3 Forschungsüberblick und Gegenstandsbestimmung

Das folgende Kapitel soll den Forschungsstand zu Mensch-Maschine-Interaktionen in der Linguistik im Überblick zusammenfassen sowie im Speziellen und detaillierter den interdisziplinären Forschungsstand zu Smart Speakern unter verschiedenen Gesichtspunkten darstellen. Das Kapitel liefert daher zunächst einen phänomenorientierten Überblick über Forschung, die sich primär aus sprachwissenschaftlicher Sicht mit dem Austausch zwischen Mensch und Maschine beschäftigt hat (Kap. 3.1). Dabei werden vorwiegend nicht Smart Speaker, sondern andere Dialogsysteme wie Chatbots, Embodied Conversational Agents, humanoide Roboter und telefonbasierte Systeme beleuchtet. Daran anschließend wird der interdisziplinäre Forschungsstand zu Smart Speakern unter verschiedenen Aspekten zusammengefasst. Es wird dazu zunächst knapp die historische Entwicklung und die technische Verarbeitung und Prozessierung der heutigen Systeme nachgezeichnet (Kap. 3.2.1), ehe auf die Funktionsweise und die technischen Hintergründe selbst eingegangen wird (Kap. 3.2.2). Eine detaillierte informatische Beschreibung ist dabei entbehrlich und auch gar nicht im Rahmen dieser Arbeit zu leisten. Wesentlich ist aber ein überblicksartiges Verständnis der ablaufenden Prozesse. Anschließend an die technische Beschreibung werden kommunikations- und medienwissenschaftliche Arbeiten vorgestellt, in denen Smart Speaker ausgehend von ihren Funktionen als Intermediäre und als Plattformen konzeptualisiert werden (Kap. 3.2.3). Dieses Kapitel liefert insofern auch einen Überblick über den Funktionsumfang der Geräte, auch wenn die vorliegende Arbeit auf deren spezifische Interfaces fokussiert. Darauf folgend werden Studien zusammengefasst, die bereits zur Nutzung und Aneignung von Smart Speakern aus unterschiedlichen disziplinären Perspektiven und mit einer noch überschaubaren Bandbreite an verschiedenen Methoden geforscht haben, darunter zentral die schwerpunktmäßig linguistisch arbeitenden Untersuchungen (Kap. 3.2.4). Ein besonderes Augenmerk soll im Anschluss auf solche Arbeiten gelegt werden, die Smart Speaker unter Gender-Gesichtspunkten beleuchtet haben (Kap. 3.2.5). Dabei wird insbesondere auf die Wahl der Stimmfarbe und die ausgehandelte Rollenverteilung zwischen dem Smart Speaker und den Nutzer*innen abgehoben. Abschließend sollen die im Fokus der öffentlichen Aufmerksamkeit stehenden Themen Datenschutz, Datenverwertung und Privatsphäre anhand der dazu erschienenen einschlägigen Studien betrachtet werden (Kap. 3.2.6). Die zum Zeitpunkt der Studie gängigen und genutzten Smart Speaker werden in Kap. 3.3 genauer beschrieben, um einen Einblick in den Aufbau und die Charakteristika der einzelnen Systeme von Amazon, Google und Apple zu bekommen.

https://doi.org/10.1515/9783111574332-003

3.1 Mensch und Maschine in der empirischen Linguistik

Der Austausch zwischen Mensch und Maschine ist von der Sprachwissenschaft in mindestens zweierlei Hinsicht betrachtet worden: Einerseits ist die Subdisziplin der Computerlinguistik selbst wesentlicher Bestandteil der vorwiegend von der Informatik geprägten Entwicklung von Interfaces, schriftlich oder mündlich: Die Hervorbringung von *Speech Processing* und *Natural Language Processing* ist nicht vorstellbar, ohne eine Beteiligung der Linguistik. Andererseits, und dies soll Fokus der folgenden Ausführungen sein, auch wenn die Unterscheidung nicht bei allen Arbeiten trennscharf vollzogen werden kann, hat die Sprachwissenschaft empirisch den Austausch selbst untersucht, ohne dabei primär die technologische Weiterentwicklung als Ziel der Forschung zu positionieren. Phänomenbezogen besteht dabei eine ‚Verwandtschaft' von Smart Speakern, die selbst bislang nur begrenzt linguistisch untersucht wurden (siehe Kap. 3.2.4), mit anderen Dialogsystemen, insbesondere Chatbots, Embodied Conversational Agents und humanoiden Robotern (vgl. Lotze 2018: 29) sowie mit telefonbasierten Sprachdialogsystemen (vgl. Thar 2015). Die nachfolgenden Ausführungen sind ein Überblick über Erkenntnisse aus der explizit interaktionslinguistischen Untersuchung des Dialogs mit Maschinen, bevor im darauf folgenden Kapitel der interdisziplinäre Forschungsstand zu Smart Speakern einschließlich einer Übersicht über die Funktionsweise und den technischen Hintergrund präsentiert wird.

