbaren Systemen erwarten [39].
Chazette et al. [11] untersuchten, wie Erklärbarkeit in modernen Softwaresystemen, insbesondere im Hinblick auf Transparenz und Fairness, praktisch umgesetzt werden kann. Sie führten eine Literaturübersicht durch, um Aktivitäten und Praktiken zur Gestaltung von Erklärungen zu identifizieren und entwickelten daraus eine Empfehlung mit sechs zentralen Aktivitäten für die Entwicklung erklärbarer Systeme. Diese Empfehlungen wurden durch Interviews mit 19 Softwarepraktikern auf ihre Durchführbarkeit hin überprüft. Die Studie betont, dass bestehende nutzerzentrierte Praktiken effektiv zur Erhebung, Implementierung und Prüfung von Erklärbarkeitsanforderungen genutzt werden können, ohne ein komplett neues Entwicklungsverfahren ein-

zuführen. Die Autoren rufen zu weiterer Forschung auf, um herauszufinden, ob andere Phasen des Software-Lebenszyklus angepasst werden müssen [11]. Dominik Kupczyk [25] untersuchte in seiner Masterarbeit die Präzision und den Recall von Datensätzen, die durch regelbasierte Filter aus Reviews bezogen wurden, um impliziten und expliziten Erklärungsbedarf zu identifizieren. Die Ergebnisse zeigen, dass Reviews mit explizitem Erklärungsbedarf mit einer Präzision von 79,8% und Reviews mit implizitem Erklärungsbedarf mit einer Präzision von 40,8% gefiltert werden können. Die Kontexteabhängigkeit von implizitem Erklärungsbedarf stelle eine Herausforderung dar, während expliziter Erklärungsbedarf aufgrund typischer Phrasenstruktur präziser gefiltert werden könne. Eine gleichmäßige Verteilung über alle Kategorien des Erklärungsbedarfs wird durch die Methodik nicht erreicht, wobei bestimmte Themen überrepräsentiert und andere unterrepräsentiert sind. Deep-Learning-Modelle zeigten bessere Ergebnisse bei der Erkennung von Erklärungsbedarf, insbesondere wenn sie binär auf impliziten und expliziten Erklärungsbedarf trainiert wurden, mit Precision- und Recall-Werten von über 90% auf ausgeglichenen Datensätzen. Auf unausgeglichenen Datensätzen verschlechterte sich der Recall erheblich. Die Arbeit integriert beide Ansätze in das von Kurtz [26] entwickelte Tool für Extrahierung und Analyse von Reviews aus App Stores. Dominik Kupczyk empfiehlt die Durchführung zukünftiger Studien zur Verbesserung der Modelle und zur detaillierten Analyse von Erklärungsbedarfskategorien [25].

3.2 Anforderungserhebung

Folgend werden verwandte Arbeiten, die sich mit der Erhebung von Erklärbarkeitsanforderungen, dem Vergleich verschiedener Erhebungsmethoden und der Anforderungserhebung ohne expliziter Erklärungsbedarfsermittlung befassen, genannt.

3.2.1 Erhebung von Erklärbarkeitsanforderungen

Mit der Erhebung von Erklärbarkeitsanforderungen wurde sich in den letzten Jahren vermehrt befasst [4] [15] [19] [40].
In dem Paper von Droste et al. [15] untersuchten die Forscher die wachsende Bedeutung von Erklärbarkeit als NFR in modernen Softwaresystemen. Um die Erklärungsbedarfe in verschiedenen Softwaresystemen zu identifizieren und zu kategorisieren, führten die Autoren eine Online-Umfrage mit 84 Teilnehmern durch. Diese wurden gebeten, ihre Erklärungsbedarfe bezüglich der drei zuletzt genutzten Softwaresysteme zu schildern. Aus den Antworten wurden 315 spezifische Erklärungsbedarfe extrahiert und kategorisiert, woraus sie dann eine Taxonomie bildeten. Die Untersuchung zeigt außerdem, dass Erklärungsbedarfe nicht nur für komplexe KI-Anwendungen relevant sind, sondern auch für alltägliche Softwaresysteme, für die Benutzer häufig

Erklärungen für bestimmte Funktionen oder Verhaltensweisen wünschen. Insbesondere stellten die Autoren fest, dass das Bedürfnis, das Verhalten eines Systems zu erklären, welches typischerweise im Fokus der Forschung zur erklärbaren Künstlichen Intelligenz (XAI) steht, nicht die häufigste Art des Erklärungsbedarfs in alltäglichen Softwaresystemen sei. Vielmehr besteht der häufigste Erklärungsbedarf darin, die Interaktionen der Endnutzer mit der Software zu erklären. Verschiedene Arten von Softwaresystemen führen zu unterschiedlichen Erklärbarkeitsanforderungen. Zudem betonen die Autoren, dass Erklärungsbedarfe explizit im RE-Prozess berücksichtigt werden müssen, um angemessene Erklärungen für jede Art von Software bereitzustellen. Die entwickelte Taxonomie hilft dabei, diese Anforderungen zu strukturieren und sicherzustellen, dass die Benutzerfreundlichkeit und das Verständnis für die Systeme verbessert werden, indem spezifische und kontextgerechte Erklärungen integriert werden [15]. Die Forscher stellten ihre in der Studie aufgestellten Daten[1] online zur Verfügung.
Chazette et al. [10] entwickelten ein Qualitätsframework für Erklärbarkeit, der externe Abhängigkeiten, Merkmale von Erklärungen und Evaluationsmethoden zusammenfasst, um die Analyse, Operationalisierung und Bewertung von Erklärbarkeitsanforderungen zu erleichtern. Das Framework wurde auf basis einer Literaturstudie erstellt und mit Hilfe einer case study in einer Navigationsapp angewendet. Das vorgestellte Framework zeigt eine gute Anwendbarkeit bei der Erhebung von Erklärbarkeitsanforderungen, der Wahl von Erklärungsdesigns und der Evaluation von dem Erfolg der Erklärung auf. Außerdem zeigen die Ergebnisse, dass ein gut durchdachter Designprozess für Erklärungen wichtig sind, um Negativeffekte zu vermeiden. So helfen bereits simple und leicht integrierte Erklärungen dabei, einen positiven Einfluss auf Qualitätsziele zu erreichen [10].
Im Zuge der Bachelorarbeit von Jonas Bohnstedt [4] wurde mittels einer Online-Umfrage der Einfluss von Domänenwissen auf den Erklärungsbedarf von Softwarenutzern analysiert. Dabei untersuchte er die beiden Domänen Office und Browser. Die Umfrage erfasste sowohl subjektives als auch objektives Domänenwissen der Teilnehmer. Die Ergebnisse zeigten, dass es keinen starken Zusammenhang zwischen Domänenwissen und Erklärungsbedarf gibt. Der Erklärungsbedarf variierte je nach Domäne oder blieb auf einem ähnlichen Niveau über verschiedene Domänen hinweg. Außerdem ergaben weitere Analysen, dass demografische Daten das Domänenwissen beeinflussen und eine höhere Anwendungssicherheit mit zunehmendem Wissen einhergeht. Teilnehmer bevorzugten weniger detaillierte Erklärungen. Zur Validierung der Ergebnisse empfiehlt Jonas Bohnstedt größere und vielfältigere Datensätze sowie die Einbeziehung von Experten [4].
Eine Onlineumfrage im Bezug auf Erklärbarkeit hat Ronja Fuchs [19] in ihrer Bachelorarbeit durchgeführt. Die Arbeit untersucht, wie das Vertrauen der

[1] `https://zenodo.org/records/10871086`

Mitarbeitenden der Firma Kraftwerk in die Vorhersagen der erzeugten elektrischen Leistung von BHKWs durch Machine-Learning-Methoden, insbesondere LSTM-Modelle, gesteigert werden kann. Acht Stakeholdergruppen, die verschiedene Motivationen an den Vorhersagen haben, wurden identifiziert. Ein Konzept für Erklärungen der Prognosen wurde entwickelt und mit Hilfe eines High-Fidelity-Prototyps evaluiert. Die Umfrage zeigte eine klare Präferenz für Erklärungen durch Entscheidungsbäume und durch die Darstellung der Wetterverhältnisse. Auf Basis der Ergebnisse wurden neun Anforderungen formuliert und in einem Paper-Prototype implementiert. Der Ausblick umfasst die praktische Umsetzung des Konzepts in die Verwaltungs-Software der BHKWs, das Training zusätzlicher Machine-Learning-Modelle zur Erklärung der LSTM-Modelle und eine Nutzerstudie zur Überprüfung der Funktionalität und Nutzbarkeit. Weitere Forschung könne die Übertragbarkeit des Konzepts auf andere Umfelder und Zeitreihendaten untersuchen [19].
Julian Voges [40] ermittelte in seiner Masterarbeit Erklärbarkeitsanforderungen zur Erhöhung der Nutzerakzeptanz des Stimmungsanalysetools RoBERTa. Zuvor wurde das Tool mithilfe eines Datensatzes mit über 4000 Einträgen vortrainiert, bevor es dann in einem Workshop präsentiert wurde. An diesem Workshop waren elf Probanden beteiligt, die alle eine Abschlussarbeit in einem Informatikstudium bereits geschrieben haben oder gerade schrieben. Zusammen erhoben sie die Erklärbarkeitsanforderungen, die sie wie folgt der beurteilten Wichtigkeit nach sortierten: *Genauigkeit der Ergebnisse, Grenzen der Software, Wahrscheinlichkeiten für die Stimmungen, Schlüsselwörter, Beispiele, Größe des Trainingsdatensatzes, Funktionsweise des Tools und Bewertung der Software.* Die Umsetzung der Erklärbarkeitsanforderungen wurde auf ihre Realisierbarkeit überprüft und teilweise softwaretechnisch bzw. prototypisch umgesetzt. In einer anschließenden Online-Umfrage wurde die Nutzerakzeptanz des Stimmungsanalysetools mit und ohne den jeweiligen Erklärungen verglichen. Eine Analyse der Umfrageergebnisse ergab, dass die Nutzung des Tools mit Erklärungen wie Beispiele, Schlüsselwörter und Grenzen der Software signifikant häufiger beabsichtigt wurde als ohne Erklärungen. Außerdem wurde festgestellt, dass das Verständnis des Tools durch diese Erklärungen erheblich verbessert wurde. Statistische Tests bestätigten, dass die Umsetzung bestimmter Erklärbarkeitsanforderungen tatsächlich eine signifikante Erhöhung der Nutzerakzeptanz bewirkte [40].

3.2.2 Vergleich von Erhebungsmethoden

Dieses Unterkapitel umfasst verwandte Arbeiten, die einen Vergleich zwischen verschiedener Erhebungsmethoden ziehen.
Younas et al. [42] adressieren die oft vernachlässigte Erhebung von NFR in agilen Softwareentwicklungsprozessen. Dabei identifizieren die Autoren die mangelnde Kenntnis über NFR bei Entwicklern und Nutzern als

Hauptgrund. Sie entwickelten eine Guideline zur Erhebung von NFR, die sowohl Entwicklern als auch Nutzern in agilen Projekten zugutekommt. Diese Guideline umfasst Methoden wie Interviews, Beobachtungen, Fokusgruppen, Brainstorming und Prototyping, angepasst an die Anforderungen agiler Entwicklungsprozesse. Zur Validierung der vorgeschlagenen Guideline wurde eine Fallstudie durchgeführt, bei der Masterstudenten mithilfe der Guideline NFR aus einem eProcurement-Dokument, das Anforderungen großer EU-Projekte enthält, extrahierten. Die Ergebnisse dieser Fallstudie waren positiv und zeigten, dass die Guideline insbesondere für Entwickler und Nutzer mit geringer Erfahrung im Umgang mit NFR nützlich ist. Die Studie betont, dass die Einbeziehung von NFR in den frühen Phasen der Softwareentwicklung entscheidend ist, da sie die Auswahl von Technologien, die Zuweisung von Hardware und die Einhaltung von Sicherheitsstandards beeinflusst [42].

Anwar und Razali [3] führten eine empirische Studie durch, um eine praktische Leitlinie für die Auswahl von Anforderungserhebungsmethoden zu erörtern. Bestehende Leitlinien seien eher theoretischer Natur, weswegen sie hiermit einen praxisnahen Bezug schufen. Sie identifizierten vier Hauptfaktoren: Technikmerkmale, Eigenschaften der Stakeholder, Quellen der Anforderungen und die Projektumgebung, die allesamt Einflüsse auf die Wahl der richtigen Erhebungsmethoden nehmen. Außerdem wurden die Arten der Methoden in gesprächsorientierte, beobachtende, analytische und synthetische Erhebungsmethoden unterteilt. Die Ergebnisse zeigen, dass Experten gesprächsorientierte Methoden, wie Interviews und Workshops präferieren, wenn die Nutzer ein tiefes Domänenwissen besitzen. Fragebögen sind besser geeignet, wenn die Analysten etwas Wissen zu dem System haben, da sie somit den Nutzern bei Fragen helfen können. Die Autoren betonen, dass die Wahl der richtigen Erhebungsmethoden ausschlaggebend für gute Anforderungen und folglich gute Software sind. [3]

In der Arbeit von Alflen et al. [2] untersuchten die Autoren verschiedene Verfahren zur Anforderungserhebung und entwickelten ein Modell, das die Auswirkungen der Teambeteiligung und der Anzahl der verwendeten Techniken analysiert. Die Autoren führten eine systematische Literaturrecherche durch, um die gängigsten und effektivsten Techniken zu identifizieren. Zu den nach Häufigkeit sortierten Techniken gehören Interviews, Fragebögen, Use Cases, Brainstorming, Szenarien, Prototypen, Fokusgruppen, Workshops, Joint Application Development (JAD) und User Stories. Basierend auf dieser Literaturrecherche entwickelten die Autoren ein Modell zur Analyse der Auswirkungen von Teammitgliedern und der gleichzeitigen Nutzung mehrerer RE-Techniken auf die Qualität der Anforderungen. Dieses Modell wurde in einem experimentellen Rahmen mit 53 Studierenden des Studiengangs Information Systems an der Universität von São Paulo angewendet und validiert. Das Paper hebt hervor, dass die Wahl und Kombination der richtigen Techniken, ergänzt durch eine aktive

Einbindung der Benutzer, die Qualität und Genauigkeit der gesammelten Anforderungen verbessern kann. Insgesamt zeigt die Arbeit, dass ein systematischer und gut geplanter Ansatz zur Anforderungserhebung wesentlich zur erfolgreichen Entwicklung von Softwareprojekten beiträgt [2]. Pacheco et al. [31] führten 2018 eine systematische Literaturrecherche durch, um eine Übersicht über die Methoden zur Anforderungserhebung zu bieten, die zwischen 1993 und 2015 verwendet wurden. Durch die Analyse von 140 relevanten Studien versuchten die Autoren herauszustellen, welche Erhebungsmethoden aktuell genutzt werden und welche davon besonders effektiv sind. 109 dieser Paper halfen dabei die Forschungsfrage zu beantworten, welche Erhebungsmethoden aktuell im Einsatz sind. Die häufigsten waren dabei Interviews, Workshops, Fokusgruppen, Joint Application Developments (JADs), Qualitätsfunktionendarstellungen, Ethnographien, Szenarien und Prototyping. Da in ca. 22% der gefunden Studien mehr als eine Erhebungsmethode genutzt wurde, ist es ratsam, dies bei der Anforderungserhebung zu berücksichtigen. Zur Beantwortung der zweiten Forschungsfrage, welche ausgereiften Methoden die Effektivität der Erhebung erhöhen, haben sie 31 Studien identifiziert. Effektive Methoden sind demnach Interviews, Szenarien, Workshops, Fokusgruppen, agile Methoden, Brainstorming, Modellierung, Prototyping und Fragebögen. Da es jedoch keinen Vergleich zwischen ihnen unter gleichen Bedingungen gab, ist es wichtig zu erwähnen, dass jede Methode ihre eigenen Vorteile in spezifischen Situationen besitzt. Die Autoren betonen, dass die Requirements Engineers die Erhebungsmethoden auswählen, die am besten zum Projekt passen [31].

3.2.3 Anforderungserhebung ohne expliziter Erklärbarkeits-anforderungserhebung

Mishra et al. [28] gehen auf die Herausforderungen des RE in der Softwareentwicklung, insbesondere in kleinen und mittleren Unternehmen, ein. Derzeitige RE-Praktiken sind oft ad hoc, und gut geplante Techniken zur Anforderungserhebung sind entscheidend, um Nacharbeit zu vermeiden und die Projektergebnisse zu verbessern. Die aktive Einbeziehung von Kunden und Stakeholdern ist für konsistente und vollständige Anforderungen von hoher Bedeutung. Eine Kombination verschiedener Erhebungsmethoden wie von Interviews, Workshops und iterativer Entwicklung hat sich als effektiv für komplexe Projekte mit häufig wechselnden Anforderungen erwiesen. Durch Brainstorming-Sitzungen konnten Informationen konsolidiert werden. Konflikte wurden erkannt und Workshops gezielt vorbereitet. Diese Erfahrungen tragen dazu bei, geeignete RE-Techniken für zukünftige Projekte auszuwählen und den RE-Prozess zu standardisieren [28].
Hadar et al. [20] untersuchen die Auswirkungen von Domänenwissen auf

die Anforderungserhebung mittels Interviews. Die Studie zeigt, dass Domänenwissen sowohl positive als auch negative Auswirkungen auf die Kommunikation und das Verständnis der Bedürfnisse der Stakeholder haben kann. Die Forschung stützt sich auf eine empirische Untersuchung, bei der Analysten mit und ohne Domänenwissen befragt wurden. Es wird gezeigt, dass Domänenwissen die Art und Weise, wie Interviews geführt werden, beeinflusst, insbesondere in Bezug auf die Kommunikation mit den Kunden und das Verständnis ihrer Anforderungen. Die Studie liefert praktische Einsichten für die Anforderungserhebung, indem sie zeigt, dass heterogene Teams aus Analysten mit unterschiedlichen Domänenwissen-Niveaus Synergien erzeugen können, die den Anforderungserhebungsprozess unterstützen. Dabei wird betont, dass es wichtig ist, nicht nur einen, sondern mehrere Experten für das jeweilige Fachgebiet in einem Team zu haben, um die Natur des Domänenwissens zu berücksichtigen. Die Ergebnisse der Untersuchung stimmen mit früheren Erkenntnissen überein, liefern jedoch neue empirische Belege für die Auswirkungen von Domänenwissen im Kontext der Anforderungserhebung. Zukünftige Forschungen können auf diesen Erkenntnissen aufbauen, um die spezifischen Auswirkungen von Domänenwissen weiter zu untersuchen und mögliche organisatorische Rahmenbedingungen zu identifizieren, die diese Effekte verstärken [20].
Natalie Dehn [14] befasste sich in ihrer Masterarbeit mit der Evaluierung von Stimmungsanalysetools im Kontext von Softwareprojekten, da bestehende Tools trotz ihrer vielfältigen Funktionen kaum in der Industrie genutzt werden. Die Studie identifizierte verschiedene Stakeholdergruppen wie Entwickler, Projektleiter und Forscher und erhebt deren Anforderungen mittels iterativer Interviews. Dabei wurden zwei Iterationen durchgeführt und als initiale Anforderungsaufnahme Ergebnisse einer nicht veröffentlichten Studie benutzt, um den Teilnehmern Hintergrundwissen zu Stimmungsanalysetools zu liefern. Folgend wurden Anforderungen in Kategorien wie Stimmungserkennung, Benutzeroberfläche und Datenschutz strukturiert und priorisiert. Die Analyse bestehender Tools zeigte erhebliche Defizite bei der Erfüllung dieser Anforderungen, insbesondere in Bezug auf Kompatibilitätsprobleme, Benutzerfreundlichkeit und Funktionalität. Zukünftige Forschungsansätze könnten weitere Stakeholder einbeziehen, die Umsetzbarkeit der erhobenen Anforderungen genauer untersuchen und rechtliche Aspekte wie Datenschutzrichtlinien in verschiedenen Ländern berücksichtigen [14].

3.3 Abgrenzung von den Verwandten Arbeiten

Während frühere Studien häufig entweder qualitative (z.B. Interviews und Workshops, Arbeit von Julian Voges [40] oder Chazette et al. [10]) oder quantitative Ansätze (z.B. Umfragen, Studie von Droste et al. [15]) priorisieren, legt diese Arbeit besonderen Wert auf die Kombination und den Vergleich beider Methoden. Dies ermöglicht eine fundiertere Analyse und ein umfassenderes Verständnis der Erhebung von Erklärbarkeitsanforderungen.
Im Gegensatz zu konzeptionellen Arbeiten wie die von Kohl et al. [23] oder Chazette et al. [12] wird in dieser Arbeit überprüft, wie sich konkrete Erklärbarkeitsanforderungen in der Praxis bei einem Unternehmen erheben lassen. In der Arbeit von Ronja Fuchs [19] werden Erklärbarkeitsanforderungen in einem Unternehmen zwar ebenfalls erhoben, jedoch wird nur eine Erhebungsmethode dafür eingesetzt. Ein systematischer Vergleich zwischen verschiedenen Erhebungsmethoden für Erklärbarkeitsanforderungen fehlt.
Darüber hinaus wird in diesem Zusammenhang der Einsatz der Taxonomie für Erklärungsbedarf von Droste et al. [15] validiert.

Kapitel 4

Konzept zum systematischen Vorgehen

In diesem Kapitel wird das Konzept zum systematischen Vorgehen in dieser Studie vorgestellt. Dabei wird zunächst auf die Forschungsmethodik samt den Forschungsfragen dieser Arbeit und der zugehörender Goal-Question-Metric eingegangen. Anschließend folgt die Auswahl der Erhebungsmethoden, woraufhin die verschiedenen Vorgehensweisen bei der Planung der unterschiedlichen Erhebungsmethoden beschrieben werden.

4.1 Auswahl der Erhebungsmethoden

Für das strukturierte Vorgehen muss sich in dieser Arbeit auf bestimmte Erhebungsmethoden entschieden werden. Forschungen ergaben, dass die Kombination von mehreren Erhebungsmethoden effektiv sind [1] [22] [36] und dass unter allen Erhebungsmethoden gesprächsorientierte Techniken wie Interviews am häufigsten vorkamen [2] [3]. Interviews werden normalerweise für die Anforderungserhebung eingesetzt, wenn das System Nutzer mit verschiedenen Rollen beinhaltet [3]. Sie gelten als effektiv in sowohl globaler Softwareentwicklung als auch traditionellen Umgebungen, weil sie dem Anforderungsanalysten ermöglichen, bessere und detailliertere Informationen zu sammeln [31]. Vor allem werden gesprächsorientierte Methoden bevorzugt, da sie von Kommunikation leben und zu tiefgründigeren Diskussionen führen [3].
Zu den gesprächsorientierten Methoden zählen auch Workshops, Fokusgruppen und Interviews. Workshops helfen dabei, Diskrepanzen bei Informationen zwischen den Nutzern zu lösen [3]. Sie sind effektiv, da sie den Stakeholdern helfen zu diskutieren und zu kollaborieren [31]. Zudem werden sie als Empfehlung für die Anforderungserhebung angesehen [38].
Eine Studie von Chazette et al. [11] zeigt, dass Interviews, Fokusgruppen bzw. Workshops, Umfragen und Personas am effektivsten von erfahrenen

IT-Mitarbeitern aus der Industrie für die Erhebung von Erklärbarkeitsanforderungen betrachtet werden. Workshops werden in Papern, die verschiedene Anforderungserhebungsmethoden behandeln, oft mit Fokusgruppen zusammengefasst [3] [11] oder separiert betrachtet, ohne sie explizit zu differenzieren [42]. Krüger definiert Fokusgruppen als eine Art von kollaborativen Gruppen mit dem Zweck, die Meinungen von verschiedenen Stakeholdern zu elaborieren [24]. Dabei schafft der Forschende eine Umgebung, die die Teilnehmenden ermutigt, ihre Meinungen zu teilen, ohne sie unter Druck zu setzen [24]. Eine Fokusgruppe besteht dabei aus 5 bis 10 Teilnehmenden, die von einem Moderator geführt wird [24]. Fokusgruppen sind besonders effektiv, weil sie Diskussionen zwischen den Stakeholdern, mit dem Ziel Anforderungen zu formalisieren, fördern [31]. Alflen et al. [2] beschreiben Workshops als eine Methode, um möglichst Effizient Unklarheiten zu lösen und Fokusgruppen als eine, die es ermöglicht, möglichst viele Meinungen zu sammeln. Dies ist in Anbetracht dessen, dass die Motivation dieser Arbeit ist, möglichst viel Erklärungsbedarf zu ermitteln, praktisch. Und da die Nutzer eine Einleitung in das Thema der Erklärbarkeit für die Durchführung der Erhebungsmethoden benötigen und ein Moderator mit Kenntnissen in diesem Bereich bei Fragen hilfreich ist, wird in dieser Arbeit eine Fokusgruppe ausgewählt.
Bei hoher Nutzeranzahl empfehlen Anwar et al. [3] Umfragen durchzuführen. Umfragen hingegen erfordern gut überlegte Fragen und Antworten, die sorgfältig geplant werden müssen [3]. Fehlmann et al. [17] betonen, dass Umfragen nicht besonders gut dafür geeignet sind, um Benutzererfahrungen zu analysieren. Da sie oft nicht die gewünschten Ergebnisse liefern, sollten sie immer mit einer anderen Methode kombiniert werden, was in dieser Arbeit getan wird [8]. Interviews sollten als primäre Erhebungsmethode eingesetzt werden, weil sie in Kombination mit anderen Methoden die Qualität der Anforderungen erhöhen [36].
Studien ergaben zudem, dass die Auswahl an Teilnehmenden, die Erfahrungen mit dem Produkt haben, einen positiven Effekt auf die Anforderungserhebung haben [16]. Darauf wird in dieser Arbeit Rücksicht genommen.
Damit die Beseitigung von möglichen Fehler in den Abläufen der einzelnen Erhebungsmethoden nicht dazu führen, dass bestimmte Methoden wiederholt werden müssen, wird in dieser Arbeit mit zwei Fokusgruppen begonnen, mit 18 Interviews fortgefahren und mit Online-Umfragen abgeschlossen. Letztere mit einer erwünschten Beteiligungsrate von über 100 Nutzern.

4.2 Forschungsmethodik

Diese Masterarbeit untersucht verschiedene Erhebungsmethoden zur Ermittlung von Erklärbarkeitsanforderungen bei Nutzern einer Personalverwaltungssoftware. Das Hauptziel besteht darin, zu verstehen, wie unter-

schiedliche methodische Ansätze die Qualität und Effizienz der erfassten Anforderungen beeinflussen und welche Rolle eine strukturierte Taxonomie bei der Erhebung von Erklärungsbedarf spielt. Des Weiteren wird die Genauigkeit der Kategorisierung untersucht, indem der Einfluss von Interrater-Reliabilität, also der Übereinstimmung der Kategorisierungen durch verschiedene Rater, analysiert wird. Dafür werden in den folgenden Unterkapiteln die Forschungsfragen aufgestellt und mittels der Goal-Question-Metric Methode präzisiert.

4.2.1 Forschungsfragen

Die relevanten Forschungsfragen werden folgend gruppiert erläutert.

Einflüsse der Erkenntnisse in die weiteren Erhebungsmethoden:

RQ1.1: Welche Erkenntnisse aus den Fokusgruppen können für die Durchführung der Interviews verwendet werden?

RQ1.2: Welche Erkenntnisse aus den Fokusgruppen und Interviews können für die Gestaltung der Onlineumfrage verwendet werden?

Die ersten beiden Forschungsfragen konzentrieren sich auf die Übertragung von Erkenntnissen zwischen den verschiedenen Erhebungsmethoden. RQ1.1 untersucht dabei, wie Erkenntnisse aus den Fokusgruppen in die Gestaltung der Interviews einfließen können, um eine tiefere und zielgerichtetere Erklärungsbedarfsermittlung zu ermöglichen. RQ1.2 befasst sich damit, wie die Erkenntnisse aus den Fokusgruppen und Interviews die Struktur und den Inhalt der Onlineumfrage beeinflussen und so die Qualität der quantitativen Daten steigern.

Vergleich von verschiedenen Erhebungsmethoden:

RQ2.1: Inwieweit sind Fokusgruppen, Interviews oder Onlineumfragen effektiver für die Erhebung von Erklärbarkeitsanforderungen bei Nutzern einer Personalverwaltungssoftware?

RQ2.2: Inwieweit sind Fokusgruppen, Interviews oder Onlineumfragen effizienter für die Erhebung von Erklärbarkeitsanforderungen bei Nutzern einer Personalverwaltungssoftware?

RQ2.3: Wie unterscheiden sich die Ergebnisse der Fokusgruppen, der Interviews und der Onlineumfrage hinsichtlich der in den jeweiligen Kategorien erfassten Erklärbarkeitsanforderungen?

Im diesem Abschnitt geht es darum, die verschiedenen Erhebungsmethoden hinsichtlich ihrer Effektivität und Effizienz zu vergleichen. RQ2.1 untersucht, welche Erhebungsmethode besser geeignet ist, um Erklärbarkeitsanforderungen effektiv zu erheben, während RQ2.2 schließlich die Effizienz der verschiedenen Methoden im praktischen Kontext beleuchtet, indem sie den Personalaufwand und die Zeit berücksichtigt, die für die Durchführung der Erhebungen bei Nutzern einer Personalverwaltungssoftware erforderlich sind. RQ2.3 analysiert die Unterschiede in den Ergebnissen der Fokusgruppen, Interviews und der Onlineumfrage hinsichtlich der Kategorien von Erklärbarkeitsanforderungen.

Anwendung einer Taxonomie:

RQ3.1: Wie unterscheiden sich die erhobenen Erklärungsbedarfe bei Anwendung einer Taxonomie?

RQ3.2: Reicht die vorgegebene Taxonomie aus, um den gesamten erhobenen Erklärungsbedarf den jeweiligen Kategorien zuzuordnen?

Diese Forschungsfragen konzentrieren sich auf den Einsatz einer Taxonomie bei der Erhebung des Erklärungsbedarfs. RQ3.1 untersucht, ob die Einführung einer Taxonomie den Erklärungsbedarf der Nutzer verstärkt, indem sie zu einer höheren Menge an Erklärungsbedarfen führt.
RQ3.2 analysiert, ob die vorhandene Taxonomie ausreicht, um den gesamten erhobenen Erklärungsbedarf korrekt zu kategorisieren oder ob Anpassungen und Erweiterungen notwendig sind.

Interrater-Reliabilität bei Zuweisung der Taxonomiekategorien

RQ4.1: Inwieweit weicht die Einordnung in die Kategorien seitens der Nutzer im Vergleich zur Einordnung durch den Anforderungsanalysten ab?

RQ4.2: Inwieweit weicht die Einordnung in die Kategorien zwischen den Nutzern ab?

Abschließend werden die Übereinstimmungen bei der Kategorisierung des Erklärungsbedarfs untersucht. RQ4.1 befasst sich mit der Frage, inwieweit die Kategorisierungen von Nutzern und Anforderungsanalysten voneinander abweichen, um die Qualität der Kategorisierung und das Verständnis der Taxonomie zu bewerten.
RQ4.2 untersucht, wie groß die Variabilität der Kategorisierungen zwischen verschiedenen Nutzern ist, um die Konsistenz und Reliabilität des Kategorisierungsprozesses zu beurteilen.

4.2.2 Goal-Question-Metric

Zur Findung von geeigneten Metriken, die bei der Beantwortung der Forschungsfragen helfen, wurde die Goal-Question-Metric (GQM) Methode[7] genutzt:

1. Einflüsse der Erkenntnisse in die weiteren Erhebungsmethoden

RQ1.1: Welche Erkenntnisse aus den Fokusgruppen können für die Durchführung der Interviews verwendet werden?

- **Goal 1.1:** Verständnis der Übertragbarkeit und Nützlichkeit von Erkenntnissen aus Fokusgruppen für die Planung und Durchführung von Interviews.
- **Question 1.1.1:** Welche spezifischen Erkenntnisse aus den Fokusgruppen sind wertvoll für die Gestaltung der Interviewfragen und -methodik?
- **Metric 1.1.1.1:** Anzahl und Art der Erkenntnisse aus den Fokusgruppen, die zur Verbesserung der Interviewführung beitragen.

RQ1.2: Welche Erkenntnisse aus den Fokusgruppen und Interviews können für die Gestaltung der Onlineumfrage verwendet werden?

- **Goal 1.2:** Feststellen, inwiefern die Ergebnisse aus Fokusgruppen und Interviews für die Gestaltung der Onlineumfrage verwendet werden können.
- **Question 1.2.1:** Welche Erkenntnisse aus den qualitativen Methoden (Fokusgruppen, Interviews) sind direkt übertragbar in die quantitative Erhebung in Form einer Onlineumfrage?
- **Metric 1.2.1.1:** Liste der spezifischen Erkenntnisse, die in Umfragefragen umgesetzt werden können.

2. Vergleich von verschiedenen Erhebungsmethoden

RQ2.1: Inwieweit sind Fokusgruppen, Interviews oder Umfragen effektiver für die Erhebung von Erklärungsbedarfen bei Nutzern einer Personalverwaltungssoftware?

- **Goal 2.1:** Evaluierung der Effektivität von Fokusgruppen und Interviews zur Erfassung von Erklärungsbedarfen.
- **Question 2.1.1:** Welche qualitative Methode (Fokusgruppen oder Interviews) liefert effektiver eine höhere Menge an Erklärungsbedarfen?

- **Metric 2.1.1.1:** Vergleich der Effektivität beider Methoden basierend auf der Anzahl der einzigartigen Erklärungsbedarfe pro Person.

RQ2.2: Inwieweit sind Fokusgruppen, Interviews oder Umfragen effizienter für die Erhebung von Erklärungsbedarfen bei Nutzern einer Personalverwaltungssoftware?

- **Goal 2.2:** Analyse der Effizienz und des Aufwands jeder Methode in Bezug auf Ergebnisse und Personalaufwand.
- **Question 2.2.1:** Wie unterscheiden sich die Ergebnisse in Bezug auf Zeit und Personalaufwand zwischen den Methoden?
- **Metric 2.2.1.1:** Anzahl der einzigartigen Erklärungsbedarfe pro Person und Zeit.
- **Metric 2.2.1.2:** Anzahl der einzigartigen Erklärungsbedarfe pro zeitlichem Personalaufwand.

RQ2.3: Wie unterscheiden sich die Ergebnisse der Fokusgruppen, der Interviews und der Onlineumfrage hinsichtlich der in den jeweiligen Kategorien erfassten Erklärungsbedarfe?

- **Goal 2.3:** Untersuchung der Übereinstimmungen und Unterschiede der Erklärungsbedarfe, die in verschiedenen Methoden erfasst werden.
- **Question 2.3.1:** In welchen Kategorien variieren die Ergebnisse zwischen Fokusgruppen, Interviews und Umfrage und inwiefern?
- **Metric 2.3.1.1:** Anzahl und Art der Erklärungsbedarfe pro Methode sowie ihre Zuordnung zu den vordefinierten Kategorien.

3. Anwendung einer Taxonomie

RQ3.1: Wie unterscheiden sich die erhobenen Erklärungsbedarfe bei Anwendung einer Taxonomie?

- **Goal 3.1:** Bewertung, wie die Einführung einer Taxonomie die Menge an erhobenem Erklärungsbedarf beeinflusst.
- **Question 3.1.1:** Verändert sich die Menge des erhobenen Erklärungsbedarfs, wenn die Erhebungsmethode mit einer Taxonomie strukturiert wird?
- **Metric 3.1.1.1:** Anzahl der Erklärungsbedarfe bei Methode mit und ohne Taxonomieeinsatz.

RQ3.2: Reicht die vorgegebene Taxonomie aus, um den gesamten erhobenen Erklärungsbedarf den jeweiligen Kategorien zuzuordnen?

- **Goal 3.2:** Beurteilung der Vollständigkeit der bestehenden Taxonomie für die Kategorisierung des Erklärungsbedarfs.
- **Question 3.2.1:** Wie sicher sind sich die Teilnehmenden bei der Zuordnung der Erklärungsbedarfe zu den bestehenden Kategorien?
- **Metric 3.2.1.1:** Durchschnittlicher Wert der Sicherheit der Teilnehmenden bei der Zuordnung (Likert-Skala 1-5).
- **Question 3.2.2:** In welchem Umfang äußern die Teilnehmer zusätzliche Kategorien?
- **Metric 3.2.2.1:** Anzahl der zusätzlichen Kategorien, die von den Teilnehmern genannt werden.
- **Metric 3.2.2.2:** Anzahl der Teilnehmer, die neue Kategorien vorschlagen.

4. Interrater-Reliabilität bei Zuweisung der Taxonomiekategorien

RQ4.1: Inwieweit weicht die Einordnung in die Kategorien seitens der Nutzer im Vergleich zur Einordnung durch den Anforderungsanalysten ab?

- **Goal 4.1:** Untersuchung der Konsistenz in der Kategorisierung von Erklärbarkeitsanforderungen zwischen Nutzern und Analysten.
- **Question 4.1.1:** Wie hoch ist die Übereinstimmung zwischen der Nutzer- und der Analystenkategorisierung der Erklärungsbedarfe?
- **Metric 4.1.1.1:** Interrater-Reliabilität in Form von Cohen's Kappa zwischen Nutzern und Analysten bei der Kategorisierung der Erklärungsbedarfe.
- **Metric 4.1.1.2:** Interrater-Reliabilität in Form von Übereinstimmungsrate zwischen Nutzern und Analysten bei der Kategorisierung der Erklärungsbedarfe.

RQ4.2: Inwieweit weicht die Einordnung in die Kategorien zwischen den Nutzern ab?

- **Goal 4.2:** Analyse der Homogenität in der Kategorisierung der Erklärungsbedarfe unter den Nutzern.

- **Question 4.2.1:** In welchem Ausmaß variieren die Kategorien, denen verschiedene Nutzer dieselben Erklärungsbedarfe zuordnen?

- **Metric 4.2.1.1:** Interrater-Reliabilität in Form von Fleiss' Kappa unter den Nutzern bei der Kategorisierung von Erklärungsbedarf.

- **Metric 4.2.1.2:** Interrater-Reliabilität in Form von Übereinstimmungsrate unter den Nutzern bei der Kategorisierung von Erklärungsbedarf.

4.3 Planung der verschiedenen Erhebungsmethoden

Für diese Arbeit werden drei unterschiedliche Erhebungsmethoden verglichen. Um Erklärungsbedarfe zu ermitteln, ist eine Referenz zu den jeweiligen Erklärungsbedarfsarten hilfreich [15]. Aus diesem Grund wird die in Abbildung 2.2 vorgestellte Taxonomie von Droste et al. [15] verwendet, da sie als eine Art Checkliste dient und dem Anforderungsspezialisten eine Leitlinie vorgibt, um die gewünschten Erklärungsbedarfe zu ermitteln [15].
Die Fokusgruppen und Interviews werden in drei Versionen durchgeführt. Die erste Version umfasst die Gruppe mit direkter Einführung der Taxonomie. Bei der zweiten Version wird die Erhebung ohne Taxonomie durchgeführt. In der dritten Version werden die Erklärungsbedarfe der Teilnehmer erst ohne Einführung Taxonomie und danach mit erhoben. Diese Version baut auf die zweite auf, weswegen es die selben Teilnehmer sind und diese Unterschiede zu Analysezwecken betrachtet werden.
Die effektivste Version wird letztlich für die Durchführung der quantitativen Studie, also der Onlineumfrage verwendet.

Fokusgruppen werden als Interviews mit Beteiligung mehrerer Teilnehmer betrachtet [33]. Aus diesem Grund werden in den folgenden Unterkapiteln die gleichen Abläufe für die Interviews und Fokusgruppen geplant. Dies ist auch erforderlich, um die beiden Erhebungsmethoden vergleichbar zu halten.
Auch bei der Onlineumfrage dient der gleiche Ablauf wie bei den Fokusgruppen und Interviews als Konzept für die Gestaltung, wobei die Erkenntnisse aus den qualitativen Erhebungsmethoden als Entscheidungsgrundlage für die Wahl der Version (mit direkter Einführung der Taxonomie / ohne Einführung der Taxonomie / mit nachträglicher Einführung der Taxonomie) der Onlineumfrage dienen.

4.3.1 Studienkonzept

Das Studienkonzept wurde mittels Flow-Methode [37] erstellt und ist in Abbildung 4.1 ersichtlich. Dabei wird zunächst der Ablaufplan vom Anforderungsanalysten erstellt. Dieser dient zur Durchführung der Erhebungsmethoden, wobei dort auch die Taxonomie zum Einsatz kommt. Aus den Fokusgruppen und Interviews entstehen kategorisierte Erklärungsbedarfslisten und Erkenntnisse, die das Gestaltungskonzept der Onlineumfrage optimieren sollen. Nachdem die Umfrage durchgeführt wurde, entsteht eine dritte kategorisierte Erklärungsbedarfsliste. Alle drei Erklärungsbedarfslisten werden zusammengefügt, analysiert und evaluiert. Erkenntnisse aus dieser Evaluation dienen dazu, die Forschungsfragen zu beantworten. Außerdem entsteht eine codierte Gesamtliste an Erklärungsbedarfen und eine potenzielle erweiterte Taxonomie.