3.1.1 Chatbots

Als eine „neue Form der Dialogizität" beschreibt die Sprachwissenschaftlerin Netaya Lotze den Dialog mit Chatbots und mit Embodied Conversational Agents (Lotze 2020: 363). Letztere basieren auf Chatbot-Systemen, sind aber umfangreich erweitert, insbesondere durch multimodale, personalisierte Avatare, über die auch eine basale Form von Emotion ausgedrückt werden kann (vgl. Lotze 2016: 39).[75] Aus einer linguistischen Perspektive setzt sich Lotze (2016) in einer umfangreichen Arbeit mit Chatbots[76] auseinander und zeigt vier methodische Zugänge zu diesen auf, die in unterschiedlichen Teilgebieten der Linguistik verwurzelt sind: Neben dem Zugriff, den die Konversationsanalyse ermöglicht, arbeitet sie mit dem Modell des „interaktiven Alignment" aus der Psycholinguistik, dem Konzept des „Computer-Talk" (Zoeppritz 1985; Krause/Hitzenberger 1992; Fischer

75 Siehe auch Krummheuer (2010: 81) und Kap. 3.1.2.

76 Zu *Social Bots*, automatisierten Accounts auf Social Media-Plattformen, siehe Kaerlein (2018).

2006),[77] das an der Schnittstelle zwischen Linguistik und Informationswissenschaft angesiedelt ist, sowie dem eher ‚kernlinguistischen' Konzept der Dialogkohärenz. Lotze (2016: 90–91) unterscheidet weitere nicht-linguistische Ansätze, die relevant sind (insbesondere informatische Forschungen zu Design und Usability und soziologische bzw. psychologische Ansätze zur Untersuchung sozialer Folgen und Einflussfaktoren). Sie betont, dass die Analyse von Mensch-Maschine-Interaktion immer eine auf mehreren Ebenen interdisziplinäre Unternehmung ist (vgl. Lotze 2016: 86).

Lotze entwickelt in ihrer Arbeit einen Untersuchungsrahmen, der auf den Erkenntnissen der Konversationsanalyse basiert und als Grundlage für die weiteren Analysen dient: Wesentliche Parameter sind dabei die Dialoggliederung, die Bestimmung von Rahmen- und Adjazenzstrukturen, die Rolle von Störungen und Reparaturen sowie der Einfluss der Form der technischen Vermittlung (vgl. Lotze 2016: 106). Diesen Rahmen wendet sie in konversationsanalytischen Auswertungen, Analysen zur Dialogkohärenz und bei der Betrachtung mit der Perspektive des Computer Talk an. Allein ihre Erkenntnisse aus der konversationsanalytischen Betrachtung der Mensch-Maschine-Dialoge sind sehr reichhaltig. So stellt sie fest, dass Turns des Chat-Systems länger sind als die Nutzer*innen-Turns und auch mehr TCUs beinhalten (vgl. Lotze 2016: 234). Dieser Befund korrespondiert mit den Erkenntnissen von Cyra/Pitsch (2017a; 2017b), die anhand eines Embodied Conversational Agents (ECA)[78] die inkrementelle Erweiterung von Nutzer*innen-Turns beleuchten und dabei von der Feststellung ausgehen, dass längere und inkrementell expandierte Turns für die Dialogsysteme eine Herausforderung darstellen.[79] Für einen gelingenden Dialog sind also kürzere Beiträge vorteilhaft. Die konventionalisierte Rahmung der von Lotze untersuchten Chatbot-Dialoge, z. B. durch Begrüßung und Verabschiedung, wird eher von den Bots eingehalten, weniger aber durch die Nutzer*innen (vgl. Lotze 2016: 239–240; siehe auch Wuenderlich/Paluch 2017). Nutzer*innen hingegen arbeiten häufig mit Adjazenzellipsen, die einen semantischen Anschluss innerhalb des Dialogs herstellen (sollen); diese jedoch erkennen die analysierten Chatbots kaum und verwenden sie auch nicht selbst (vgl. Lotze 2016: 242). Im Vergleich mit anderen Strategien aus der computervermittelten schriftlichen Kommunikation wie der Verwendung von Emoticons und Emojis, Buchstabeniteration, kreativer Satzzeichengebrauch oder Reduktionsformen (siehe etwa Androutsopoulos/Busch 2020) stellt sie fest, dass diese in ihren Daten kaum vorhanden sind, und sieht eine Ursache dafür in der Standardnähe der Beiträge