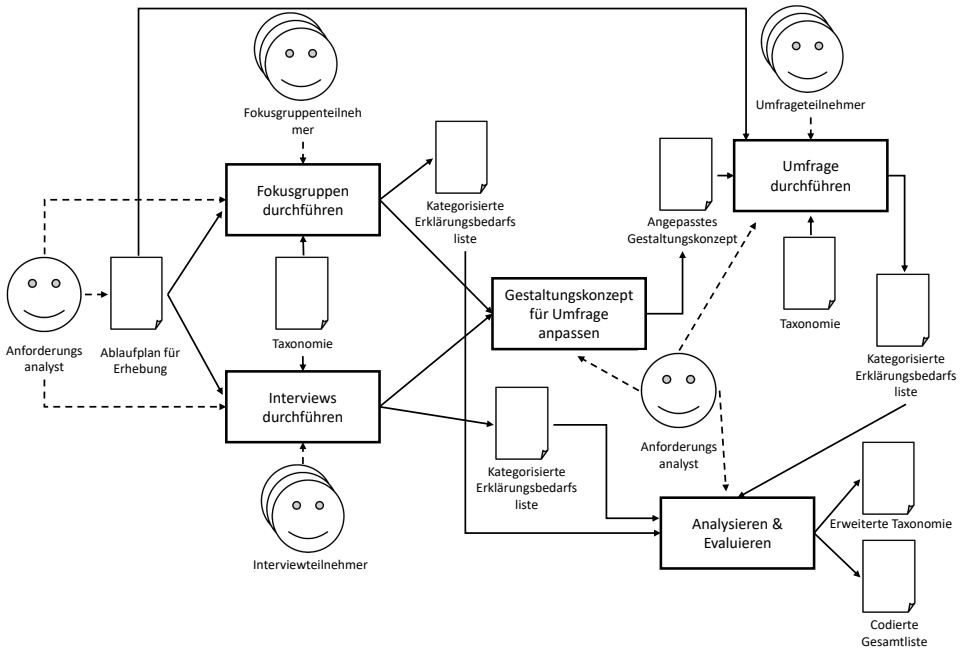

Abbildung 4.1: Studienkonzept in Flow-Methode

4.3.2 Erhebungsmethoden mit direkter Einführung der Taxonomie

Die erste Version der Erhebungsmethode ist die mit direkter Einführung der Taxonomie aus Abbildung 2.2. Der Ablaufplan für diese Version ist in Abbildung 4.2 einsehbar.
Detailliert betrachtet umfasst der Ablauf für die Erhebungsmethoden mit

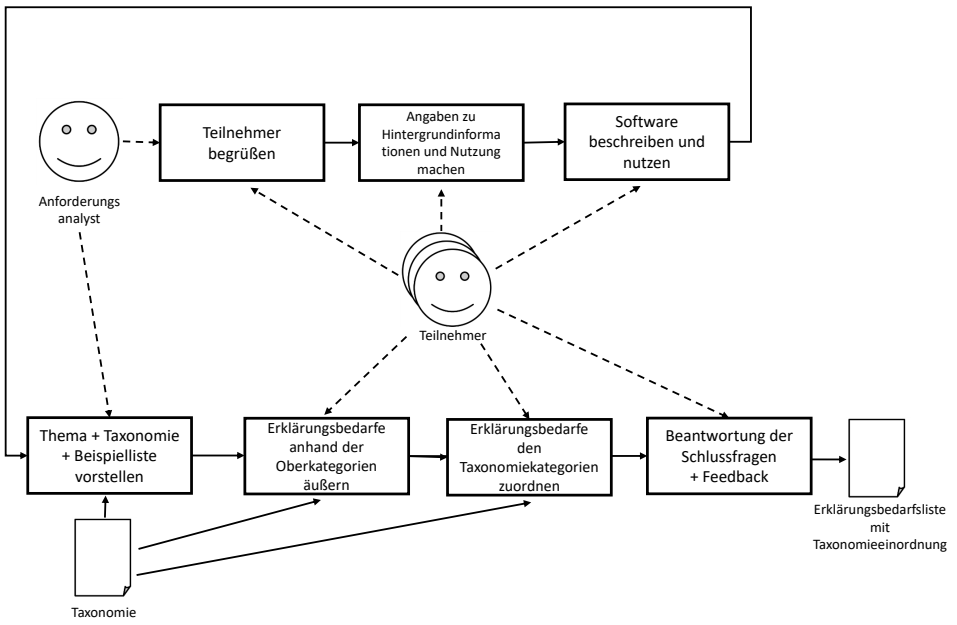

Abbildung 4.2: Ablauf der Erhebungsmethoden mit direkter Einführung der Taxonomie in FLOW-Methode

direkter Einführung der Taxonomie die in Tabelle 4.1 aufgeführten Schritte:

1. Einleitung

Zunächst werden die bzw. wird der Teilnehmer begrüßt, wobei eine kurze Vorstellung stattfindet.

Der Zweck der Fokusgruppe wird erklärt und es wird eine kurze Einleitung in das Thema dieser Arbeit geboten. Dabei wird darauf eingegangen, dass der Zweck dieser Erhebung ist, Erklärungsbedarfe zu sammeln und diese dann dazu dienen können, die Software zu optimieren.

Die Dauer des Interviews wird mit 45 Minuten angegeben, die der Fokusgruppe mit 45 bis 60 Minuten.

Anschließend wird das Einverständnis zur Aufzeichnung der Erhebung eingeholt und die Aufzeichnung gestartet.

2. Hintergrundinformationen

Die Hintergrundinformationen der jeweiligen Teilnehmer werden aufgenommen. Das umfasst ihre Rolle in dem Unternehmen, ihr Geschlecht und ihr Alter.

3. Allgemeine Nutzungsangaben

Allgemeine Nutzungsangaben werden dokumentiert. Dabei wird abgefragt, wie lange sie adpeople bereits nutzen und wie häufig sie dies tun in folgenden Angaben: täglich / wöchentlich / monatlich / jährlich.

Schritt	**Details**	**Vsl. Dauer**
1. Einleitung	- Begrüßung + Vorstellung - Zweck + Dauer der Erhebung	2 min
2. Hintergrundinformationen	- Rolle - Geschlecht + Alter	2 min
3. Allgemeine Nutzungsangaben	- Dauer der Nutzung - Häufigkeit der Nutzung	2 min
4. Softwarebeschreibung und -nutzung	- Software vorstellen lassen - Software 3 min. nutzen lassen	5 min
5. Vorstellung Erklärbarkeit	- Kurze Einführung in das Thema - Einführung in Erklärungsbedarf	5 min
6. Vorstellung Taxonomie mit Beispielen	- Durchgehen der Taxonomie + zugehörige Beispielliste	4 min
7. Erklärungsbedarfsermittlung	- Sammeln der Bedarfe mit Hilfe der Taxonomie	15 min
8. Kategorisierung der Bedarfe	- Kategorisierung der genannten Bedarfe mittels Taxonomie	7 min
9. Schlussfragen	- Beendigung der Erhebung - Einholen von Feedback	3 min

Tabelle 4.1: Übersicht der Schritte der Methoden mit Details und Zeitangaben bei der Version mit direkter Einführung der Taxonomie

4. Softwarebeschreibung und -nutzung
Daraufhin werden die Teilnehmer gebeten, adpeople mit ihren wichtigsten Funktionen in eigenen Worten kurz zu beschreiben. Anschließend bekommen sie drei Minuten, adpeople aufzurufen und zu nutzen.

5. Vorstellung Erklärbarkeit und Erklärungsbedarf Den Teilnehmern wird das Oberthema der Arbeit erklärt, sprich Erklärbarkeit. Dabei wird auch auf das Thema Erklärungsbedarf eingegangen.

6. Vorstellung der Taxonomie mit Beispielen
Die Taxonomie aus Abbildung 2.2 wird mitsamt der darunter beschriebenen Beispielen zu den Unterkategorien vorgestellt. Es wird sich erkundigt, ob Fragen dazu bestehen.

7. Erklärungsbedarfe ermitteln
Die Taxonomie wird entlang der Oberkategorien durchgegangen und die Erklärungsbedarfe der Teilnehmer werden ermittelt.

8. Kategorisierung der Erklärungsbedarfe
Die Teilnehmer werden aufgefordert, ihren zuvor genannten Erklärungsbedarfen eine passende Unterkategorie bzw. Oberkategorie, wenn zu dieser keine Unterkategorien existieren, zuzuweisen.

9. Schlussfragen
Zum Schluss werden folgende Fragen gestellt:

- Hast du noch Vorschläge, wie adpeople leichter verständlich gemacht werden könnte?
- Welche zusätzlichen Hilfsmittel oder Erklärungen würdest du dir wünschen?
- Gibt es noch etwas, das du über deine Erfahrungen mit adpeople mitteilen möchtest?
- Auf einer Skala von 1 bis 5, wobei 1 „gar nicht" repräsentiert und 5 „sehr" repräsentiert, wie sehr willst du, dass deine geäußerten Erklärungsbedarfe auch von den Entwicklern angegangen werden?
- Fallen dir noch weitere Kategorien für die Taxonomie ein?
- Auf einer Skala von 1 bis 5, wobei 1 „nicht sicher" und 5 „sehr sicher" repräsentiert, wie sicher warst du dir bei der Vergabe der Kategorien?
- Feedback zur Umfrage: Das Ziel war es, Erklärungsbedarfe zu ermitteln - findest du, dass diese Erhebung das Ziel erreicht hat?
- Wie fandest du die Struktur des Ablaufs dieser Erhebung?

4.3.3 Erhebungsmethoden mit nachträglicher Einführung der Taxonomie

Die zweite und dritte Version der Erhebungsmethoden ist die mit nachträglicher Einführung der Taxonomie. Der Ablaufplan für diese Version ist in Abbildung 4.3 einsehbar.
Tabellarisch zusammengefasst sind die Unterschiede zur ersten Version fett markiert und in Tabelle 4.2 einsehbar.
Auf diese Unterschiede wird hier detailliert eingegangen:
1. Einleitung
2. Hintergrundinformationen
3. Allgemeine Nutzung
4. Softwarebeschreibung und -nutzung
5. Vorstellung Erklärbarkeit und Erklärungsbedarf
Hier werden zusätzlich zu den Themen noch drei Beispiele für Erklärungsbedarf genannt, damit die Teilnehmer ohne Taxonomie bereits eine Vorstellung darüber haben, wie Erklärungsbedarf aussieht. Diese Beispiele sind Teil der Taxonomie und wurden zufällig gewählt. Sie werden hier ohne Kategorie gelistet:

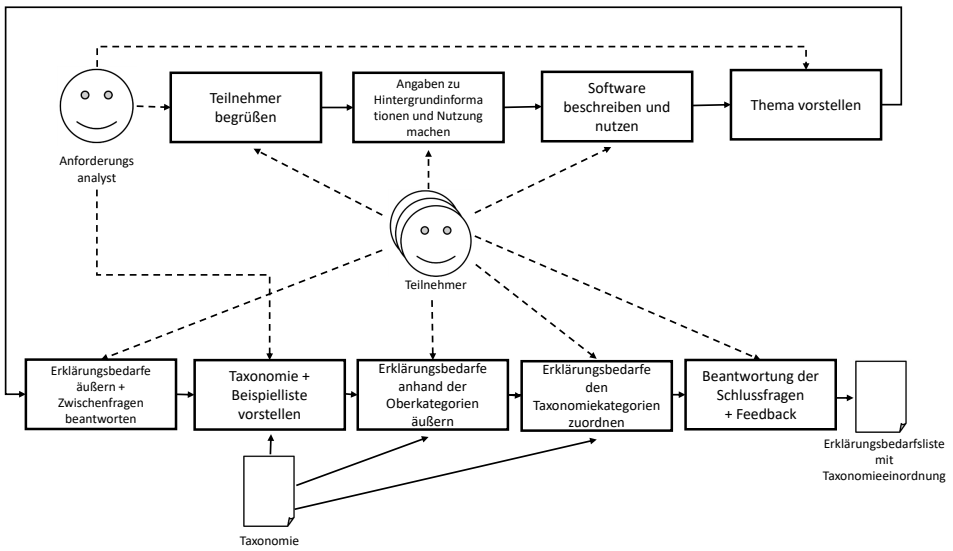

Abbildung 4.3: Ablauf der Erhebungsmethoden mit nachträglicher Einführung der Taxonomie in FLOW-Methode

- „Wieso kann ich mich nicht einloggen?"
- „Warum ist die App teurer geworden?"
- „Warum ist die Suchleiste unten?"

6. Erklärungsbedarfe ermitteln ohne Taxonomie
7. Zwischenfragen
Hier werden vier Zwischenfragen gestellt, die ebenfalls in den Schlussfragen vorkommen. Dies ist dafür da, um zu schauen, ob sich durch die Einführung der Taxonomie etwas geändert hat:

- Hast du noch Vorschläge, wie die Software leichter verständlich gemacht werden könnte?
- Welche zusätzlichen Hilfsmittel oder Erklärungen würdest du dir wünschen?
- Gibt es noch etwas, das du über deine Erfahrungen mit der Software mitteilen möchtest?
- Auf einer Skala von 1 bis 5, wobei 1 „gar nicht" repräsentiert und 5 „sehr" repräsentiert, wie sehr willst du, dass deine geäußerten Erklärungsbedarfe auch von den Entwicklern angegangen werden?

8. Vorstellung der Taxonomie mit Beispielen
9. Erklärungsbedarfe ermitteln

Schritt	**Details**	**Vsl. Dauer**
1. Einleitung	- Begrüßung + Vorstellung - Zweck + Dauer der Erhebung	2 min
2. Hintergrundinformationen	- Rolle - Geschlecht + Alter	2 min
3. Allgemeine Nutzungsangaben	- Dauer der Nutzung - Häufigkeit der Nutzung	2 min
4. Softwarebeschreibung und -nutzung	- Software vorstellen lassen - Software 3 min. nutzen lassen	5 min
5. Vorstellung Erklärbarkeit	- Kurze Einführung in das Thema - Einführung in Erklärungsbedarf	5 min
6. Erklärungsbedarfsermittlung ohne Taxonomie	**- Sammeln der Bedarfe**	**10 min**
7. Zwischenfragen	**- Beantwortung der Zwischenfragen**	**3 min**
8. Vorstellung Taxonomie mit Beispielen	- Durchgehen der Taxonomie + zugehörige Beispielliste	4 min
9. Erklärungsbedarfsermittlung	- Sammeln der Bedarfe mit Hilfe der Taxonomie	**5 min (-10)**
10. Kategorisierung der Bedarfe	- Kategorisierung der genannten Bedarfe mittels Taxonomie	7 min
11. Schlussfragen	**- Beendigung der Erhebung mit einer Zusatzfrage**	3 min

Tabelle 4.2: Übersicht der Schritte der Methoden mit Details und Zeitangaben bei der Version mit nachträglicher Einführung der Taxonomie mit Markierung der Änderungen in Bezug auf die Version mit direkter Taxonomieeinführung

10. Kategorisierung der Erklärungsbedarfe
Hier sollen nun alle genannten Erklärungsbedarfe einer Kategorie zugeordnet werden.
11. Schlussfragen
Zusätzlich zu den Schlussfragen der Erhebungsmethode mit direkter Einführung der Taxonomie kommt folgende Frage hinzu:

- Auf einer Skala von 1 bis 5, wobei 1 „gar nicht" und 5 „sehr" repräsentiert, inwiefern hat dir die Taxonomie geholfen, weitere Bedarfe zu finden?

Kapitel 5

Durchführung der Erhebungsmethoden

Im Folgenden werden die Abläufe der einzelnen Erhebungsmethoden vorgestellt. Auf Erkenntnisse und Anpassungen der Methoden wird ebenfalls eingegangen. Die Teilnehmer der Erhebungen sind Mitarbeiter des Beratungs- und IT-Dienstleistungsunternehmens adesso SE[1] mit einer Mitarbeiterzahl von über 10.200 (Stand 2023), das in mehreren Standorten, unter anderem in Hannover und Dortmund, vertreten ist. Die Nutzer der Personalverwaltungssoftware übernehmen vielfältige Rollen, die von Softwareentwicklern über IT-Berater und -Vertriebsmitarbeiter bis hin zu Projektleitern, HR-Managern und Führungskräften aus den entsprechenden Fachbereichen, reichen.

5.1 Fokusgruppen

Für die erste Fokusgruppe in Hannover wurde innerhalb der firmeninternen Email-Verteilerliste für diesen Standort akquiriert. Die voraussichtliche Dauer wurde mit 60 Minuten angegeben, wobei sich an die ungefähren Zeiten aus Tabelle 4.1 orientiert und eine Pufferzeit von 15 Minuten aufgeschlagen wurde. Es wurden nur vier Teilnehmer für den ersten und zwei für den zweiten Termin gefunden, weswegen sich dafür entschieden wurde, nur die erste Fokusgruppe stattfinden zu lassen. Um weitere Teilnehmer zu finden, wurden zwei Mitarbeiter im direkten Gespräch akquiriert. Somit wurde die erste Fokusgruppe mit sechs Mitarbeitern des Unternehmens durchgeführt. Die Dauer betrug 73 Minuten. Angewandt wurde dabei die erste Version, mit direkter Einführung der Taxonomie.
Für die zweite Fokusgruppe wurden Teilnehmer an einem anderen Standort mit mehr potenziellen Teilnehmern gesucht. Dafür kam Dortmund in Frage. Dort konnten dann ebenfalls sechs Mitarbeiter gefunden werden. Bei dieser

[1] `https://adesso.de/`

Fokusgruppe betrug die Dauer 71 Minuten und es wurde die Version mit nachträglicher Einführung der Taxonomie angewandt.
Beide Fokusgruppen wurden am selben Wochentag, einem Dienstag und zur gleichen Tageszeit, 13 Uhr nach der Mittagspause, durchgeführt.
Bei der Durchführung der Fokusgruppen ist aufgefallen, dass die Teilnehmer auffällig oft die Kategorien *Nicht vorhandenes Feature* und *Einführung* vergeben haben.
Ein Beispiel dient bei folgendem Erklärungsbedarf:
„Wie kann ich ToDo tasks erstellen?"
Drei Rater haben diesen Bedarf als *Nicht vorhandenes Feature*, zwei als *Einführung* kategorisiert. Eine Person hat es korrekt als *Operation* interpretiert, da es eine Frage zu einer Funktionsweise ist.

Eine Erkenntnis ist, dass für die Interviews genauer erklärt werden muss, dass die Teilnehmer ihre Erklärungsbedarfe präzise äußern sollen. Dies soll verhindern, dass der Erklärungsbedarf potenziell falsch gelöst wird, indem der Entwickler eine falsche Erklärung liefern würde, wenn der Erklärungsbedarf in einer uneindeutigen Form geäußert wird. Auch bei der Kategorisierung hat dies Auswirkungen, da somit möglicherweise eine genauere Vergabe stattfindet. Wenn die Teilnehmer den Bedarf haben zu erfahren, wieso ein bestimmtes Feature nicht vorhanden ist, sollen sie es auch in der Form äußern. Wünschen sich die Teilnehmer eine Einführung, so soll der Bedarf das auch so ausdrücken: „Ich benötige eine Anleitung, um ...". Dies wird beim Erläutern der Beispielliste besonders betont.
Bei der ersten Fokusgruppe haben einige Teilnehmer bei manchen Bedarfen mehr als eine Kategorie vergeben, weil sie sich nicht entscheiden konnten. Sie wurden daraufhin gebeten, sich für eine Kategorie zu entscheiden. Dieser Hinweis wird auch für die Interviews relevant sein.
Da die Teilnehmer zu viel subjektive Interpretationen bei der Kategorievergabe einfließen lassen habe, wurde ihnen gesagt, die Kategorien getrennt von ihren Wünschen zu vergeben und die Bedarfe so zu betrachten, wie sie geäußert wurden.
Die Teilnehmer wurden darauf hingewiesen, darauf noch mal zu achten und ihre Kategorien dahingehend anzupassen.

Außerdem wurden die Teilnehmer bei der Fokusgruppe mit direkter Einführung der Taxonomie, wie in dem Ablaufplan in Abbildung 4.2 dargestellt, nach ihren Erklärungsbedarfen anhand der Oberkategorien befragt. Dabei hatten sie jedoch bereits Erklärungsbedarf, den sie äußern wollten, aber noch nicht konkret wussten, welchen Oberkategorien sie ihm zuordnen sollten. Dementsprechend hatten die Teilnehmer zuerst ihre Erklärungsbedarfe genannt, bevor sie speziell nach den Oberkategorien befragt wurden. Dieses Vorgehen hilft auch bei den Interviews. Damit wird auch das Risiko verhindert, dass sie ihre Erklärungsbedarfe vergessen, wenn

sie gezwungen sind, erst nach den Oberkategorien zu gehen.
Bei der zweiten Fokusgruppe, also der Version mit nachträglicher Einführung der Taxonomie, wurde vor dem Beginn der Kategorisierungsphase erwähnt, dass nur eine Kategorie vergeben werden darf, um den redundanten Aufwand zu verhindern.

5.2 Interviews

Für die Durchführung der Interviews wurden zunächst Teilnehmer akquiriert, wobei es zwei Wochen dauerte, um die benötigten 18 Teilnehmer zu finden. Dabei wurden alle Mitarbeiter, die die Software nutzen per Email-Verteiler im BCC mit einer persönlichen Nachricht angeschrieben. Eine voraussichtliche Dauer wurde mit 45 Minuten angegeben, resultierend aus den geschätzten Zeiten aus Tabelle 4.1. Die Interviews wurden alle online durchgeführt. Dabei wurde bei der Hälfte der Teilnehmer, also in neun Interviews, die Version mit und bei den anderen neun Interviews die Version mit nachträglicher Einführung der Taxonomie angewandt.