77 Ausführlicher dazu siehe Kap. 6.1.2.1.

78 Siehe dazu Kap. 3.1.2.

79 Die Studien fokussieren sich in einem „Wizard-of-Oz"-Szenario auf die Entstehung und Klassifikation der Expansionen sowie darauf aufbauende Vorschläge für Design-Lösungen.

des Chat-Systems, sodass „im Zuge dynamischer Anpassungsprozesse“ auch die User*innen solche sprachlichen Formen selten übertrügen (vgl. Lotze 2016: 251–252). Lotze (2016: 247) zeigt aber auch, dass die Orientierung an sprachlichem Material, das (den Auffassungen der Nutzer*innen nach) leicht von den Chat-Systemen verarbeitet (geparsed) werden kann, eine wesentliche Reparaturstrategie ist; es könnte entsprechend auch eine Strategie sein, Ausdrücke aus der zwischenmenschlichen Chat-Kommunikation in diesem Zusammenhang zu vermeiden – auch, weil der Konventionalisierungsgrad hierbei noch deutlich geringer ist als im standardsprachlichen Bereich.

Psycholinguistisch arbeitet Lotze mit dem Konzept des „interaktiven Alignment“ (Pickering/Garrod 2004), in dem die im Dialog verwendeten sprachlichen Konstruktionen mit den kognitiven Repräsentationen in Verbindung gebracht werden – die Wiederaufnahme sprachlicher Strukturen (u. a. auf syntaktischer, lexischer, prosodischer und multimodaler Ebene) ist ein Zeichen für ein solches Alignment, d. h. für die „Angleichung“ kognitiver Repräsentationen. Lotze (2016: 257–262) weist nach, dass in Dialogen mit Chatbots weder auf lexikalischer noch auf syntaktischer Ebene ein vergleichbar häufiges Alignment stattfindet wie in Mensch-Mensch-Interaktionen. Im Dialog mit Chatbots kommt Alignment weniger häufig zustande und ist wesentlich seltener über mehrere Dialogzüge persistent. Neuere Dialogsysteme zeigen zwar höhere Alignment-Werte als ältere Systeme, doch eine Reihe von Störfaktoren und Variablen scheint das Auftreten der Alignments in HCI weiterhin niedrig zu halten (vgl. Lotze 2016: 281–282). Im Hinblick auf Kohärenz und Kohäsion kann Lotze zeigen, dass die Chat-Systeme einen deutlich niedrigeren Anteil kohärenter Anschlüsse produzieren als in verglichenen Mensch-Mensch-Interaktionen, in denen sämtlich Kohärenz hergestellt wird. Dies ist umso bemerkenswerter, als die Anzahl der von den Systemen verwendeten Kohäsionsmittel sehr hoch ist und teilweise sogar höher ist als die der Nutzer*innen (vgl. Lotze 2016: 312). Fehlender Common Ground, fehlende semantische Referenzdefinitionen und nicht ausdefinierte thematische Progression in Frames und Skripts sowie mangelhafte Verarbeitung von Konjunktoren und turnübergreifenden Konnektiva sind dafür die Hauptursache – sie liegen alle auf der Ebene „Tiefenstruktur“ der Chat-Systeme (Lotze 2016: 314–315).