Aufgrund der Erkenntnisse aus den Fokusgruppen in Unterkapitel 5.1, dass die Teilnehmer auffällig oft die Kategorien *Nicht vorhandenes Feature* und *Einführung* vergeben haben, wurde die Taxonomieübersicht aus Abbildung 2.2 angepasst. Die Reihenfolge der Kategorien wurde angepasst, um dem gehäuften Vergeben von *Nicht vorhandenes Feature* entgegenzuwirken, da es zuvor an der letzten Position möglicherweise zu prägnant war und deswegen oft in den Fokusgruppen vergeben wurde. Somit rutschte es direkt nach *Interaktion*. Die Reihenfolge der Beispiele wurde dementsprechend angepasst.

Um dem häufig inkorrekten Vergeben von *Einführung* entgegenzuwirken, wurde darauf hingewiesen, dass ein Erklärungsbedarf, der zu dieser Kategorie gehört, den Wunsch einer Einführung beinhalten muss.
Bei der Erklärungsbedarfsermittlungsphase innerhalb der Interviews wurde die Erkenntnis aus den Fokusgruppen in Unterkapitel 5.1 berücksichtigt, dass nach dem Durchgehen der Taxonomie die Teilnehmer bereits Erklärungsbedarf hatten, den sie noch nicht spezifisch einer Oberkategorie zuordnen konnten. Deswegen sollten sie nach der Einführung der Taxonomie zu erst ihre Bedarfe äußern, bevor anhand der Oberkategorien der Taxonomie Bedarfe ermittelt wurden.
Zu Beginn der Kategorisierungsphase wurde bei den Interviews ebenfalls erwähnt, dass nur eine Kategorie vergeben werden soll.

5.3 Onlineumfrage

Aufgrund erster Auswertungen der verschiedenen Versionen der Durchführungen der Erhebungsmethoden, wurde sich in dieser Arbeit dafür entschieden, die Umfrage in der Version, in der man die Taxonomie nachträglich einführt, durchzuführen. Ausschlaggebend für diese Entscheidung ist die erhöhte durchschnittlichen Anzahl an Bedarfen pro Person aus Tabelle 6.1. Für die Gestaltung der Onlineumfrage wurde LimeSurvey[2] des Instituts Software Engineering der Leibniz Universität Hannover verwendet.

Zur Teilnehmerakquise wurde zunächst seitens des Managements der Software alle Mitarbeiter von adesso per Email-Verteiler angeschrieben. Nach sechs Tagen war die Anzahl an Teilnehmern bei 50, wobei nur neun Teilnehmer die Umfrage vollständig abgeschlossen hatten. Aus diesem Grund wurden die Teilnehmer im BCC noch mal persönlich von dem Anforderungsanalysten mit der Notwendigkeit der Studie angesprochen. Dies führte nach weiteren acht Tagen zu insgesamt 895 Teilnahmen, wovon 277 vollständig waren. Nach Filterung von Teilnahmen, bei denen sowohl vor als auch nach Einführung der Taxonomie keine Erklärungsbedarfe eingegeben wurden, kamen 188 valide Teilnahmen zusammen. Dabei wurden auch solche, die Äußerungen wie „ich habe keine Bedarfe" oder „???????" getätigt haben, rausgefiltert. Erkenntnisse aus den Fokusgruppen und Interviews wurden hier ebenfalls umgesetzt:
Es wird darauf hingewiesen, dass die Teilnehmer ihre Erklärungsbedarfe so äußern sollen, dass ein Entwickler sie möglichst geeignet umsetzen bzw. beantworten kann. Der Teilnehmer soll seinen Bedarf also so äußern, dass die mögliche Antwort zu der Frage auch die Information ist, die er wissen will.

Die Analysen über die einzelnen Interrater-Reliabilitäten bei der Vergabe der Taxonomiekategorien in Tabelle 6.7 haben gezeigt, dass besonders die Kategorien *Unerwartetes Systemverhalten, Algorithmus, Operation, Einführung* und *Nicht vorhandenes Feature* falsch interpretiert wurden.
Aus diesen Gründen wurden mit der Einführung der Taxonomie und der Beispielliste auch Negativbeispiele genannt:

Einführung

- **Negativbeispiel:** „Wofür ist die Funktion xyz da?"
  Passende Kategorie: *Operation*
  Begründung: Es handelt sich um eine Frage zur Funktion selbst, nicht um eine Einführung oder ein Tutorial. Möchte man ein Tutorial, müsste man dies explizit äußern.

[2] `https://survey.se.uni-hannover.de/`

Nicht vorhandenes Feature

- **Negativbeispiel:** „Wieso kann ich keine xyz erstellen?“
  Passende Kategorie: *Operation*
  Begründung: Diese Aussage beschreibt ein Problem bei der Durchführung einer Funktion. Möchte man eine Einführung, müsste dies explizit formuliert werden.

Unerwartetes Systemverhalten

- **Negativbeispiel:** „Warum gibt mir die Funktion x statt y aus?“
  Passende Kategorie: *Algorithmus*
  Begründung: Es handelt sich um das Ergebnis einer Funktion und ist daher eher der Kategorie Algorithmus zuzuordnen als unerwartetem Systemverhalten.

Zugleich wurde darauf hingewiesen, dass Bedarfe die der Kategorie *Einführung* angehören, bestimmter expliziter Signalwörter bedürfen, wie „ich brauche eine Einführung / ein Tutorial / ein Guide”.
Gehören Erklärungsbedarfe der Kategorie *Nicht vorhandenes Feature* an, so müssen sie Signalwörter wie „Ich brauche eine Erklärung, wieso es das Feature xyz nicht gibt...” oder „Wieso gibt es die Funktion xyz nicht?...” etc. beinhalten.

Bei der Kategorisierung wurde darauf hingewiesen, dass die Teilnehmer darauf achten sollen, dass sie den Bedarf so kategorisieren, dass nur das explizit genannte in dem von ihnen formulierten Bedarf drin steht, getrennt von ihren unterbewussten Wünschen.
Zusätzlich wurde die Schlussfrage „Würdest du eine (Unter)-kategorie aus der Taxonomie entfernen und wenn ja, welche und warum?” gestellt, da in den zuvor durchgeführten Erhebungen manche Kategorien wie Business oder Konsequenzen fast nie genannt wurden.

Die Gestaltung der Onlineumfrage, die in der Version der nachträglichen Taxonomieeinführung durchgeführt wurde, sah somit wie folgt aus:

- Begrüßung des Teilnehmers mit Angabe des Zwecks der Erhebung und einer voraussichtlichen Dauer von 15-20 Minuten.
- „Was ist deine Rolle bei adesso?”
- „Bist du als Führungskraft oder im HR-Bereich tätig?”
- „Bitte gebe dein Alter + Geschlecht (m/w/d) an:”
- „Seit wann nutzt du adpeople?”

- „Wie häufig nutzt du adpeople?"
  - Angabemöglichkeiten: täglich / wöchentlich / monatlich / jährlich
- „Bitte beschreibe kurz in wenigen Stichpunkten, was adpeople ist und welche (Haupt-)Funktionen vorhanden sind."
  - Beispielantworten: Abwesenheiten, Personalstammdaten, Zielvereinbarungen und Verträge
- „Bitte rufe nun adpeople auf und nutze die Software ca. 3 Minuten lang. Klicke dich durch verschiedene Funktionen der Seite und kehre zur Umfrage zurück und drücke auf „Weiter", wenn du so weit bist."
- Vorstellung Erklärbarkeit / Erklärungsbedarf
- Beispiele für Erklärungsbedarfe
  - „Wieso kann ich mich nicht einloggen?"
  - „Warum ist die App teurer geworden?"
  - „Warum ist die Suchleiste unten?"
- „Gebe hier nun deine Erklärungsbedarfe zu adpeople ein. Die Textfelder werden automatisch hinzugefügt und somit erweitert. Bedenke, dass es sich auf alles Mögliche zur Software beziehen kann. Schreibe deine Erklärungsbedarfe möglichst präzise auf, also optimalerweise so, dass ein potenzieller Entwickler sie möglichst gut oder für dich zufriedenstellend beantworten kann, ohne interpretieren zu müssen, was du eigentlich wissen willst. Das heißt, formuliere die Frage so, dass die Antwort zu der Frage auch die Information ist, die du wissen willst."
  - Beispielantwort: „Warum ist der Speicher-Button klein und ausgegraut?"
  - Beispielantwort: „Wieso wird der Urlaub immer erst fast ein halbes Jahr zu spät dem Urlaubskonto gutgeschrieben?"
- Zwischenfragen aus Kapitel 4.3.3.
- Vorstellung der Taxonomie mit Aufführung der besagten Beispielliste samt Negativbeispielen.
- „Fällt dir noch Erklärungsbedarf zu Systemverhalten, Interaktion, Nicht vorhandenes Feature, Domänenwissen, Business, Geheimhaltung & Sicherheit oder Benutzeroberfläche ein?"

- „Hier sind deine genannten Erklärungsbedarfe aufgelistet. Bitte weise ihnen eine Unterkategorie der vorhin aufgestellten Taxonomie zu. Achte dabei darauf, dass du nur das kategorisierst, was explizit in dem von dir formulierten Bedarf drin steht, getrennt von deinen Wünschen. Klicke hier, um die Taxonomie aufzurufen, falls du sie geschlossen hast."
- „Bitte beantworte noch die letzten Schlussfragen:"
  - „Fallen dir noch weitere Kategorien für die Taxonomie ein?"
  - „Auf einer Skala von 1 bis 5, wobei 1 „nicht sicher" und 5 „sehr sicher" repräsentiert, wie sicher warst du dir bei der Vergabe der Kategorien?"
  - „Auf einer Skala von 1 bis 5, wobei 1 „gar nicht" und 5 „sehr" repräsentiert, inwiefern hat dir die Taxonomie geholfen, weitere Bedarfe zu finden?"
  - „Würdest du eine (Unter)-kategorie aus der Taxonomie entfernen und wenn ja, welche und warum?"
  - „Feedback zur Umfrage: Das Ziel war es, Erklärungsbedarfe zu ermitteln - findest du, dass diese Umfrage das Ziel erreicht hat? Bitte mit Begründung."
  - „Wie fandest du die Struktur des Ablaufs dieser Umfrage?"

Kapitel 6

Ergebnisse

In diesem Kapitel werden die Ergebnisse der gesamten Studie vorgestellt. Zunächst wird ein Überblick über die statistischen Ergebnisse geboten, bevor auf die Ergebnisse der einzelnen Erhebungsmethoden eingegangen wird.

6.1 Überblick über die statistischen Ergebnisse der Erhebungsmethoden

Um einen Überblick über die statistischen Ergebnisse der Erhebungsmethoden zu liefern, werden diese in Tabelle 6.1 dargestellt. Zur Identifizierung der einzigartigen Bedarfe wurde das Codingsystem nach Nikollau [29] verwendet. In dieser Tabelle ist ersichtlich, dass bei der Fokusgruppe mit direkter Taxonomieeinführung 27 Erklärungsbedarfe, von denen 23 einzigartig sind, bei der ohne Einführung 20 (19 einzigartige) und bei der Fokusgruppe mit nachträglicher Taxonomieeinführung 28 (27 einzigartige) Bedarfe von jeweils sechs Teilnehmern erhoben wurden. Somit ist ersichtlich, dass in der Version der direkten Einführung vier und in den anderen beiden jeweils ein Bedarf doppelt vorkam.

Bei der Durchführung der neun Interviews mit direkter Taxonomieeinführung kamen 109 (105 einzigartige) Bedarfe zu Stande. Bei der Version ohne Taxonomieeinführung 103 mit 96 einzigartigen Bedarfe. Die Version mit nachträglicher Taxonomieeinführung kommt hier ebenfalls auf die meisten Erklärungsbedarfe mit 147 (133 einzigartige).
Die Umfrage, an der 188 Mitarbeiter teilnahmen, kam in der Version ohne Taxonomieeinführung auf 409 (327 einzigartige) Erklärungsbedarf, bei der Version mit nachträglicher Taxonomieeinführung auf 471 (364 einzigartige) Erklärungsbedarfe.

Insgesamt zeigt sich, dass die Version ohne Taxonomieeinführung 17,39% weniger einzigartige Bedarfe bei der Fokusgruppe und 8,57%

	Fokusgruppe			Interviews			Umfrage	
Taxonomie-einführung	direkt	ohne	nachtrgl.	direkt	ohne	nachtrgl.	ohne	nachtrgl.
n	6	6	6	9	9	9	188	188
#Bedarfe	27	20	28	109	103	147	409	**471**
Personal-aufwand [h]	6,34	3,22	6,07	10,35	7,56	11,58	35,41	**39,01**
ø t_gesamt [min]	**54,30**	27,60	52,00	33,99	23,47	38,21	11,40	14,14
Median t_gesamt [min]	**54,30**	27,60	52,00	34,5	25,20	38,60	11,30	12,45
ø t Einführung + Ermittlung [min]	18,80	12,50	**25,90**	20,30	10,66	20,33	4,10	6,83
Median t Einführung + Ermittlung [min]	18,80	12,50	**25,90**	20,60	9,70	20,77	2,42	3,57
einzigartige Bedarfe	23	19	27	105	96	133	327	**364**
ø einzigartige Bedarfe p.P.	3,83	3,17	4,50	11,67	10,67	**14,78**	1,74	1,94
einzigartige Bedarfe p.P. p. ø t [min]	0,20	0,25	0,17	0,57	**1,10**	0,71	0,42	0,28
einzigartige Bedarfe p.P. p. Median t [min]	0,20	0,25	0,17	0,57	**1,10**	0,71	0,72	0,54
einzigartige Bedarfe pro Personal-aufwand [h]	3,63	5,90	4,45	10,14	**12,70**	11,49	9,24	9,33

Tabelle 6.1: Statistischer Vergleich der Erhebungsmethoden mit Unterscheidung der Versionen der Taxonomieeinführung und mit Markierung der höchsten Werte pro betrachtetem Aspekt, wobei n für die Anzahl der Teilnehmer steht

weniger bei den Interviews erzielt als bei der direkten Einführung. Die nachträgliche Einführung der Taxonomie führte zu den meisten einzigartigen Erklärungsbedarfen. Konkret erzielte sie in der Fokusgruppe 27 solcher Bedarfe, was rund 17,39% mehr sind als bei der direkten Einführung. In den Interviews zeigte sich ein ähnliches Bild: Mit 133 einzigartigen

Bedarfen war dies etwa 26,67 % mehr als bei der direkten Einführung. Bei der Umfrage resultierte die nachträgliche Einführung in 364 einzigartige Erklärungsbedarfe, was einer Steigerung von 11,31 % entspricht.
Ein wichtiges Maß für die Effizienz einer Methode ist hierbei die Angabe der einzigartigen Bedarfe pro Person pro Median Zeit in Minuten, wobei für die Berechnung der Zeit nur die reine Bedarfsermittlungszeit und die Zeit für die Einführung der Taxonomie bei den Versionen, die eine Taxonomieeinführung beinhalten, betrachtet wird. Hierbei hat die Fokusgruppenversion mit direkter Taxonomieeinführung 0,20 einzigartige Bedarfe pro Person pro Minute, bei der ohne Taxonomieeinführung rund 25% mehr mit 0,25 und bei der Fokusgruppe mit nachträglicher Taxonomieeinführung am wenigsten mit 0,17 einzigartige Bedarfe pro Person pro Minute.

Bei Betrachtung dieser Werte bei den anderen Erhebungsmethoden fällt auf, dass dies die niedrigsten sind. Bei den Interviews hat die Version mit direkter Taxonomieeinführung 0,57 einzigartige Bedarfe pro Person pro Minute, bei der ohne Taxonomieeinführung fast doppelt so viele (1,1) und bei den Interviews mit nachträglicher Taxonomieeinführung 0,71 einzigartige Bedarfe pro Person pro Minute.
Bei der Umfrage ohne Taxonomieeinführung kommt dieser Wert auf 0,72. Bei der Version mit nachträglicher Taxonomieeinführung ist es ein kleinerer Wert (0,54). Hierbei muss man beachten, dass die Erhebungs- und Einführungsdauer mit 2,42 Minuten bzw. 3,57 Minuten im Median deutlich geringer ausfällt als bei den anderen Erhebungsmethoden. Deutlich stechen die Interviews in der Version ohne Taxonomieeinführung aus, wobei der Wert für die nachträgliche Taxonomieeinführung mit 0,71 einzigartigen Bedarfen pro Person pro Minute der zweithöchste aus allen Erhebungen ist.

Der Personalaufwand besteht aus der Anzahl der Teilnehmer, wobei da jeweils der Anforderungsanalyst bei den Fokusgruppen und Interviews mit dazu gerechnet wird, multipliziert mit der Gesamtzeit. Die Gesamtzeit berechnet sich dabei mithilfe der Durchführungszeit der einzelnen Erhebungen abzüglich der Zeit für das Kategorisieren der Bedarfe, der Schluss- und der Zwischenfragen. Aus der Anzahl der einzigartigen Erklärungsbedarfe geteilt durch den Personalaufwand resultieren die niedrigsten Werte bei der Fokusgruppe mit 3,63 einzigartigen Bedarfen pro stündlichem Personalaufwand bei der Version mit direkter Taxonomieeinführung, 5,90 bei der ohne und 4,45 bei der mit nachträglicher Taxonomieeinführung.
Die Interviews haben dort die höheren Werte erzielt mit 10,14 bei den Interviews mit direkter Taxonomieeinführung, 12,70 bei der Version ohne und 11,49 einzigartige Bedarfen pro stündlichem Personalaufwand bei der Version mit nachträglicher Taxonomieeinführung.
Bei der Umfrage sind es 9,24 einzigartige Bedarfe pro stündlichem

Personalaufwand bei der Version ohne Taxonomieeinführung und leicht höher bei nachträglicher Taxonomieeinführung mit 9,33.

Die Zeiten, die für die Einführung der Taxonomie benötigt wurden, ist mit 4,4 Minuten bei der Fokusgruppe am höchsten, mit 3,14 Minuten bei den Interviews und 0,69 Minuten im Median am niedrigsten.

Bezüglich der Kategorisierung konnten alle Erklärungsbedarfe sowohl von dem Anforderungsanalysten als auch von den Nutzern abgedeckt werden. Bei der Frage nach der Sicherheit bei der Kategorievergabe, gaben die Fokusgruppenteilnehmer einen durchschnittlichen Wert von 3,33 auf einer Likert-Skala von 1-5, wobei 1 für „nicht sicher" und 5 für „sehr sicher" steht, an. Bei den Interviews ist der durchschnittliche Wert bei 3,69 und bei der Umfrage 3,15. Dabei wurden die Nutzer auch nach weiteren Kategorien gefragt.
In den Fokusgruppen wurden keine Vorschläge dazu geäußert. Bei den Interviews und der Umfrage kamen die in Tabelle 6.2 einsehbaren Antworten auf die Frage, ob dem Teilnehmer noch weitere Kategorien für die Taxonomie einfallen.

Auf die Frage, ob die Teilnehmer das Gefühl hatten, dass die Umfrage das Ziel, also die Ermittlung von Erklärungsbedarfe, erreicht hat, kam bei den Fokusgruppen und Interviews nur positives Feedback. Bei der Umfrage war dies gemischt, denn viele gaben an, dass die Form der Umfrage zu wissenschaftlich sei: „Leider nein, die Umfrage ist zu lang und nicht alle Leute können das verstehen (zu technisch und zu viel)". Einige Teilnehmer merkten aber auch an, dass ihnen die Taxonomie bei der Formulierung von ihren Erklärungsbedarfen geholfen hat: „Ja, das hat geholfen anders darüber nachzudenken. Zu Beginn war das ganze so offen, dass es mir schwer fiel, Themen zu finden. Ich habe das erst duch Aufrufen der Seite hinbekommen. Ich denke, mithilfe der Taxonomie kann man da besser formulierte Bedarfe finden."

125 Teilnehmer haben die Umfrage bei Seite 2 abgebrochen, auf der sie nach den allgemeinen Nutzungsangaben gefragt haben. Dicht gefolgt mit den zweitmeisten Abbrüchen hatte die Seite 5, auf der Erklärbarkeit vorgestellt wurde.

Antworten aus den Interviews	**Antworten aus der Umfrage**
Redundantes Feature	Bedienung
Sprache der UI	Verhalten
Barrierefreiheit	Betrieb
Schlechte Bezeichnungen	Prozesse
Hintergrundinformationen	Performanz unter Systemverh.
Anwenderverhalten	Komplexität
Berechtigungen	Administration
Support	Weitere Unterkategorien zur UI: - Grafisch - Funktionell
Geschäftsprozess	Workflows / Prozesse
Texte und Bildliches als Unterteilung von Benutzeroberfläche	Schönheit in der UI-Gestaltung
UX	Systemverhalten / Fehlertoleranz
	Nutzerrollen & Berechtigungen
	Einführung um Erklärung ergänzen, wenn ein dauerhafter Hinweistext gewünscht ist
	Nachhaltigkeit & Umweltfreundlichkeit
	Barrierefreiheit
	Validierung

Tabelle 6.2: Kategorievorschläge aus Interviews und Umfragen

Abbildung 6.1 zeigt das prozentuale Auftreten der Unterkategorien innerhalb der Erhebungsmethoden mit nachträglicher Taxonomieeinführung. Auffällig ist das fehlende Auftreten von *Unerwartetes Systemverhalten* bei der Fokusgruppe, während es bei den Interviews und Umfragen ähnlich häufig mit 14,29% bzw. 13,59% auftrat.
In den Interviews wurden zudem sehr selten Erklärungsbedarfe, die der Kategorie *Algorithmus* zugeordnet werden, genannt.
Während die Unterkategorie *Konsequenzen* in den anderen beiden Erhebungsmethoden zwar auch selten auftrat, ist die Verteilung innerhalb der Umfrage mit 0,21% sehr niedrig.
Die Unterkategorie *Operation* kam bei den Interviews mit nur 1,36% auffallend selten vor, während sie besonders häufig bei der Fokusgruppe mit 7,14% vorkam. *Einführung* wurde in der Fokusgruppe nicht geäußert.
Die Kategorie *Nicht vorhandenes Feature* kam auf einen Wert von 21,43% in der Fokusgruppe und war somit deutlich höher als bei den Interviews mit

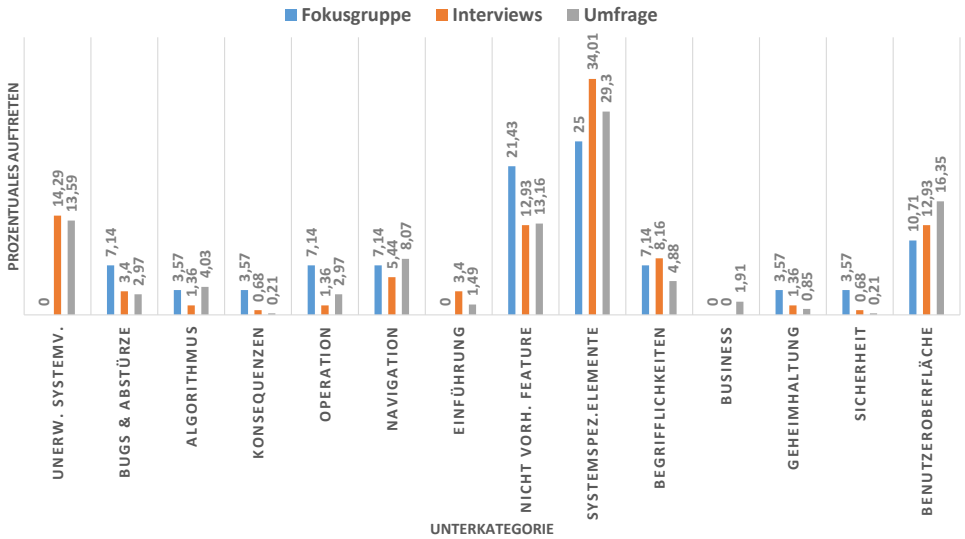

Abbildung 6.1: Prozentuale Auftreten der Unterkategorien innerhalb der Erhebungsmethoden mit nachträglicher Taxonomieeinführung

12,93% und der Umfrage mit 13,16%.
Systemspezifische Elemente traten bei den Umfragen mit 29,3%, bei der Fokusgruppe mit 25% und bei den Interviews mit 34,01% am häufigsten innerhalb dieser Methoden in Bezug auf alle Unterkategorien auf.
Die Kategorie *Business* trat im allgemeinen sehr selten auf mit 1,91% bei den Umfragen, aber gar nicht bei der Fokusgruppe und den Interviews.
Erklärungsbedarf zur *Benutzeroberfläche* trat bei den Interviews mit 12,93% auf und bei der Fokusgruppe nur etwas seltener mit 10,71%. Die Umfrage erreicht dabei den höchsten Wert mit einem Auftreten von 16,35% für diese Unterkategorie.

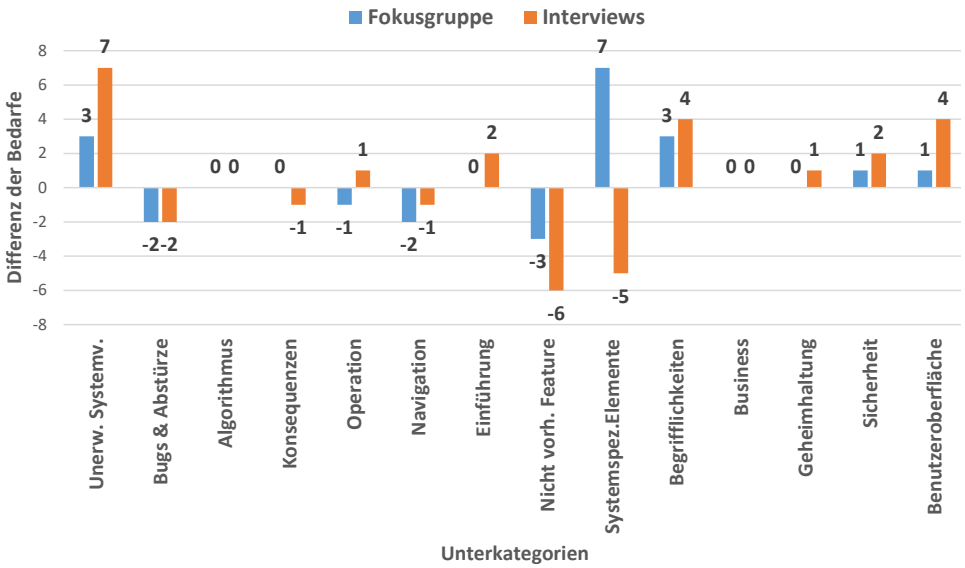

Abbildung 6.2: Differenzen der Kategorieverteilungen zwischen der Taxonomieversion ohne Einführung und der mit der direkten Einführung bei den Fokusgruppen und Interviews

Die Differenzen der Kategorieverteilungen zwischen der Taxonomieversion ohne Einführung und der mit der direkten Einführung bei den Fokusgruppen und Interviews ist in Abbildung 6.2 einsehbar. Während es bei den meisten Kategorien kaum Änderungen in der Verteilung gab, sind deutliche Trends bei *Unerwartetes Systemverhalten*, *Nicht vorhandenes Feature*, *Systemspezifische Elemente* und bei *Begrifflichkeiten* zu sehen. Erstere hat sowohl bei Fokusgruppen als auch Interviews stark zugenommen, mit drei bzw. sieben Bedarfen mehr bei direkter Taxonomieeinführung.
Nicht vorhandenes Feature ist bei beiden Erhebungsmethoden gesunken um -3 und -6 Bedarfen.
Bei *Systemspezifische Elemente* sind gegenteilige Bewegungen erkennbar: Während in den Interviews deutlich weniger Bedarfe dieser Kategorie geäußert wurden, war es bei den Fokusgruppen deutlich mehr.
Weitere Erhöhungen sind bei beiden Methoden in den Kategorien *Begrifflichkeiten* und *Benutzeroberfläche* zu sehen.

Abbildung 6.3 hingegen zeigt die Differenzen der Kategorieverteilungen zwischen der Taxonomieversion ohne Einführung und der mit der nachträglichen Einführung bei den Fokusgruppen, Interviews und der Umfrage. Erkennbar ist, dass durch die Taxonomieeinführung bei Fokusgruppe keine Erklärungsbedarfe zu *Unerwartetes Systemverhalten* hinzugekommen sind, während die Interviews und die Umfrage dort

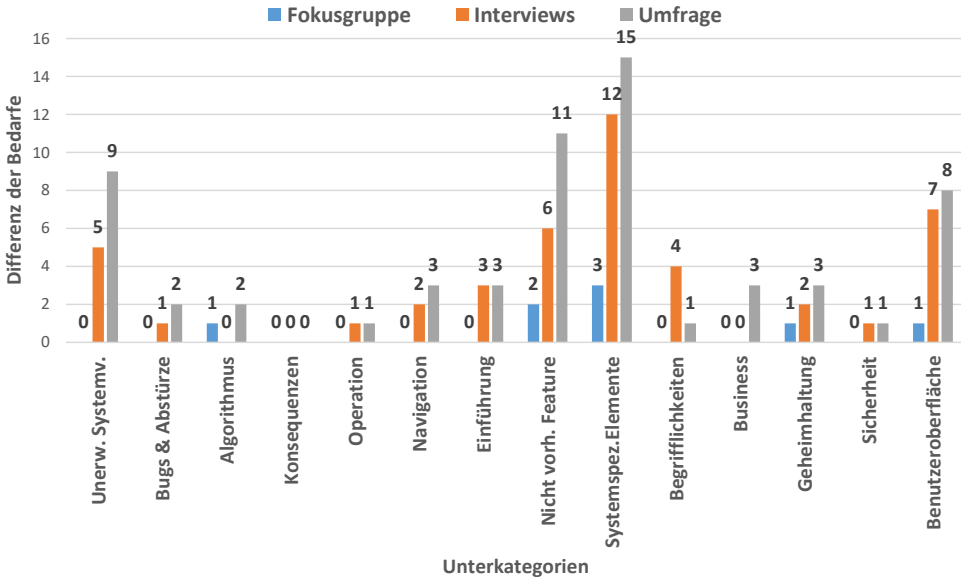

Abbildung 6.3: Differenzen der Kategorieverteilungen zwischen der Taxonomieversion ohne Einführung und der mit der nachträglichen Einführung bei den Fokusgruppen, Interviews und der Umfrage

einen Zuwachs verzeichnen. Die größten Steigungen sind für alle drei Erhebungsmethoden bei den Kategorien *Nicht vorhandenes Feature* und *Systemspezifische Elemente* zu sehen. Bei ersterem haben die Fokusgruppen zwei Bedarfe, die Interviews sechs und die Umfrage elf Erklärungsbedarfe dazugewonnen. Die höchsten positiven Differenzen sind bei *Systemspezifische Elemente* erkennbar. Bedarfe der Kategorie *Begrifflichkeiten* nahmen moderat bei der Fokusgruppe, aber sehr stark bei den Interviews mit sieben und der Umfrage mit acht Bedarfen zu.

6.2 Ergebnisse der Fokusgruppen

Hier werden die Ergebnisse der Fokusgruppen präsentiert.

Kategorie	**Mit dir. Taxonomieeinf.**			**Mit nachträgl. Taxonomieeinf.**		
	Übereinst.	**Fleiss' K**	**n**	**Übereinst.**	**Fleiss' K**	**n**
Systemverhalten						
Systemverhalten (Oberkat.)	0%	0,07	10	50,00%	0,83	6
Unerwartetes Systemverh.	0%	0,03	5	0%	0,38	3
Bugs & Abstürze	0%	0,17	4	0%	0,61	2
Algorithmus	0%	-0,01	1	25,00%	0,68	4
Konsequenzen	0%	-0,01	2	0%	0,19	1
Interaktion						
Interaktion (Oberkat.)	11,11%	0,34	18	10,00%	0,47	20
Operation	0%	0,02	11	0%	0,17	9
Navigation	0%	0,10	6	0%	0,32	7
Einführung	0%	0,26	14	7,69%	0,46	13
Nicht vorhandenes Feature						
Nicht vorh. Feature (Oberkat.)	12,50%	0,50	8	0%	0,49	7
Domänenwissen						
Domänenwissen (Oberkat.)	0%	0,20	14	0%	0,41	5
Systemspez. Elemente	0%	0,17	12	0%	0,22	3
Begrifflichkeiten	0%	0,23	6	0%	0,08	3
Business						
Business (Oberkat.)	-	-	0	0%	-0,01	2
Geheimhaltung & Sicherheit						
Geheimh. & Sicherh. (Oberkat.)	50,00%	0,84	4	100%	1,00	2
Geheimhaltung	0%	-0,01	2	50,00%	0,90	2
Sicherheit	0%	0,64	4	0%	-0,01	1
Benutzeroberfläche						
Benutzeroberfläche (Oberkat.)	0%	0,19	4	7,14%	0,38	14
Gesamt Oberkategorien	**20,00%**	**0,36**	25	**28,57%**	**0,55**	28
Gesamt Unterkategorien	**4,00%**	**0,26**	25	**14,29%**	**0,43**	28

Tabelle 6.3: Interrater-Reliabilitäten zwischen den Ratern bei beiden Fokusgruppen, wobei n für die Anzahl der relevanten Erklärungsbedarfe steht

Tabelle 6.3 zeigt die Interrater-Reliabilitäten zwischen den Ratern bei beiden Fokusgruppen. Dabei steht n für die Anzahl der relevanten Erklärungsbedarfe, in denen die jeweilige Kategorie von mindestens einem Teilnehmer genannt wird. Eine Übereinstimmung findet zwischen den Ratern nur statt, wenn alle sechs Teilnehmer die gleiche Kategorie vergeben haben.
Betrachtet man die Oberkategorie Systemverhalten, zeigt sie zwischen den Ratern eine leichte Übereinstimmung mit einem Fleiss' Kappa-Wert von 0,07 bei einem Vorkommen von 10 Erklärungsbedarfe in der Version mit direkter Taxonomieeinführung auf. Dabei haben auch die Unterkategorien ein leichtes bzw. schwaches Maß an Übereinstimmung. Die Version mit nachträglicher Taxonomieeinführung zeigt jedoch einen fast perfekten Fleiss' Kappa-Wert mit 0,83 für die Oberkategorie *Systemverhalten*. Die Unterkategorien *Bugs & Abstürze* und *Algorithmus* weisen eine erhebliche Übereinstimmung auf.

Die Version mit direkter Taxonomieeinführung zeigt für die Oberkategorie *Interaktion* eine mäßige Übereinstimmung, bei der Version mit nachträglicher Taxonomieeinführung eine moderate. Besonders uneinig waren sich die Teilnehmer hierbei in den Kategorie *Operation* mit einem Wert von 0,02 bzw. 0,17.
Nicht vorhandenes Feature hat die zweitbesten Werte mit 0,5 bzw. 0,49, während *Geheimhaltung & Sicherheit* mit 0,84 und 1,0 die besten Fleiss' Kappa-Werte erzielt. Hierbei war die prozentuale Übereinstimmung auch am höchsten mit 50% bzw. 100%.
In der *Benutzeroberfläche* war die Übereinstimmung mit direkter Einführung sehr gering mit 0%, bei einem Fleiss' Kappa von 0,19. Ohne Einführung stieg die Übereinstimmung auf 7,14% und das Kappa auf 0,38, was auf eine etwas höhere Konsistenz hindeutet.
Die Kategorie *Business* kam bei der ersten Fokusgruppe gar nicht vor, bei der zweiten nur zwei mal und hat einen negativen Median Fleiss' Kappa-Werte erzielt.
Insgesamt lässt sich sagen, dass die Fleiss' Kappa-Werte zwischen den Ratern in der Version mit nachträglicher Taxonomieeinführung in sowohl den Unterkategorien als auch in den Oberkategorien etwas besser waren. Dies spiegelt sich auch im Gesamt-Fleiss' Kappa wider. In den Oberkategorien beträgt dieser 0,36 zu 0,55, was einer schon fast erheblichen Übereinstimmung entspricht und in den Unterkategorien 0,26 zu 0,43.
Die Analyse der Interrater-Reliabilitäten zwischen einem Mehrheitsentscheid und dem Anforderungsanalysten in der Tabelle A.1 im Anhang ergab ähnliche Unterschiede in den Versionen mit direkter Taxonomieeinführung und mit nachträglicher Taxonomieeinführung. Ein Mehrheitsentscheid definiert sich durch das erste, häufigste Auftreten einer Kategorie zwischen den sechs Ratern. Wenn also zwei Rater *Algorithmus* nennen, zwei andere *Operation* und weitere zwei Rater *Systemspezifische Elemente*, so ist *Algorithmus* das Ergebnis des Mehrheitsentscheids. Zum Vergleich zwischen den einzelnen Methoden sind die Interrater-Reliabilitäten bezüglich den einzelnen Ratern und dem Anforderungsanalysten aussagekräftiger, weswegen auf die Statistiken bezüglich des Mehrheitsentscheid nicht weiter eingegangen wird.

Tabelle 6.4 zeigt die Übereinstimmungsraten pro Kategorie zwischen den sechs Ratern und dem Anforderungsanalysten bei beiden Fokusgruppen, wobei n für die Anzahl an betrachteten Teilnehmern steht.
In der Kategorie *Systemverhalten* zeigt sich ein großer Unterschied zwischen der direkten und nachträglichen Taxonomieeinführung. Bei direkter Einführung liegt der Median der Übereinstimmungsrate bei 20%, wohingegen er bei nachträglicher Einführung auf 63,33% ansteigt. Unter der Unterkategorie *Unerwartetes Systemverhalten* ist bei direkter Einführung ein Median von 0% zu beobachten, was darauf hindeutet, dass die Rater

	Mit direkter Taxonomieeinf.					Mit nachträgl. Taxonomieeinf.				
Kategorie	**Min**	**Max**	**Avg**	**Median**	**n**	**Min**	**Max**	**Avg**	**Median**	**n**
Systemverhalten										
Systemv. (Oberkat.)	0%	28,58%	18,10%	20,00%	6	60,00%	100%	71,11%	63,33%	6
Unerwart. Systemverh.	0%	33,33%	9,72%	0%	6	0%	0%	0%	0%	6
Bugs & Abstürze	0%	0%	0%	0%	6	0%	100%	58,33%	50,00%	6
Algorithmus	0%	0%	0%	0%	6	25,00%	100%	62,50%	50,00%	6
Konsequenzen	0%	50,00%	8,33%	0%	6	0%	100%	33,33%	0%	6
Interaktion										
Interaktion (Oberk.)	0%	14,20%	6,80%	6,97%	6	0%	44,44%	28,92%	32,05%	6%
Operation	0%	10,00%	1,67%	0%	6	0%	100%	36,27%	26,67%	6
Navigation	0%	0%	0%	0%	6	0%	66,67%	38,89%	41,67%	6
Einführung	0%	0%	0%	0%	6	0%	0%	0%	0%	6
Nicht vorh. Feature										
Nicht vorh. Feature	0%	33,33%	20,04%	19,64%	6	14,29%	71,43%	43,95%	43,75%	6
Domänenwissen										
Domänenw. (Oberkat.)	0%	47,06%	27,52%	34,52%	6	0%	33,33%	16,67%	16,67%	6
Systemspez. Elemente	0%	44,44%	18,52%	11,67%	6	0%	14,29%	9,23%	13,39%	6
Begrifflichkeiten	0%	40,00%	23,65%	24,29%	6	0%	50,00%	8,33%	0%	6
Business										
Business	-	-	-	-	0	-	-	-	-	0
Geheimhaltung & Sicherheit										
Geheimh. & Sicherh. (Oberkat.)	50,00%	100%	91,67%	100%	6	100%	100%	100%	100%	6
Geheimhaltung	0%	0%	0%	0%	6	50,00%	100%	58,33%	50,00%	6
Sicherheit	0%	100%	75,00%	100%	6	0%	100%	16,67%	0%	6
Benutzeroberfläche										
Benutzeroberfl. (Oberkat.)	0%	50,00%	16,67%	12,50%	6	20,00%	50,00%	32,82%	30,95%	6
Gesamt Oberkat.	16,00%	56,00%	37,33%	38,00%	6	39,28%	67,86%	54,17%	55,36%	6
Gesamt Unterkat.	16,00%	36,00%	28,00%	28,00%	6	21,43%	64,29%	41,67%	41,07%	6

Tabelle 6.4: Übereinstimmungsrate pro Kategorie zwischen den sechs Ratern und dem Anforderungsanalysten bei beiden Fokusgruppen, wobei n für die Anzahl an betrachteten Teilnehmern steht

keine einheitlichen *unerwarteten Systemverhalten* identifizieren konnten. Bei nachträglicher Taxonomieeinführung bleibt der Wert ebenfalls bei 0%. Sowohl bei *Bugs & Abstürze* als auch bei *Algorithmus* bringt die nachträgliche Einführung eine Steigerung von 0% Übereinstimmung auf auf 50%. Der Median von *Konsequenzen* blieb bei beiden bei 0%.
Die Übereinstimmungen bei *Operation* und *Einführung* waren bei der Version mit direkter Taxonomieeinführung bei 0% und bei der Fokusgruppe mit nachträglicher Taxonomieeinführung mit 26,67% bei *Operation* immer noch sehr niedrig, wohingegen *Einführung* immer noch so blieb.
Nicht vorhandenes Feature erzielte mit 19,64% bei direkter Einführung eine niedrige Übereinstimmung, bei nachträglicher eine etwas höhere mit 43,75%.
Business wurde nicht betrachtet aufgrund geringen Vorkommens.
Bei *Geheimhaltung & Sicherheit* waren sich die Rater mit dem Anforderungsanalysten in beiden Versionen einig.
Die *Benutzeroberfläche* erhielt sehr geringe Übereinstimmung mit 12,50%, jedoch etwas höhere bei nachträglicher Taxonomieeinführung mit 39,95%.
Betrachtet man die Gesamtkappa-Werte, so bildet sich ein ähnliches Bild wie bei der Interrater-Reliabilität zwischen den Mehrheitsentscheid

und dem Anforderungsanalysten ab. Die Version mit nachträglicher Taxonomieeinführung erzielt insgesamt eine höhere Übereinstimmungsrate.

Tabelle 6.5 hingegen zeigt die Cohen's Kappa-Werte dafür an. Dort sind ähnliche Ergebnisse ersichtlich. Auch hier fallen besonders die Kategorien *Unerwartetes Systemverhalten, Algorithmus, Operation* und *Einführung* auf.