Netaya Lotze zeigt mit ihrer Monografie umfangreiche Anknüpfungspunkte für die Untersuchung von Mensch-Maschine-Dialogen auf: Unterschiede, die durch die verbale Bedienweise der Benutzerschnittstelle entstehen, sind ebenso relevant wie die Frage danach, wie die sprachlichen Besonderheiten im Dialog mit Chatbots sich in die Praxis der Nutzer*innen einbetten, inwieweit sie in die sprachlichen Register übergehen oder diese beeinflussen und zu einer Bewertung der Systeme insgesamt beitragen. Lotze (2018: 31) stellt fest: „Wenn [...] ein Mensch mit einem künstlichen Gegenüber in Interaktion tritt, ist das keine Kom-

munikation unter Gleichen“. Wie in Kap. 2.2 dargestellt, soll bei einem Austausch zwischen Mensch und Maschine wie Lotze ihn beschreibt, in der vorliegenden Arbeit nicht von „Interaktion“ die Rede sein. Denn auch wenn das Design der Dialogsysteme (durch die verbale Benutzerschnittstelle bei Smart Speakern noch einmal mehr) die „Illusion eines menschenähnlichen Gegenübers [unterstützt] und [...] dadurch soziale Wirkungen [evoziert]“ (Lotze 2016: 73; siehe auch Bucher 2014), zeigt sich auch auf sprachlicher Ebene die Bearbeitung wesentlicher Differenzen zwischen Mensch und Maschine. Lotze (2018: 30) unterscheidet passend zwischen „Dialogkompetenz und -performanz“ und hat in ihren Arbeiten gezeigt, dass bei Chat-Systemen durchaus ein Grad von Dialogperformanz vorliegt, sie aber „von einer eigenen Dialogkompetenz weit entfernt“ sind (Lotze 2018: 30). Vielmehr führen aufgebaute Illusionsformen eines menschlichen Gegenübers zu übersteigerten Erwartungen und in besonderer Weise zu Brüchen im Dialog (vgl. Lotze 2018: 46).[80] Als Perspektive verweist Lotze (2018: 46) weniger auf weitere Annäherung zwischen Mensch-Mensch- und Mensch-Maschine-Dialogen, sondern vielmehr auf das wechselseitige Erlernen einer „anwendungsadäquate[n] Interaktionsform“. Zu untersuchen, ob und wie sich dieser Prozess vollzieht, ist wesentlicher Teil des Ansinnens der vorliegenden Arbeit.

3.1.2 Embodied Conversational Agents

Eine erweiterte Form von Chatbots sind Embodied Conversational Agents (ECAs). Sie haben prinzipiell nach Cassell (2000: 29) folgende Eigenschaften:

- the ability to recognize and respond to verbal and nonverbal input
- the ability to generate verbal and nonverbal output
- the ability to deal with conversational functions such as turn taking, feedback, and repair mechanisms
- the ability to give signals that indicate the state of the conversation, as well as to contribute new propositions to the discourse

80 Es soll dabei nicht übersehen werden, dass es auf Social Media-Plattformen wie Twitter durchaus gelingt, Bots einzusetzen, die von menschlichen Accounts nur schwer zu unterscheiden sind (siehe dazu Kaerlein 2018), was bereits eine ethische Debatte über die Pflicht zur Identifikation nicht-menschlicher Benutzerkonten ausgelöst hat (siehe Graff 2016; Williams 2018). Diese werden jedoch in durch die Plattformen präfigurierten und stark konventionalisierten Umgebungen eingesetzt, in denen interaktive Vorgänge grundlegend anders gesteuert werden (Aiello et al. 2012; Howard/Woolley/Calo 2018). In diesen ist, so argumentiert Kaerlein (2018), die Ununterscheidbarkeit von Mensch und Maschine gerade eine Folge der spezifischen medialen Bedingungen, unter denen die Kommunikation stattfindet.