	Mit direkter Taxonomieeinf.					**Mit nachträgl. Taxonomieeinf.**				
Kategorie	**Min**	**Max**	**Avg**	**Median**	**n**	**Min**	**Max**	**Avg**	**Median**	**n**
Systemverhalten										
Systemv. (Oberkat.)	-0,12	0,32	0,20	0,25	6	0,71	1,00	0,79	0,73	6
Unerwart. Systemverh.	-0,06	0,47	0,10	-0,03	6	0	0	0	0	5
Bugs & Abstürze	0	0	0	0	4	0	1,00	0,66	0,65	6
Algorithmus	0	0	0	0	1	0,36	1,00	0.72	0,65	6
Konsequenzen	0	0,65	0,10	0	6	0	1,00	0,33	0	6
Interaktion										
Interaktion (Oberk.)	-0,08	0,19	0,06	0,06	6	-0,21	0,52	0,28	0,34	6
Operation	-0,06	0,12	0,01	0	6	-0,05	1,00	0,41	0,36	6
Navigation	0	0	0	0	3	0	0,78	0,50	0,56	6
Einführung	0	0	0	0	5	0	0	0	0	6
Nicht vorh. Feature										
Nicht vorh. Feature	-0,06	0,47	0,27	0,28	6	0,16	0,78	0,52	0,53	6
Domänenwissen										
Domänenw. (Oberkat.)	0	0,33	0,20	0,27	6	0	0,41	0,21	0,21	6
Systemspez. Elemente	-0,16	0,51	0,14	0,06	6	0	0,20	0,12	0,16	6
Begrifflichkeiten	0	0,52	0,30	0,31	6	-0,05	0,65	0,08	-0,03	6
Business										
Business	-	-	-	-	0	0	0	0	0	1
Geheimhaltung & Sicherheit										
Geheimh. & Sicherh. (Oberkat.)	0,63	1,00	0,94	1,00	6	1,00	1,00	1,00	1,00	6
Geheimhaltung	0	0	0	0	1	0,65	1,00	0,71	0,65	6
Sicherheit	0	1,00	0,78	1,00	6	0	1,00	0,17	0	6
Benutzeroberfläche										
Benutzeroberfl. (Oberkat.)	0	0,62	0,22	0,17	6	0,21	0,63	0,40	0,39	6
Gesamt Oberkat.	0,11	0,38	0,26	0,26	6	0,29	0,62	0,46	0,48	6
Gesamt Unterkat.	0,13	0,25	0,21	0,22	6	0,15	0,61	0,37	0,36	6

Tabelle 6.5: Cohen's Kappa pro Kategorie zwischen den sechs Ratern und dem Anforderungsanalysten bei beiden Fokusgruppen, wobei n für die Anzahl der betrachteten Teilnehmer steht

6.3 Ergebnisse der Interviews

Für die Interviews wurden ebenfalls die Übereinstimmungsrate und die Cohen's Kappa-Werte zwischen den Ratern und dem Anforderungsanalysten berechnet. Tabelle 6.6 geht auf die Übereinstimmungsrate pro Kategorie

	Mit direkter Taxonomieeinf.					Mit nachträgl. Taxonomieeinf.				
Kategorie	**Min**	**Max**	**Avg**	**Median**	**n**	**Min**	**Max**	**Avg**	**Median**	**n**
Systemverhalten										
Systemv. (Oberkat.)	0%	100%	45,75%	62,50%	7	0%	100%	59,58%	66,67%	8
Unerwart. Systemverh.	0%	77,78%	34,39%	39,29%	6	0%	80%	45,42%	58,33%	8
Bugs & Abstürze	0%	100%	44,44%	33,33%	3	0%	66,67%	16,67%	0%	4
Algorithmus	0%	0%	0%	0%	3	0%	100%	50%	50%	2
Konsequenzen	0%	0%	0%	0%	2	0%	0%	0%	0%	3
Interaktion										
Interaktion (Oberk.)	0%	50%	25%	25%	7	22,22%	100%	45,04%	36,36%	7
Operation	0%	0%	0%	0%	5	0%	0%	0%	0%	5
Navigation	0%	50%	15,48%	0%	7	0%	50%	30,28%	33,33%	6
Einführung	0%	100%	75%	100%	4	33,33%	100%	66,67%	50%	5
Nicht vorh. Feature										
Nicht vorh. Feature	0%	50%	26,67%	33,33%	5	33,33%	100%	63,33%	50%	7
Domänenwissen										
Domänenw. (Oberkat.)	0%	100%	32,3%	25%	9	0%	100%	43,76%	50%	9
Systemspez. Elemente	0%	100%	24,18%	14,29%	9	0%	100%	34,98%	33,33%	9
Begrifflichkeiten	0%	66,67%	19,44%	0%	6	0%	71,43%	24,11%	12,50%	4
Business										
Business	-	-	-	-	0	0%	0%	0%	0%	1
Geheimhaltung & Sicherheit										
Geheimh. & Sicherh. (Oberkat.)	100%	100%	100%	100%	3	0%	100%	37,5%	25%	4
Geheimhaltung	100%	100%	100%	100%	1	100%	100%	100%	100%	1
Sicherheit	100%	100%	100%	100%	2	0%	50%	16,67%	0%	3
Benutzeroberfläche										
Benutzeroberfl. (Oberkat.)	0%	100%	50,4%	50%	9	0%	100%	53,33%	46,67%	8
Gesamt Oberkat.	0%	87,50%	53,53%	57,89%	9	46,15%	100%	65,30%	63,64%	9
Gesamt Unterkat.	0%	75%	44,18%	45%	9	42,31%	90,91%	57,87%	54,55%	9

Tabelle 6.6: Übereinstimmungsrate pro Kategorie zwischen den je 9 Ratern pro Version und dem Anforderungsanalysten bei allen 18 Interviews, wobei n für die Anzahl der betrachteten Teilnehmer steht

zwischen den je 9 Ratern pro Version und dem Anforderungsanalysten bei allen 18 Interviews ein, wobei n für die Anzahl der betrachteten Teilnehmer steht. Dort sind höhere Werte bei den meisten Oberkategorien als zwischen den sechs Ratern der Fokusgruppe und dem Analysten erkennbar. So war die Oberkategorie *Systemverhalten* bei den Interviews mit direkter Einführung der Taxonomie mit 62,50% ähnlich hoch wie bei der mit nachträglicher mit 66,67%. Dabei ist die Übereinstimmung bei den Unterkategorien *Bugs% Abstürze, Algorithmus* und *Konsequenzen* auf einem sehr niedrigen Niveau. Auch zeigt die Anzahl der Teilnehmer, die für die Berechnung der Median-Werte in Betracht gezogen wurden, dass diese Kategorien relativ selten im ausreichenden Maße geäußert wurden.
Interaktion zeigt geringere Übereinstimmungsraten auf. So wurde *Operation*

von allen zehn Teilnehmern, bei denen diese Kategorie relevant war, komplett falsch eingeschätzt. *Navigation* war mit 0% bzw. 33,33% auch sehr niedrig. Einzig *Einführung* kam auf gute Werte, mit 100% bzw. 50%.
Die gesamten Übereinstimmungsraten sind hier höher bei den Interviews mit nachträglicher Taxonomieeinführung, wobei die Differenzen zwischen ihnen nicht so hoch sind wie bei den Fokusgruppen mit 17,36% in den Oberkategorien und 13,07% in den Unterkategorien. Hier sind es nur 10,11% in den Oberkategorien und 9,45% in den Unterkategorien. Zu beachten sind hier die Median-Werte, die sich auf einem hohen Niveau bewegen mit 63,64% Übereinstimmung in den Ober- und 54,55% in den Unterkategorien.
Die Tabelle 6.7 zeigt ähnliche Resultate auf. So befinden sich alle Median-Werte des Gesamtkappas und der Oberkategorien bis auf Benutzeroberfläche bei den Interviews mit nachträglicher Taxonomieeinführung in einer positiven Differenz zu denen bei der Version mit direkter Einführung.

	Mit direkter Taxonomieeinf.					**Mit nachträgl. Taxonomieeinf.**				
Kategorie	**Min**	**Max**	**Avg**	**Median**	**n**	**Min**	**Max**	**Avg**	**Median**	**n**
Systemverhalten										
Systemv. (Oberkat.)	-0,09	1,00	0,41	0,47	7	0	1,00	0,60	0,69	8
Unerwart. Systemverh.	0	0,79	0,36	0,37	6	0	0,78	0,48	0,61	8
Bugs & Abstürze	0	1,00	0,49	0,46	3	0	0,74	0,19	0	4
Algorithmus	0	0	0	0	3	0	1,00	0,50	0,50	2
Konsequenzen	0	0	0	0	2	0	0	0	0	3
Interaktion										
Interaktion (Oberk.)	0	0,50	0,26	0,29	7	0,29	1,00	0,48	0,41	7
Operation	-0,10	0	-0,04	0	5	-0,09	0	-0,03	0	5
Navigation	0	0,63	0,18	0	7	0	0,62	0,38	0,41	6
Einführung	0	1,00	0,75	1,00	4	0,47	1,00	0,75	0,65	5
Nicht vorh. Feature										
Nicht vorh. Feature	0	0,64	0,30	0,23	5	0,31	1,00	0,69	0,62	7
Domänenwissen										
Domänenw. (Oberkat.)	0	1,00	0,23	0,03	9	0	1,00	0,42	0,53	9
Systemspez. Elemente	-0,24	1,00	0,15	0	9	0	1,00	0,35	0,20	9
Begrifflichkeiten	-0,07	0,67	0,19	0	6	0	0,80	0,27	0,14	4
Business										
Business	-	-	-	-	0	0	0	0	0	1
Geheimhaltung & Sicherheit										
Geheimh. & Sicherh. (Oberkat.)	1,00	1,00	1,00	1,00	3	0	1,00	0,40	0,30	4
Geheimhaltung	1,00	1,00	1,00	1,00	1	1,00	1,00	1,00	1,00	1
Sicherheit	1,00	1,00	1,00	1,00	2	0	0,60	0,20	0	3
Benutzeroberfläche										
Benutzeroberfl. (Oberkat.)	0	1,00	0,48	0,44	9	0	1,00	0,54	0,49	8
Gesamt Oberkat.	0	0,79	0,40	0,45	9	0,32	1,00	0,53	0,47	9
Gesamt Unterkat.	0	0,68	0,34	0,37	9	0,29	0,88	0,47	0,43	9

Tabelle 6.7: Cohen's Kappa pro Kategorie zwischen den je 9 Ratern pro Version und dem Anforderungsanalysten bei allen 18 Interviews, wobei n für die Anzahl der betrachteten Teilnehmer steht

	Übereinstimmung				
Kategorie	**Min**	**Max**	**Avg**	**Median**	**n**
Systemverhalten					
Systemv. (Oberkat.)	0%	100%	47,99%	50%	91
Unerwart. Systemverh.	0%	100%	34,72%	0%	72
Bugs & Abstürze	0%	100%	40,62%	0%	16
Algorithmus	0%	100%	12%	0%	25
Konsequenzen	0%	100%	8,33%	0%	12
Interaktion					
Interaktion (Oberk.)	0%	100%	27,53%	0%	99
Operation	0%	100%	11,84%	0%	38
Navigation	0%	100%	38,33%	0%	60
Einführung	0%	100%	15,38%	0%	26
Nicht vorh. Feature					
Nicht vorh. Feature	0%	100%	45,41%	33,33%	57
Domänenwissen					
Domänenw. (Oberkat.)	0%	100%	22,06%	0%	101
Systemspez. Elemente	0%	100%	9,8%	0%	92
Begrifflichkeiten	0%	100%	13,02%	0%	32
Business					
Business	0%	100%	33,33%	0%	12
Geheimhaltung & Sicherheit					
Geheimh. & Sicherh. (Oberkat.)	0%	100%	40,74%	0%	9
Geheimhaltung	0%	100%	44,44%	33,33%	6
Sicherheit	0%	100%	33,33%	0%	3
Benutzeroberfläche					
Benutzeroberfl. (Oberkat.)	0%	100%	38,83%	0%	94
Gesamt Oberkat.	0%	100%	51,86%	50%	188
Gesamt Unterkat.	0%	100%	43,44%	50%	188

Tabelle 6.8: Übereinstimmungsrate zwischen den 188 Umfrageteilnehmern und dem Anforderungsanalysten pro Kategorie Unter der Betrachtung der Version mit nachträglicher Taxonomieeinführung, wobei n für die Anzahl der betrachteten Teilnehmer steht

6.4 Ergebnisse der Umfrage

Die Onlineumfrage in der Version mit nachträglicher Taxonomieeinführung liefert die in Tabelle 6.8 aufgeführten statistischen Werte für die Übereinstimmungsrate zwischen jedem einzelnen Rater und dem Anforderungsanalysten. Die Cohen's Kappa-Werte sind hierbei zu

uneindeutig und werden nicht aufgeführt, da bei der Umfrage, wie in Tabelle 6.1 zu sehen, im Durchschnitt nur 2,51 Erklärungsbedarfe pro Person aufgestellt wurden. Da zur Berechnung von validen Cohen's Kappa-Werten mindestens zwei Aussagen benötigt werden, konnten an vielen Stellen keine Berechnungen dafür durchgeführt werden.

Die Übereinstimmungsraten bei der Umfrage zeigen im Median in allen Ober- und Unterkategorien niedrigere Werte als bei der Fokusgruppe und den Interviews mit nachträglicher Taxonomieeinführung an. Auf die Gesamtübereinstimmung bezogen hat die Umfrage einen Median-Wert von je 50% für sowohl Oberkategorien als auch Unterkategorien ergeben. Dies ist bei den Unterkategorien um 8,33% höher als bei der Fokusgruppe, jedoch 4,55% niedriger als bei den Interviews.

Eine Möglichkeit zur Analyse der Werte bietet auch die Betrachtung der Durchschnittswerte, da oftmals vermehrt eine Übereinstimmung von 0% bei den einzelnen Kappa-Werten erzielt wird und die Median-Werte daher etwas ungenau sind. Da zeigt sich für die Kategorien, für die Negativbeispiele zusätzlich erwähnt wurden, wie für *Unerwartetes Systemverhalten*, eine leicht niedrigere durchschnittliche Übereinstimmung von 34,72% als bei den Interviews, aber eine höhere als bei den Fokusgruppen. *Algorithmus* ist bei der Umfrage mit 12% deutlich niedriger als bei den anderen beiden Erhebungsmethoden in der selben Version (Fokusgruppe 62,50%, Interviews 0%).
Für die Kategorie *Operation* gilt eine Übereinstimmungsrate von nur 11,84% bei der Umfrage, dies ist höher als bei den Interviews, bei denen es gar keine Übereinstimmung gab. Bei den Fokusgruppen war in dieser Version der Taxonomieeinführung die Übereinstimmung mit durchschnittlich 36,27% höher.
Nachdem die Kategorie *Einführung* bei den Fokusgruppen gar keine Übereinstimmung erzielte, war diese mit 66,67% in den Interviews hoch. Bei der Umfrage wurden nur 15,38% erreicht.
Nicht vorhandenes Feature ist bei der Umfrage mit einem Durchschnitts-Übereinstimmungswert von 45,41% knapp höher als bei der Fokusgruppe (43,95%), aber niedriger als bei den Interviews, die auf 63,33% kommen.

Kapitel 7

Diskussion

Dieses Kapitel behandelt die Diskussion der Ergebnisse, indem zunächst die Forschungsfragen dieser Arbeit betrachtet werden. Um die Ergebnisse aus Kapitel 6 korrekt einzuordnen, folgt eine Diskussion mittels Interpretation dieser. Mögliche Validity Threats werden benannt.

7.1 Beantwortung der Forschungsfragen

7.1.1 Einflüsse der Erkenntnisse in die weiteren Erhebungsmethoden

RQ1.1: *Welche Erkenntnisse aus den Fokusgruppen können für die Durchführung der Interviews verwendet werden?*

Zur Beantwortung dieser Forschungsfrage wird auf die bereits genannten Punkte in Unterkapitel 5.1 verwiesen.

- Aufgrund des gehäuften fehlerhaften Auftretens der Kategorievergabe von *Nicht vorhandenes Feature* und *Einführung* bei der ersten Fokusgruppe wurde bereits bei der Durchführung dieser darauf aufmerksam gemacht, dass die Teilnehmer ihre Erklärungsbedarfe präziser äußern sollen. Um außerdem dieser gehäuften Falschvergabe zu entgehen, wurde die Reihenfolge der Kategorien innerhalb der Taxonomie aus Abbildung 2.2 mitsamt der zugehörigen Beispielen angepasst. Dabei wurde *Nicht vorhandenes Feature* von der letzten Position 3 Plätze weiter nach oben positioniert.
- Um die gehäufte Falschvergabe von *Einführung* zu verhindern, wurde darauf hingewiesen, dass ein Erklärungsbedarf, der zu dieser Kategorie gehört, den Wunsch einer Einführung beinhalten muss.
- Bei der Kategorisierung ist ebenfalls aufgefallen, dass die Teilnehmer zu viel subjektive Interpretationen in die Vergabe der Kategorien

einfließen lassen haben. Deswegen wurde ihnen Nahe gebracht, ihre Erklärungsbedarfe unabhängig von ihren dahinterliegenden Wünschen und Gedanken zu kategorisieren.

- Da die Teilnehmer der ersten Fokusgruppe versucht haben, mehrere Kategorien für einen Erklärungsbedarf anzugeben, wurde diese Restriktion den Teilnehmern vorher klargestellt.

- Der in Kapitel 4.3.1 vorgestellte Ablaufplan sah einen strikten Vorgang bei der Erklärungsbedarfsermittlung vor, indem nach der Taxonomieeinführung anhand der Oberkategorien gesammelt wurde. Es hat sich als vorteilhafter herausgestellt, wenn man zunächst die bereits offenen Erklärungsbedarfe notiert.

RQ1.2: *Welche Erkenntnisse aus den Fokusgruppen und Interviews können für die Gestaltung der Onlineumfrage verwendet werden?*

Die erwähnten Erkenntnisse für die Interviews flossen auch in die Gestaltung der Onlineumfrage ein.

- Es wird darauf hingewiesen, dass die Teilnehmer ihre Fragen und Erklärungsbedarfe so formulieren sollen, dass ein Entwickler sie optimal umsetzen oder beantworten kann. Die Teilnehmer sollten ihre Anliegen daher so ausdrücken, dass die mögliche Antwort genau die Informationen enthält, die sie benötigen.

- Der fehlerhaften Kategorievergabe von *Nicht vorhandenes Feature* wurde entgegengewirkt, indem die angepasste Reihenfolge innerhalb der Taxonomie aus Abbildung 2.2, die auch schon für die Interviews verwendet wurde, innerhalb der Umfrage abgebildet wurde. Zu der Vergabe von *Einführung* wurde erwähnt, dass Erklärungsbedarfe, die zur Kategorie Einführung gehören, auch explizit Signalwörter wie „ich brauche eine Einführung / ein Tutorial / ein Guide" enthalten müssen.

- Zusätzlich wurden den Kategorien, die schlechte Interrater-Reliabilitäten aufwiesen oder bei der ersten Fokusgruppe auffallend oft falsch vergeben wurden, Negativbeispiele unterhalb der Beispielliste gezeigt. Diese sind *Unerwartetes Systemverhalten*, Algorithmus, Operation, Einführung, und Nicht vorhandenes Feature.

- Bei der Gestaltung der Onlineumfrage konnte die Möglichkeit, einem Erklärungsbedarf mehrere Kategorien vergeben zu können, mittels passender Gestaltung der Eingabefelder vermieden werden.

- Zusätzlich wurde die Schlussfrage „Würdest du eine (Unter)-kategorie aus der Taxonomie entfernen und wenn ja, welche und warum?" gestellt, um zu überprüfen, ob die Teilnehmer eine Kategorie als überflüssig ansahen.

7.1.2 Vergleich von verschiedenen Erhebungsmethoden

RQ2.1: *Inwieweit sind Fokusgruppen, Interviews oder Umfragen effektiver für die Erhebung von Erklärbarkeitsanforderungen bei Nutzern einer Personalverwaltungssoftware?*

Zur Beantwortung dieser Frage werden die drei unterschiedlichen Versionen der Taxonomieeinführung betrachtet. Die Interviews mit direkter Taxonomieeinführung sind mit 11,67 durchschnittlichen einzigartigen Bedarfen pro Person deutlich effektiver als die Fokusgruppe (3,83). Vergleicht man die drei Erhebungsmethoden in der Version ohne Taxonomieeinführung, so schneidet die Umfrage am schlechtesten mit nur 1,74 einzigartigen Bedarfen pro Person (Fokusgruppe = 3,17, Interviews = 10,67).
Bei Betrachtung der Version mit nachträglicher Einführung der Taxonomie, sind Interviews mit 14,78 ebenfalls am effektivsten (Fokusgruppe = 4,50, Umfrage = 1,94).