Eine Übersicht zu ECA-Design-Projekten findet sich bei Cassell/Sullivan/Prevost (2000) und bei Cassell et al. (2000). ECAs wurden intensiv in der bereits andiskutierten Arbeit von Krummheuer (2010) untersucht. Sie geht insbesondere der Frage nach, ob der Austausch mit ECAs als Interaktion bezeichnet werden kann. Diese Frage verneint sie und charakterisiert den Austausch als hybrid und ambig zugleich.[81] Krummheuer stößt dabei auch sprachlich auf interessante Strategien, auch wenn die Arbeit sich mit einem primär soziologischen Interesse den ECAs widmet. Ihre Befunde von 2010 und 2011 stehen im Einklang mit den Erkenntnissen von Lotze: Die Orientierung auf Konventionalisierungen und rituelle Klammern des ECA ist stärker als die des Menschen (vgl. Krummheuer 2010: 187). Sie kann ebenfalls an der Schnittstelle von „Spiel, Provokation und Test" (Krummheuer 2010: 305) liegende Formen von unangemessenem und beleidigendem Verhalten feststellen, mit dem Nutzer*innen die technische Beschaffenheit des Geräts betonen und es ‚herausfordern'. Diese Situationen werden vom Publikum in den Daten Krummheuers als unterhaltsam, aber auch als irritierend aufgefasst und teilweise mit Lachen bzw. Kichern kommentiert. Krummheuer betont hier, aber auch z. B. mit Bezug auf Störungen und Reparaturen, dass der Beitrag der Nutzer*innen in ihren Settings immer ein interaktional und situativ ausgehandelter Beitrag ist, der unter Beobachtung der Umstehenden zustande kommt, während der ECA diese Prozesse nicht beobachten kann – der direkte Dialog ‚nur' mit dem ECA ist also nur eine kleine Facette des Austauschs. Der ECA seinerseits simuliert eine Form „intersubjektiven Verstehens" (Krummheuer 2010: 261), die aber (wie bei Chatbot-Systemen) auf Basis von Inputs und Skripts erfolgt und insofern keine interaktionale Aushandlung darstellt.

Die Mündlichkeit, die bei der Steuerung zum Einsatz kommt, bringt aber auch weitere Spezifika des Austauschs mit sich. So ist z. B. die Aushandlung des Rederechts und die Gestaltung von Sprecher*innen-Wechseln ebenfalls Teil des Austauschs. Krummheuer (2010: 209) beobachtet dabei v. a. eine Anpassung der Nutzer*innen, auf die längere Pausen und eine sehr geringe Überlappungsrate hindeuten. Auch machen die Nutzer*innen die Produktionsplanung des nächsten Beitrags gegenüber anderen Anwesenden *accountable* (vgl. Krummheuer 2010: 208), was ebenfalls für eine ‚Anpassung' der menschlichen Akteur*innen spricht.

In neueren Arbeiten zu ECAs argumentiert Cassell (2020), allerdings weniger linguistisch fundiert und eher aus einer Design-Perspektive, dass mit neueren Systemen durchaus eine Form von sozialer Beziehung zwischen Mensch und Maschine entstehen kann. Bei der gemeinsamen Bearbeitung von Aufgaben, im Beispiel von Cassell Lesen- und Schreiben-Lernen, habe sie beobachtet, „that they

81 Siehe Kap. 2.2.3.

[Conversational Agents, T.H.] do in fact improve relations between people and the systems“ (Cassell 2020: 17) und dass sie auch wechselseitige Erfolge bei der Bearbeitung der Lese- und Schreib-Aufgaben als gemeinsam bearbeitetes Projekt erzielen konnten. Allerdings fehlt eine sequentielle Betrachtung der Dialoge in alltäglichen Umgebungen; Cassell stützt sich auf Experimente aus Settings, die primär aus dem Bereich der Design-Evaluation stammen und auch einen vortheoretischen Begriff von „social bond“ verwenden. Gleichwohl bleibt die Frage, inwieweit eine Beziehung zu einem ECA eingegangen werden kann, relevant, wie auch Studien zeigen, die nicht auf schriftliche Chats oder ECAs schauen, sondern verkörperte, menschenähnliche Robotik in den Blick nehmen.