RQ2.2: *Inwieweit sind Fokusgruppen, Interviews oder Umfragen effizienter für die Erhebung von Erklärungsbedarfen bei Nutzern einer Personalverwaltungssoftware?*

Bei der Version mit direkter Taxonomieeinführung sind Interviews mit 0,57 einzigartigen Bedarfen pro Person pro Minute deutlicher effizienter als die Fokusgruppe (0,20). Im Vergleich zwischen den Versionen ohne Taxonomieeinführung führen die Interviews mit 1,10 einzigartigen Bedarfen pro Person pro Minute, gefolgt von der Umfrage (0,72) und der Fokusgruppe (0,25). Bei Betrachtung der Version mit nachträglicher Taxonomieeinführung ist die Reihenfolge gleich mit 0,71 einzigartigen Bedarfen pro Person pro Minute bei den Interviews, 0,54 bei der Umfrage und nur 0,17 bei der Fokusgruppe. Bezogen auf die Anzahl an einzigartigen Bedarfen pro stündlichem Personalaufwand sind ebenfalls Interviews am effizientesten mit 11,49 einzigartigen Bedarfen pro stündlichem Personalaufwand, die Umfrage (9,33) am zweiteffizientesten und die Fokusgruppe am geringsten (4,45).

RQ2.3: *Wie unterscheiden sich die Ergebnisse der Fokusgruppen, der Interviews und der Onlineumfrage hinsichtlich der in den jeweiligen Kategorien erfassten Erklärbarkeitsanforderungen?*

In den Erhebungsmethoden mit nachträglicher Taxonomieeinführung wurden Differenzen zwischen den Verteilungen der Unterkategorien innerhalb der Methoden deutlich. *Unerwartetes Systemverhalten* trat in der Fokusgruppe gar nicht auf, wobei Interviews und die Umfrage relativ hohe Verteilung erzielten. Auffallend war auch, dass Erklärungsbedarfe zu Konsequenzen bei den Umfragen sehr selten genannt wurde mit einem Auftreten von 0,21%. Dafür überwiegt deutlich der Anteil an Bedarfen der Unterkategorie Systemspezifische Elemente. Die Interviews erzielten den Höchstwert (34,01%), während die Umfrage (29,3%) und die Fokusgruppe (25%) ein ähnlich hohes Aufkommen vorweisen. Die Kategorie Business kam sehr selten vor bei der Umfrage, aber gar nicht bei den Interviews und der Fokusgruppe. Die Umfrage erreicht mit 16,35% den Höchstwert bei der Kategorie Benutzeroberfläche im Vergleich zu den Interviews und der Fokusgruppe.

7.1.3 Anwendung einer Taxonomie

RQ3.1: *Wie unterscheiden sich die erhobenen Erklärungsbedarfe bei Anwendung einer Taxonomie?*

Zwischen den Fokusgruppen und Interviews zeigt sich, dass eine direkte Einführung der Taxonomie zu 23 einzigartigen Bedarfen bei der Fokusgruppe und zu 105 bei den Interviews geführt hat. Die Version ohne Taxonomie brachte es nur zu 17,39% weniger einzigartigen Bedarfen bei der Fokusgruppe und zu 8,57% weniger bei den Interviews. Die meisten einzigartigen Erklärungsbedarfe erzielten die Durchführungen mit nachträglicher Einführung der Taxonomie. Dies sind mit 27 bei der Fokusgruppe rund 17,39% mehr einzigartige Bedarfe als bei der Version mit direkter Einführung und bei den Interviews mit 133 rund 26,67% mehr einzigartige Bedarfe. Bei der Umfrage hat die nachträgliche Einführung der Taxonomie mit 364 zu 11,31% mehr einzigartigem Erklärungsbedarf verholfen.

Bezüglich der Verteilungen der Kategorien ist erkennbar, dass eine direkte Taxonomieeinführung bei der Fokusgruppe starke Zuwächse in den Kategorien *Unerwartetes Systemverhalten*, *Systemspezifische Elemente* und *Begrifflichkeiten* erzielt hat, aber Verluste vor allem bei *Nicht vorhandenes Feature*. Bei den Interviews sind Steigungen ebenfalls bei *Unerwartetes Systemverhalten* und *Begrifflichkeiten* erkennbar, aber auch bei *Benutzeroberfläche*. Ebenfalls ist bei *Nicht vorhandenes Feature* der stärkste Verlust zu sehen. Im Gegensatz zur Fokusgruppe haben die Interviews starke Verluste bei der Kategorie *Systemspezifische Elemente* verbüßt.

Bei Betrachtung der Differenzen zwischen der Version ohne und der mit nachträglicher Taxonomieeinführung bei allen drei Erhebungsmethoden, so sind vor allem Zuwächse in den Kategorien *Unerwartetes Systemverhalten*, *Nicht vorhandenes Feature*, *Systemspezifische Elemente* und *Benutzeroberfläche* erkennbar, wobei die Fokusgruppe dies für letztere und *Unerwartetes Systemverhalten* nicht teilt.

RQ3.2: *Reicht die vorgegebene Taxonomie aus, um den gesamten erhobenen Erklärungsbedarf den jeweiligen Kategorien zuzuordnen?*

Sowohl der Anforderungsanalyst als auch die Teilnehmer der Erhebungen konnten allen Erklärungsbedarfen eine Kategorie zuweisen. Die durchschnittlichen Werte zur Frage, wie sicher sich die Teilnehmenden bei der Zuordnung zu den Kategorien fühlten, wurden auf einer Likert-Skala von 1 bis 5 erhoben, wobei 1 „nicht sicher" und 5 „sehr sicher" bedeutete. Die Ergebnisse zeigen, dass die Teilnehmenden bei den Umfragen eine etwas geringere Sicherheit angaben (M = 3,15) als in den Fokusgruppen (M = 3,33), während sie sich in den Interviews am sichersten fühlten (M = 3,69). Acht Teilnehmer der Interviews äußerten elf Kategorievorschläge zur Erweiterung der Taxonomie, während 16 Umfrageteilnehmer je einen Vorschlag oder den Wunsch einer Spezifizierung bestimmter Kategorien nannten. Ein Beispiel dafür ist die Unterteilung von *Benutzeroberfläche* in grafisch und funktionell. Ob solche Unterteilungen sinnvoll sind, muss weiter evaluiert werden.

7.1.4 Interrater-Reliabilität bei Zuweisung der Taxonomiekategorien

RQ4.1: *Inwieweit weicht die Einordnung in die Kategorien seitens der Nutzer im Vergleich zur Einordnung durch den Anforderungsanalysten ab?*

Zwischen den beiden qualitativen Erhebungsmethoden, also den Fokusgruppen und Interviews, sind Übereinstimmungswerte von 28% für die Fokusgruppe und 45% für die Interviews mit direkter Taxonomieeinführung bei Betrachtung der Unterkategorien berechnet worden. Die Cohen's Kappa-Werte sind auf einem ähnlichen Differenzniveau (0,22 zu 0,37). Für die Fokusgruppe und für die Interviews mit nachträglicher Einführung der Taxonomie ergaben sich insgesamt bessere Übereinstimmungswerte (41,07% zu 54,55%). Die Kappa-Werte sind hier ebenfalls erhöht (0,36 zu 0,43).
Zwischen der Fokusgruppe, den Interviews und der Umfrage mit nachträglicher Taxonomieeinführung sind für die Unterkategorien durchschnittliche Übereinstimmungsraten von 41,67%, 57,87% und 43,44% berechnet worden.

RQ4.2: *Inwieweit weicht die Einordnung in die Kategorien zwischen den Nutzern ab?*

Die Interrater-Reliabilitäten zwischen den Ratern bei beiden Fokusgruppen zeigen, dass sie sich bei der Fokusgruppe mit direkter Taxonomieeinführung bei Betrachtung der Oberkategorien einigermaßen und bei der Fokusgruppe mit nachträglicher Taxonomieeinführung mit einem Fleiss' Kappa-Wert von 0,55 fast erheblich einig waren. In den Unterkategorien sind die Werte niedriger (0,26 bzw. 0,43).

7.2 Interpretation der Ergebnisse

Bei Betrachtung der Ergebnisse in Kapitel 6 wird deutlich, dass zwischen den drei durchgeführten Erhebungsmethoden Interviews am effektivsten für die Erhebung von Erklärbarkeitsanforderungen bei Nutzern einer Personalverwaltungssoftware sind. Bei Verfolgung des Ziels der Effektivität bei einer Erhebung von Erklärbarkeitsanforderungen sind Interviews somit am besten geeignet. Am wenigsten einzigartige Bedarfe pro Person hat die Umfrage erzielt, weswegen sie bezüglich der Effektivität nicht empfohlen wird.
Besitzt der Anforderungsanalyst die Priorität, besonders effizient die Erklärbarkeitsanforderungen zu erheben, so sollte die Wahl ebenfalls auf die Interviews fallen. Dabei war die Umfrage nur etwas weniger effizient und kann auch in Betracht gezogen werden. Die Fokusgruppen schnitten deutlich schlechter ab. Zudem muss man den Organisationsaufwand einberechnen, eine Fokusgruppe an einem Ort zur gleichen Zeit mit mehreren Teilnehmern durchzuführen. Der Aufwand dafür ist zwar je nach Unternehmen variabel, jedoch zeigen die zuvor genannten Werte, dass die Durchführung einer Fokusgruppe deutlich ineffizienter ist und nicht bevorzugt werden sollte.
Ist die Wahl der Erhebungsmethode eine Kostenfrage, so muss man die Werte für die Anzahl der einzigartigen Bedarfe pro stündlichem Personalaufwand untersuchen. Diese Resultate zeigen, dass selbst mit Einberechnung des Aufwands des Anforderungsanalysten, die bei der Umfrage wegfällt, die Interviews die meisten einzigartigen Bedarfe am kostengünstigsten liefern. Ein Unternehmen, das der Wirtschaftlichkeit nachgeht, sollte ebenfalls die Interviews bevorzugen.

Für die Frage, ob der Einsatz einer Taxonomie seitens eines Anforderungsanalysten bei der Erhebung von Erklärbarkeitsanforderungen in Betracht gezogen werden soll, müssen mehrere Aspekte betrachtet werden. So ist mit 12,70 einzigartigen Bedarfen pro stündlichem Personalaufwand die Version ohne Taxonomieeinführung bei Interviews

am effizientesten. Jedoch fällt bei Betrachtung der durchschnittlichen Anzahl an einzigartigen Erklärungsbedarfen pro Person auf, dass die nachträgliche Taxonomieeinführung 38,52% mehr Bedarfe liefert, obwohl die Effizienz mit 11,49 Bedarfen pro stündlichem Personalaufwand nur 9,53% niedriger ist. Daraus lässt sich schlussfolgern, dass es lohnenswert ist, bei der Durchführung einer Erhebung an Erklärbarkeitsanforderungen eine Taxonomie als Unterstützung in Betracht zu ziehen. Es zeigt sich zudem, dass die Version mit nachträglicher Taxonomieeinführung besser dafür geeignet ist als die mit direkter.
Die Ergebnisse der Umfrage unterstützen diese Aussage, da dort die Werte für die Effizienz bezüglich des Personalaufwands bei der Version mit nachträglicher Taxonomieeinführung (9,33) sogar höher sind als bei der ohne (9,24). Die Effektivität ist dort ebenfalls höher.
Die Ergebnisse legen nahe, dass zukünftige Forschungen auf die Entwicklung von klareren und präziseren Taxonomien für Anforderungsanalysen fokussieren sollten. Dies könnte insbesondere für Kategorien wie *Systemverhalten*, *Operation* und *Einführung* von Vorteil sein. Eine weitere Verbesserung könnte durch die Schulung der Rater in Bezug auf die Taxonomie und die Erstellung detaillierterer Definitionen samt Negativbeispielen für die Kategorien erreicht werden.

Dass es bei den Fokusgruppen auch nicht einzigartige Bedarfe gab zeigt, dass die Teilnehmer nicht aufmerksam darauf achten, was die anderen Teilnehmer sagen und es so zu doppelten Bedarfsäußerungen kam. Dies zeigt, dass bei Fokusgruppen die Identifizierung von einzigartigen Erklärungsbedarfen auch wichtig ist, um diese zu berücksichtigen.
Ein möglicher Grund für die wenigen Erklärungsbedarfe pro Teilnehmer bei den Fokusgruppen ist, dass die Personen die Erklärungsbedarfe nicht kopieren wollen, wobei sie eventuell einen ähnlichen Erklärungsbedarf zu der selben Funktion gehabt hätten. Aufgrund der schlechten Interrater-Reliabilitäten ist ersichtlich, dass das Verständnis der Taxonomie der Teilnehmer nicht präzise ist und sie viel Subjektivität in die Erklärungsbedarfe interpretieren. Deswegen ist es möglich, dass die Teilnehmer denken, ihre Erklärungsbedarf ist der selbe, wobei er einer anderen Kategorie angehören und auch anders gelöst werden würde, weswegen sie ihre Bedarfe in diesen Fällen dann nicht mehr nennen. Dies spricht ebenfalls dagegen, Fokusgruppen als primäre Erhebungsmethode für Erklärbarkeitsanforderungen in einem Unternehmen durchzuführen.

Die variierenden Interrater-Reliabilitäten zwischen den Versionen mit direkter und mit nachträglicher Taxonomieeinführung innerhalb der einzelnen qualitativen Erhebungsmethode zeigen, dass die Teilnehmer bei der Version mit nachträglicher Taxonomieeinführung ein besseres Verständnis von der Taxonomie haben. Auch unter sich sind sie sich bei

der Fokusgruppe einiger. Der einzige Unterschied zu dieser Form der Taxonomieeinführung ist, dass die Teilnehmer ihre Erklärungsbedarfe vor der Taxonomieeinführung bereits äußern. Dies kann als Grund für die gestiegenen Reliabilitäten gesehen werden.
Da die Interrater-Reliabilitäten für die Interviews am höchsten waren, könnte die Interaktivität und der direkte Dialog zum Teilnehmer der Grund für die bessere Übereinstimmung sein. Die Umfrage erzielte dabei die niedrigsten Übereinstimmungsraten, was vermutlich auf die geringe Anzahl an Erklärungsbedarfen pro Person und die isolierte Art der Umfrage zurückzuführen ist.
Da sowohl bei direkter als auch nachträglicher Taxonomieeinführung mehr einzigartige Bedarfe in jeder Methode erhoben wurden, können auch die gestiegenen Interrater-Reliabilitäten ein Grund dafür sein. Also sollte ein Anforderungsanalyst darauf achten, bei Einführung der Taxonomie für ein besseres Verständnis dieser für die Teilnehmer zu sorgen. Dies hat wohlmöglich Auswirkungen auf die Effektivität der Erhebung.

Die Teilnehmer der Erhebungsmethoden nannten verschiedene Vorschläge zur Erweiterung der Taxonomie von Droste et al. [15].
Redundantes Feature wäre dabei eine geeignete Kategorie, die von der Taxonomie von Droste et al. [15] bisher nicht abgedeckt wird und unter einem Punkt mit Nicht vorhandenes Feature zusammengefasst werden könnte.
Performanz wurde als mögliche Unterkategorie für Systemverhalten genannt. Dies wird bei dieser Taxonomie ebenfalls bisher nicht betrachtet, doch nach dem Codingsystem von Kliti Nikollau [29] wird dieser Punkt als Performanzverschlechterung unter *Unerwartetes Systemverhalten* definiert.
Die genannten Vorschläge Berechtigungen und Barrierefreiheit werden ebenfalls von dem Codingsystem unterstützt. Eine Unterteilung der Kategorie Benutzeroberfläche in grafisch und funktionell wäre unter Umständen für bestimmte Softwaretypen sinnvoll. Für diese Studie und spezifische Software war es aber nicht nötig, da durch die Anwendung des Codingsystems die mehrfach vorkommenden Erklärungsbedarfe auch ohne diese Unterscheidung identifiziert wurden.
Die anderen genannten Kategorien können bereits unter die bestehenden Kategorien eingeordnet werden, dabei oftmals unter Domänenwissen. Diese würden jedoch nicht für alle Softwaretypen dienen und sind daher zur Erweiterung der Taxonomie nicht brauchbar.

Für die Kategorien der Fokusgruppen und Interviews, in denen niedrige Interrater-Reliabilitäten erfasst wurden, wurden Negativbeispiele für die Umfrage aufgeführt, um ein besseres Verständnis dieser für die Nutzer zu gewährleisten. Während *Unerwartetes Systemverhalten* bei der Umfrage eine leicht niedrigere durchschnittliche Übereinstimmung als bei den

Interviews erzielte, war sie dennoch höher als bei den Fokusgruppen. *Algorithmus, Operation* und *Einführung* haben dabei sehr niedrige Übereinstimmungsraten erzielt. *Nicht vorhandenes Feature* befand sich zwischen den beiden anderen Methoden und war auf einem ähnlich hohen Niveau.
Dies zeigt, dass die Einführung von Negativbeispielen bei der Umfrage keine Verbesserungen des Verständnisses der zugehörigen Kategorien erzielt hat. Da jedoch die Median Zeit für die Betrachtung der Taxonomie bei der Umfrage gerade ein mal 0,69 Minuten lang war, ist es sehr wahrscheinlich, dass diese von den Teilnehmern nicht ausreichend lange beachtet wurde. Diese lag mit 4,4 Minuten bei der Fokusgruppe bzw. 3,14 Minuten bei den Interviews deutlich höher. Somit müssen die Effekte der Einführung von Negativbeispielen noch genauer untersucht werden. Möglicherweise ist es hilfreich, detailliert auf die potenziellen Lösungen des Erklärungsbedarfs einzugehen, um dem Nutzer zu zeigen, bei welcher Form der Äußerung die passende Antwort resultiert. Auch könnte eine andere Form der Darstellung der Taxonomie samt Beispielen in der Onlineumfrage die Teilnehmer dazu bewegen, sich intensiver damit zu befassen. Dies würde jedoch den Zeitaufwand erhöhen und die Effizienz der Umfrage somit verringern.
Die Umfrage hatte außerdem die meisten Abbrüche (125) nach der ersten Seite bei den Angaben zur Nutzung der Software und fast genau so viele Abbrüche (124) bei der Vorstellung von Erklärbarkeit. Kombiniert mit dem Feedback von einigen Teilnehmern, dass ihnen die Umfrage zu lang und zu wissenschaftlich sei, kann man davon ausgehen, dass eine Umfrage in einem Unternehmen zum Erheben von Erklärbarkeitsanforderungen in dieser Form nicht gut geeignet ist. Denn möglicherweise gehen die Teilnehmer bei einer Umfrage davon aus, dass sie sich kein Wissen aneignen müssen und nur Felder auswählen müssen. Eine andere Möglichkeit ist, dass die Teilnehmer, da sie auch Arbeitnehmer sind, die erforderliche Zeit dafür nicht aufbringen können und deswegen die Umfrage bei erhöhter Komplexität abbrechen.

7.3 Validity Threats

Folgend werden mögliche Validity Threats benannt. Die Kategorisierung dieser wird mittels der Threats nach Wohlin [41] erfolgen. Dabei wird zwischen *Threats to Construct, Internal, Conclusion* und *Threats to External Validity* unterschieden.

Bei der ersten Fokusgruppe fiel auf, dass die Teilnehmer die Kategorien *Einführung* und *Nicht vorhandenes Feature* oft falsch vergaben. Daraufhin bat der Anforderungsanalyst die Teilnehmer, ihre Wahl noch mal zu überlegen mit dem Hinweis, die Bedarfe getrennt von ihren Wünschen und nur das explizit Gesagte zu kategorisieren. Während dies dann auch bei

den anderen Erhebungen praktiziert wurde, hat es bei der Fokusgruppe etwas mehr Zeit benötigt. Dies hat Auswirkungen auf die Effizienz dieser Methode, da es den Personalaufwand erhöht hat. Die Vergleichbarkeit mit dieser Fokusgruppe ist in gewissem Maße bedroht (Threat to Construct Validity).

Die Kategorisierung der Ergebnisse wurde von einer einzelnen Person vorgenommen, die zugleich Nutzer der Personalverwaltungssoftware ist. Trotz der Bemühung, die Kategorisierung so objektiv wie möglich durchzuführen, lässt sich eine Beeinflussung durch persönliche Subjektivität nicht ausschließen. Dies könnte potenziell zu Verzerrungen in den Ergebnissen führen und sollte bei der Interpretation berücksichtigt werden (Threat to Construct Validity).

Die Bestimmung einzigartiger Erklärungsbedarfe birgt das Risiko, durch eine unpassende Wahl der Subkategorien fehlerhaft zu sein. Zwar wurden alle mehrfach auftretenden Erklärungsbedarfe auf ihre Richtigkeit geprüft, doch lässt sich nicht ausschließen, dass weitere zugehörige Bedarfe übersehen wurden. Solche nicht erkannten Zuordnungen könnten dazu führen, dass die Anzahl einzigartiger Erklärungsbedarfe innerhalb einer Methode überschätzt wird (Threat to Construct Validity).

Ein potenzieller Threat to Validity ergibt sich aus den unterschiedlichen Rahmenbedingungen der beiden Fokusgruppen. Während die Teilnehmer der ersten Fokusgruppe an einem gemeinsamen, kleineren Standort arbeiteten und sich bereits kannten, wurde die zweite Fokusgruppe aufgrund geringer Teilnahmebereitschaft an einem anderen Standort mit Personen durchgeführt, die sich untereinander nicht so gut vertraut waren. Diese Unterschiede könnten die Dynamik der Diskussion beeinflusst haben, da sich vertraute Personengruppen möglicherweise offener oder anders verhalten als fremde Gruppen. Diese Variabilität in den Rahmenbedingungen könnte die Vergleichbarkeit der Ergebnisse zwischen den beiden Gruppen beeinträchtigen und sollte bei der Interpretation der Ergebnisse berücksichtigt werden (Threat to Internal Validity).