3.1.3 Humanoide Roboter

Die Ergebnisse von Lotze und Krummheuer zu Chatbots bzw. ECAs lassen sich gut an die auf sprachliche Aspekte fokussierenden Arbeiten zu Dialogen mit humanoiden Robotern anknüpfen. Auch hier zeigt sich, dass der Dialog mit einem Roboter strukturelle Unterschiede im Vergleich mit Mensch-Mensch-Interaktionen aufweist (vgl. etwa Habscheid et al. 2020: 174; Fischer 2011; Pitsch/Gehle/Wrede 2013). Um diese herauszuarbeiten, ist eine Reihe an Studien unterschiedlicher Art mit „museum-guide-robots“ durchgeführt worden. Solche gelten als eine der „principal areas applications considered in research on human-robot interaction“ (Kuzuoka et al. 2008: 201). Diese „real-world-settings“ ermöglichen es einerseits, Erkenntnisse über die Gestaltung von Dialogen mit Robotern zu gewinnen, die aus Design-Perspektive verwertet werden können. Sie ermöglichen aber gleichsam das Gewinnen von Erkenntnissen über zwischenmenschliche Interaktion (vgl. Pitsch 2015: 29–30). In diesen Settings stoßen die Roboter auf Herausforderungen, die in Laborbedingungen nicht gegeben sind: die Anbahnung eines Dialogs mit dem Roboter (siehe dazu Scheffler/Pitsch 2020) und situationsadäquate Eröffnungen (vgl. Gehle et al. 2017) sind hier ebenso zu nennen wie mehrere Dialogbeteiligte (vgl. Pitsch/Gehle/Wrede 2013), der angemessene Umgang mit Nicht-Verstehen und Reparaturen (vgl. Gehle et al. 2015) und die durchaus bedeutungstragende Bewegung innerhalb der Ausstellungsräumlichkeiten (vgl. Pitsch/Gehle/Wrede 2013: 2). Auch die bereits erwähnte „Kontextblindheit“ stellt die Dialogsysteme vor Herausforderungen und führt zudem zu Ängsten und Sorgen auf Seiten der Nutzer*innen die auch diskursiv tradiert werden (siehe Habscheid et al. 2020). Hierzu entwickeln Teile der genannten Studien in sog. *Wizard-of-Oz*-Experimenten (WOZ-Studien) Designvorschläge für eine Weiterentwicklung der Dialogmodelle. In dieser Form der Studie wird der Roboter-Prototyp im Hintergrund durch einen Menschen gesteuert, während den Teilnehmer*innen ‚vorgetäuscht‘ wird, mit einem voll funktionsfähigen

Roboter in Dialog zu treten. Dabei wurden v. a. Fragen der Gesprächseröffnung (Gehle et al. 2017; Kuzuoka et al. 2008) und der Anbahnung (Scheffler/Pitsch 2020) bearbeitet. Es konnte festgestellt werden, dass mit passenden Kombinationen von Pausen und *Restarts* die Responsivität und damit die Wahrscheinlichkeit für einen länger anhaltenden Dialog erhöht werden kann (vgl. Pitsch et al. 2009a). Außerdem wurde v. a. die Notwendigkeit multimodaler Rezeptions- und Ausdrucksmöglichkeiten in den Bereichen Blick, Raum und Körperbewegung festgestellt (Gehle et al. 2017; Scheffler/Pitsch 2020). Parallelen zu Chatbots und ECAs zeigen sich auch hier im Hinblick auf die Ritualisierung bzw. Konventionalisierung der Begrüßung, die seitens der humanoiden Roboter deutlich stärker hergestellt wird als von den Nutzer*innen, auch wenn solche Klammern in den hier untersuchten Settings das Zusammenbrechen des Dialogs unwahrscheinlicher und das Eintreten in *topic talk* wahrscheinlicher machen (vgl. Pitsch 2015: 254).

Die Situiertheit und grundlegende Offenheit menschlicher Interaktion als Schwierigkeit für den Roboter treten gerade dann als Problem auf, wenn mehrere Personen am Dialog mit dem Roboter beteiligt sind. In Studien, in denen der Roboter nicht in einem WOZ-Setting kontrolliert, sondern im Modus der ‚autonomen' Steuerung eingesetzt wurde, konnte auch hierfür besonders die Rolle von Blick, Kopfbewegungen und der Koordination der Personen untereinander als relevant ausgemacht werden. So orientieren sich Teilnehmer*innen an Dialogen mit dem Roboter beim Auftreten von Fehlern koordiniert zueinander, unterbrechen den Blick zum Roboter und suchen die Orientierung zum Museumspersonal (vgl. Gehle et al. 2015: 410).