Die zweite Fokusgruppe wurde zeitlich nach der ersten durchgeführt und zeigte eine sicherere Moderation. Dies könnte die besseren Interrater-Reliabilitäten und höheren Anzahl an einzigartigen Erklärungsbedarfen bei der zweiten Fokusgruppe erklären. Obwohl bei der Planung darauf geachtet wurde, beide Fokusgruppen möglichst einheitlich durchzuführen, ist nicht auszuschließen, dass eine gesteigerte Routine des Anforderungsanalysten die Ergebnisse beeinflusst haben könnte (Threat to Internal Validity).

Die unterschiedlichen Gesamtzeiten der Erhebungsmethoden, trotz

Bemühungen um eine möglichst einheitliche Gestaltung, könnten die Vergleichbarkeit der Ergebnisse beeinträchtigen. So kann es sein, dass mit längerer Dauer der Erhebung die Konzentration der Teilnehmer nachlässt, was sich vor allem auf die letzten Teile der Erhebungen wie der Kategorisierung auswirkt (Threat to Internal Validity).

Die Erhebungen wurden ausschließlich innerhalb eines einzigen Unternehmens in Deutschland durchgeführt. Dies könnte die Generalisierbarkeit der Ergebnisse einschränken, da sie möglicherweise nicht ohne Weiteres auf Unternehmen in anderen Ländern oder global übertragen werden können (Threat to External Validity).

Vergleiche zwischen den qualitativen und quantitativen Erhebungsmethoden hinsichtlich ihrer Effektivität und Effizienz können lediglich für Fokusgruppen, Interviews und Onlineumfragen gezogen werden, da ausschließlich diese Methoden im Rahmen dieser Masterarbeit angewendet wurden. Um ein umfassenderes Bild zu erhalten, wären weiterführende Analysen erforderlich, die zusätzliche Erhebungsmethoden einbeziehen und deren Ergebnisse miteinander vergleichen (Threat to External Validity).

Obwohl bei der Durchführung der Erhebungen darauf geachtet wurde, eine ausreichende Teilnehmerzahl für aussagekräftige Ergebnisse sicherzustellen, besteht die Frage nach der Angemessenheit der Teilnehmerzahl. Die Auswahl von zwei Fokusgruppen mit je sechs Teilnehmern, 18 Interviewteilnehmern und einer Onlineumfrage mit 188 Teilnehmern könnte in Frage gestellt werden, insbesondere im Hinblick auf die Generalisierbarkeit der Ergebnisse. Kleinere Stichproben, wie in den Fokusgruppen und Interviews, könnten Einschränkungen in der Repräsentativität der Ergebnisse mit sich bringen (Threat to External Validity).

Eine potenzielle Bedrohung der Validität besteht darin, dass aufgrund der spezifischen Domäne der untersuchten Software keine allgemeinen Schlussfolgerungen auf andere Softwaretypen gezogen werden können. Es sind weitere Untersuchungen erforderlich, um zu klären, in welchem Maße die Ergebnisse der angewandten Erhebungsmethoden auf Softwarelösungen anderer Domänen, abseits von Personalverwaltungssoftware, übertragbar sind (Threat to External Validity).

Kapitel 8

Zusammenfassung und Ausblick

Dieses Kapitel liefert einen abschließenden Überblick über die Ergebnisse dieser Arbeit. Anschließend folgt ein Ausblick, der auf weitere Anwendungsmöglichkeiten und offene Probleme eingeht.

8.1 Zusammenfassung

Erklärbarkeit ist eine wichtige NFR im Bereich des Software-Engineerings. Damit Nutzern eine gute Bedienung der Software gewährleistet werden kann, ist es hilfreich, deren Erklärungsbedarfe zu erheben. Diese können folgend in Erklärbarkeitsanforderungen umformuliert werden. Aus diesem Grund wurden 2 Fokusgruppen mit je sechs Teilnehmern, 18 Interviews und eine Onlineumfrage mit 188 Teilnehmern bei dem IT- und Beratungsunternehmen adesso SE durchgeführt. Für diese Arbeit wird die dort eingesetzte Personalverwaltungssoftware betrachtet.
Es wurden die Forschungsfragen in Bezug auf die verschiedenen Erhebungsmethoden behandelt, wobei auch deren Effektivität sowie Effizienz in Betracht gezogen wurde.
Die Ergebnisse zeigen, dass Interviews am effektivsten und effizientesten sind, sowohl in Bezug auf die Anzahl der einzigartigen Erklärungsbedarfe als auch hinsichtlich des Personalaufwands. Umfragen sind in beiden Aspekten weniger effektiv, während Fokusgruppen die schlechtesten Ergebnisse liefern.

Die Anwendung einer Taxonomie führte zu einer größeren Anzahl an einzigartigen Erklärungsbedarfen und war insbesondere bei der nachträglichen Einführung der Taxonomie von Vorteil. Bei nachträglicher Einführung der Taxonomie wurde in den verschiedenen Erhebungsmethoden eine erhöhte Übereinstimmung bei der Kategorisierung erzielt, was auf ein besseres Verständnis der Taxonomie bei den Teilnehmern hinweist.

Es wurden von den Teilnehmern der Erhebungsmethoden Vorschläge zur Erweiterung der Taxonomie gemacht, von denen einige in zukünftigen Studien berücksichtigt werden sollten.

Zusammenfassend lässt sich sagen, dass Interviews die bevorzugte Methode für die Erhebung von Erklärbarkeitsanforderungen sind, da sie sowohl bezüglich der Effektivität als auch der Effizienz am vorteilhaftesten sind. Eine nachträgliche Taxonomieeinführung unterstützt die Erhebung von Erklärbarkeitsanforderungen, sollte jedoch in Zukunft weiter präzisiert werden. Die Diskussion deutet auch darauf hin, dass der Einsatz von Negativbeispielen in der Umfrage möglicherweise nicht ausreichend war, um die Verständlichkeit der Taxonomie zu verbessern. Dies regt weitere Forschungen zu diesem Thema an.

8.2 Ausblick

Die in dieser Arbeit evaluierten Erkenntnisse der Fokusgruppen, Interviews und der Umfrage können als Basis dafür dienen, eine Entscheidungsgrundlage bei der Auswahl einer Methode zur Erhebung von Erklärbarkeitsanforderungen zu treffen. Dabei kann das erstellte Konzept mit den in dieser Arbeit erlangten Erkenntnisse zur Durchführung dieser als Ablaufplan dienen.

Die Ergebnisse dieser Arbeit, wie die Wahl von Interviews als qualitative Erhebungsmethode, sollten in weiteren Softwaretypen und in Unternehmen aus anderen Ländern überprüft werden, um deren Gültigkeit zu bestätigen. Dabei können weitere Erhebungsmethoden verglichen werden, wobei die Ergebnisse dieser Arbeit darauf schließen lassen, das Fokusgruppen nicht geeignet sind und außer Acht gelassen werden können.
Die erfassten Erklärungsbedarfe werden nach der Häufigkeit der einzigartigen Bedarfe priorisiert und folglich gebündelt an das Management der Software übermittelt. Ein nächster Schritt ist dabei zu entscheiden, für welche Erklärungsbedarfe Erklärungen geliefert werden und werden können.

Ein weiterer wichtiger Forschungsbereich könnte die Weiterentwicklung und Verfeinerung der verwendeten Taxonomie darstellen. Es wurde festgestellt, dass die Einführung der Taxonomie zu besseren Ergebnissen führte, was auf die Notwendigkeit hinweist, diese Taxonomie weiter zu präzisieren und in der Praxis zu testen. Eine mögliche Weiterentwicklung könnte darin bestehen, Negativbeispiele und spezifische Anwendungsfälle intensiver in die Taxonomie zu integrieren, um so die Verständlichkeit und Anwendbarkeit weiter zu verbessern. Die Integration dieser konnte aufgrund der kurzen Betrachtungszeit seitens der Umfrageteilnehmer nicht ausreichend analysiert werden. Die Vorschläge für weitere Kategorien müssen für die Taxonomie von Droste et al. [15] weiter analysiert werden, um diese zu erweitern und somit eine detailliertere Kategorisierung von Erklärungsbedarf ermöglicht wird. Eine mögliche Untersuchung könnte sein, das Codingsystem nach Kliti Nikollau [29] für die Erhebung von Erklärbarkeitsanforderungen zu nutzen, da diese die Kategorisierung präzisiert und zu einer effektiveren Erhebung führen kann.

Die erstellten Datensätze beinhalten weitere Informationen wie Arbeitsrolle, Alter oder Nutzungsdauer und Häufigkeit der Software, die für diese Studie nicht weiter untersucht wurden. Diese stehen eventuell mit der Äußerung von Erklärungsbedarf in einem Zusammenhang, wofür Korrelationsanalysen angewendet werden könnten. Solche Zusammenhänge könnten dazu beitragen, die Erhebung von Erklärbarkeitsanforderungen zu optimieren.

Anhang A

Anhang

A.1 Statistische Auswertung zwischen einem Mehrheitsentscheid der Fokusgruppen und dem Anforderungsanalysten

Kategorie	Mit dir. Taxonomieeinf.			Mit nachträgl. Taxonomieeinf.		
	Übereinst.	Cohen's K	n	Übereinst.	Cohen's κ	n
Systemverhalten						
Systemverhalten (Oberkat.)	0%	0	4	80,00%	0,87	5
Unerwartetes Systemverh.	33,33%	0,47	3	0%	0	1
Bugs & Abstürze	0%	0	1	50,00%	0,65	2
Algorithmus	-	-	0	50,00%	0,65	2
Konsequenzen	0%	0	1	0%	0	1
Interaktion						
Interaktion (Oberkat.)	10,00%	0,12	10	40,00%	0,46	10
Operation	0%	-0,04	2	100%	1,00	2
Navigation	-	-	0	100%	1,00	2
Einführung	0%	0	9	0%	0	5
Nicht vorhandenes Feature						
Nicht vorh. Feature (Oberkat.)	16,67%	0,23	6	57,41%	0,66	7
Domänenwissen						
Domänenwissen (Oberkat.)	35,71%	0,33	14	11,11%	0,15	9
Systemspez. Elemente	11,11%	0,14	9	14,29%	0,20	7
Begrifflichkeiten	16,67%	0,19	6	0%	0	2
Business						
Business (Oberkat.)	-	-	0	-	-	0
Geheimhaltung & Sicherheit						
Geheimh. & Sicherh. (Oberkat.)	100%	1,00	2	100%	1,00	2
Geheimhaltung	-	-	0	50,00%	0,65	2
Sicherheit	100%	0,37	4	0%	0	1
Benutzeroberfläche						
Benutzeroberfläche (Oberkat.)	25,00%	0,34	4	33,33%	0,42	6
Gesamt Oberkategorien	**40,00%**	**0,29**	25	**60,71%**	**0,54**	28
Gesamt Unterkategorien	**28,00%**	**0,23**	25	**50,00%**	**0,45**	28

Tabelle A.1: Interrater-Reliabilitäten zwischen einem Mehrheitsentscheid und dem Anforderungsanalysten bei beiden Fokusgruppen, wobei n für die Anzahl der relevanten Erklärungsbedarfe steht

A.2 Übersicht über die in dem Codingsysteme nach Kliti Nikollau erweiterten Subkategorien

time off	role designation	role description	personal data
Report Centre	file upload	french	employee files
site loading	search function	employee details	review & performance
employee profile	development plan	active candidate	language
sap jam	todo tasks	quick actions	notifications
position information	absences	sport subsidy	highest education
error message	organigram	menu	vacation
loading	logout	search results	family subsidy
tutorial	teamportfolio	language settings	payroll forms
career path	login	data correction	refusal
onboarding	profile picture	downtimes	information hiding
session timeout	payment	self evaluation	for you today
study certificates	shortcuts	settings	adtime
profile header	workflows	Permissions	mobile app
working place	links	filtering	functionalities
company id	url	data safety	absences overview
data	reminders	favorites	user groups
quick actions - shortcuts	relationships	dropdown	contract
job classification	quick actions - shortcuts - links	input	shortcuts - menu
admin center	pdf export	email security	as of today
favorites - shortcuts	overview	quick actions - shortcuts - menu	data editing
Corporate Benefits	proof of activities	daily sight	recruiting page
successfactors - adpeople	documents	workpackages	differences between systems
fields	location	certificate	contact to HR
process changes	shortcuts - links	subsidies	allignment
application	retired employees	arrangement	chatbot
mouseoverview	suggestions	support	employee evaluation
emergency contact	info buttons	breadcrumbs	explanations
trainings	tooltips	bookings	working time
customer working time	SAP success factors	multi select	development plan - ecampus
prefill	myforms	navigation	navigation possibilities
glasses subsidy	personal interviews	separate windows	absences - time off
calender	vitamin a	data transfer	changes
Sap	interaction concept	shares	adding employees
yearly formular	career path - learning activities	product feedback	learning activities
overtime	welcome text	icons	buttons
quick actions - menu	personalization	links - intern function buttons	tip
notifications - tasks	system	labels	header
responsive	info mails	scrolling	project name
consistent design	quick actions -shortcuts - links - menu	design	starting page
font	hierarchy	auditor status	validated
dependents	faq	social security number	health insurance type
view my profile - adesso profiler	reward statements	adpeople	pending workflows
employee profile - view my profile - employee file	fit	ACE	job requisitions
employee self evaluation	examples	travel times	jam report
adtravel	adup	bookr	shortcuts - menu - links
organigram - azure ad			

Tabelle A.2: Liste der verwendeten Subkategorien

Literaturverzeichnis

[1] N. C. Alflen, L. Cássia, E. P. Prado, and A. Grotta. Using combined techniques for requirements elicitation: A brazilian case study. In *ICEIS (2)*, pages 241–248, 2021.

[2] N. C. Alflen, E. P. Prado, and A. Grotta. A model for evaluating requirements elicitation techniques in software development projects. In *ICEIS (2)*, pages 242–249, 2020.

[3] F. Anwar and R. Razali. A practical guide to requirements elicitation techniques selection-an empirical study. *Middle-East Journal of Scientific Research*, 11(8):1059–1067, 2012.

[4] J. Bohnstedt. Untersuchung des einflusses von domänenwissen auf den erklärungsbedarf der nutzenden von software. *Bachelorarbeit, Gottfried Wilhelm Leibniz Universität Hannover, FG Software Engineering*, 2024.

[5] E. Börger, B. Hörger, D. Parnas, and D. Rombach. Requirements capture, documentation, and validation. In *Dagstuhl Seminar*, volume 99241, 1999.

[6] W. Brunotte, L. Chazette, V. Klös, and T. Speith. Quo vadis, explainability?–a research roadmap for explainability engineering. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pages 26–32. Springer, 2022.

[7] V. R. B. G. Caldiera and H. D. Rombach. The goal question metric approach. *Encyclopedia of software engineering*, pages 528–532, 1994.

[8] N. M. Carod and A. Cechich. Cognitive profiles in understanding and prioritizing requirements: a case study. In *2010 Fifth International Conference on Software Engineering Advances*, pages 341–346. IEEE, 2010.

[9] L. Chazette, W. Brunotte, and T. Speith. Exploring explainability: a definition, a model, and a knowledge catalogue. In *2021 IEEE 29th international requirements engineering conference (RE)*, pages 197–208. IEEE, 2021.

[10] L. Chazette, V. Klös, F. Herzog, and K. Schneider. Requirements on explanations: a quality framework for explainability. In *2022 IEEE 30th International Requirements Engineering Conference (RE)*, pages 140–152. IEEE, 2022.

[11] L. Chazette, J. Klünder, M. Balci, and K. Schneider. How can we develop explainable systems? insights from a literature review and an interview study. In *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering*, ICSSP'22, page 1–12, New York, NY, USA, 2022. Association for Computing Machinery.

[12] L. Chazette and K. Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25(4):493–514, 2020.

[13] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[14] N. Dehn. Anforderungen verschiedener stakeholdergruppen an die stimmungsanalyse in softwareprojekten. *Masterarbeit, Gottfried Wilhelm Leibniz Universität Hannover, FG Software Engineering*, 2023.

[15] J. Droste, H. Deters, M. Obaidi, and K. Schneider. Explanations in everyday software systems: Towards a taxonomy for explainability needs. *arXiv preprint arXiv:2404.16644*, 2024.

[16] P. Engelbrektsson, Ö. Yesil, and I. Karlsson. Eliciting customer requirements in focus group interviews: can efficiency be increased. In *7th International Product Development Management Conference*, pages 29–30. Citeseer, 2000.

[17] T. Fehlmann and E. Kranich. Customer-driven software product development software products for the social media world–a case study. In *Systems, Software and Services Process Improvement: 20th European Conference, EuroSPI 2013, Dundalk, Ireland, June 25-27, 2013. Proceedings 20*, pages 300–312. Springer, 2013.

[18] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[19] R. Fuchs. Erhebung von erklärbarkeitsanforderungen für machine-learning-methoden von blockheizkraftwerken. *Bachelorarbeit, Gottfried Wilhelm Leibniz Universität Hannover, FG Software Engineering*, 2023.

[20] I. Hadar, P. Soffer, and K. Kenzi. The role of domain knowledge in requirements elicitation via interviews: an exploratory study. *Requirements Engineering*, 19:143–159, 2014.

[21] A. Herrmann, E. Knauss, R. WeißBach, R. Fahney, T. Gartung, J. Glunde, A. Hoffmann, and U. Valentini. *Requirements Engineering und Projektmanagement*. Springer, 2013.

[22] L. Jiang, A. Eberlein, and B. H. Far. Combining requirements engineering techniques-theory and case study. In *12th IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS'05)*, pages 105–112. IEEE, 2005.

[23] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender. Explainability as a non-functional requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 363–368. IEEE, 2019.

[24] R. A. Krueger. *Focus groups: A practical guide for applied research*. Sage publications, 2014.

[25] D. Kupczyk. Automatisierte detektion von erklärungsbedarf in nutzerfeedback zu software. *Masterarbeit, Gottfried Wilhelm Leibniz Universität Hannover, FG Software Engineering*, 2023.

[26] T. Kurtz. Entwicklung einer software zur extrahierung und analyse von reviews aus app stores. *Bachelorarbeit, Gottfried Wilhelm Leibniz Universität Hannover, FG Software Engineering*, 2023.

[27] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[28] D. Mishra, A. Mishra, and A. Yazici. Successful requirement elicitation by combining requirement engineering techniques. In *2008 First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT)*, pages 258–263. IEEE, 2008.

[29] K. Nikollau. Identification of criteria for classifying the explanation needs in apps. *Masterarbeit, Gottfried Wilhelm Leibniz Universität Hannover, FG Software Engineering*, 2024.

[30] M. Obaidi. Dataset: Gold standard dataset for explainability need detection in app reviews., Sept. 2024. `https://doi.org/10.5281/zenodo.11522828`.

[31] C. Pacheco, I. García, and M. Reyes. Requirements elicitation techniques: a systematic literature review based on the maturity of the techniques. *IET Software*, 12(4):365–378, 2018.

[32] H. Partsch. *Requirements-Engineering systematisch: Modellbildung für softwaregestützte Systeme*. Springer-Verlag, 2010.

[33] K. Pitula and T. Radhakrishnan. On eliciting requirements from end-users in the ict4d domain. *Requirements Engineering*, 16:323–351, 2011.

[34] K. Pohl and C. Rupp. *Basiswissen Requirements Engineering : Aus- und Weiterbildung nach IREB-Standard zum C̈ertified Professional for Requirements Engineering:̈ Foundation Level nach IREB-Standard.* dpunkt.verlag, Heidelberg, 2015.

[35] C. Rupp, M. Simon, and F. Hocker. Requirements engineering und management. *HMD Praxis der Wirtschaftsinformatik*, 46(3):94–103, 2009.

[36] N. Sabahat, F. Iqbal, F. Azam, and M. Y. Javed. An iterative approach for global requirements elicitation: A case study analysis. In *2010 International Conference on Electronics and Information Engineering*, volume 1, pages V1–361. IEEE, 2010.

[37] K. Stapel and K. Schneider. Flow-methode-methodenbeschreibung zur anwendung von flow. *arXiv preprint arXiv:1202.5919*, 2012.

[38] S. Tiwari and S. S. Rathore. A methodology for the selection of requirement elicitation techniques. *arXiv preprint arXiv:1709.08481*, 2017.

[39] M. Unterbusch, M. Sadeghi, J. Fischbach, M. Obaidi, and A. Vogelsang. Explanation needs in app reviews: Taxonomy and automated detection. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 102–111. IEEE, 2023.

[40] J. Voges. Ermittlung von erklärbarkeitsanforderungen zur erhöhung der nutzerakzeptanz eines stimmungsanalysetools. *Masterarbeit, Gottfried Wilhelm Leibniz Universität Hannover, FG Software Engineering*, 2024.

[41] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering.* Springer Science & Business Media, 2012.

[42] M. Younas, D. Jawawi, I. Ghani, and R. Kazmi. Non-functional requirements elicitation guideline for agile methods. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(3-4):137–142, 2017.