Besonders relevant sind Arbeiten, in denen die Interaktionsdynamiken innerhalb der Nutzer*innen-Gruppe genauer beleuchtet werden. So zeigt sich, ähnlich wie die bereits erwähnte Studie von Porcheron et al. (2018) es auch für häusliche VUI festgestellt hat,[82] dass Antworten an den Roboter, die Adressierung dessen und die Beteiligung am Dialog interaktional zwischen den Beteiligten ausgehandelt wird, der Roboter diese Vorgänge aber nicht registriert: „a time slot which appears to the robot as ‚silence' turns out in fact as a locus of high interactional activity and negotiation between the visitors" (Pitsch et al. 2017: 396). Für die Aushandlung der Partizipation werden neben leise gesprochenen verbalen Ausdrücken auch Körperhandlungen (z. B. Berührungen an der Schulter für die Zuweisung des Rederechts im Dialog mit dem Roboter) relevant (Pitsch et al. 2017: 395; siehe auch Pitsch 2020: 141–142). Bei Nutzer*innen-Gruppen mit Kindern zeigt sich, dass Erwachsene häufig die Rolle eines „Participation facilitators" einneh-

82 Diese Studie wird zusammen mit anderen Studien zu Smart Speakern in Kap. 3.2.4 genauer diskutiert.

men bzw. interaktional zugeschrieben bekommen und Kinder unterstützen, die Rolle der primären Dialogpartner*innen einzunehmen (Gehle et al. 2017; Pitsch 2020). Auch hier wird die Relevanz von Blick, räumlicher Positionierung der Beteiligten und insbesondere Kopfbewegungen herausgestellt (vgl. Pitsch/Gehle/Wrede 2013).

Die situationsabhängige, spontane, dynamische und interaktiv ausgehandelte Partizipation und Positionierung zum Roboter zeigt sich – auch unabhängig von diskursiven (teils verfestigten Einstellungen) zu humanoider Robotik – auch in den Daten von Habscheid et al. (2020). So kann es auch durch das anthropomorphe Design der humanoiden Roboter zu einer Form der „emotionalen Befremdung" (Habscheid et al. 2020: 184) kommen, die aus der Situation heraus entsteht und von den Beteiligten verbal expliziert und somit accountable gemacht wird. Insgesamt lässt sich feststellen, dass der Dialog mit autonomen, humanoiden Robotern auf Seiten der menschlichen Beteiligten als hochgradig situiert, von der räumlichen Umgebung geprägt und interaktional sowohl verbal wie non-verbal ausgehandelt betrachtet werden muss.

3.1.4 Telefonbasierte Sprachdialogsysteme

Auch Sprachdialogsysteme, die etwa in Call-Centern eingesetzt werden und einen Mensch-Maschine-Dialog evozieren, sind sprachwissenschaftlich beleuchtet worden. Eine Arbeit an der Schnittstelle von Gesprächslinguistik und Design-Forschung zu telefonbasierten Sprachdialogsystemen legt Thar (2015) vor. Die Analyse stützt sich auf Tonaufzeichnungen von sechs verschiedenen Dialogsystemen aus dem kommerziellen Call-Center-Einsatz (vgl. Thar 2015: 18). Dabei wurden 40 Personen in zwei Testreihen aufgezeichnet, wie sie Gespräche mit einem telefonbasierten VUI führen; dabei ist zu bemerken, dass es sich um Daten aus einem experimentellen Setting handelt (vgl. Thar 2015: 25), die nicht spontan entstanden sind, sondern durch einen entsprechenden Versuchsaufbau elizitiert wurden – dies diente mit Blick auf das Ziel der Arbeit, Empfehlungen für die Optimierungen des Designs der VUIs zu generieren, der Vergleichbarkeit. Gleichwohl geht Thar (2015: 25) davon aus, dass „die relevanten Verhaltensmuster durch die Untersuchungssituation nicht beeinflusst wurden".

Thar (2015: 17) nimmt an, dass Nutzer*innen entweder bereits „vorhandene Gesprächskompetenz" nutzen oder ein „neues Regelwerk" entwickeln – genau dieses Spannungsfeld will auch die vorliegende Arbeit für Smart Speaker beleuchten, wenn auch mit einer weniger auf die Kompetenz der Nutzer*innen als vielmehr auf den praktischen Vollzug scharfgestellten Perspektive, wobei diese auch

implizites Wissen über Gesprächsorganisation sichtbar machen kann.[83] Thar arbeitet ethnomethodologisch-konversationsanalytisch. Den Teil ihrer Untersuchung zu Interaktionsphänomenen gliedert sie entlang klassischer konversationsanalytischer Kategorien und greift dabei auf Forschungsergebnisse zum Sprecher*innen-Wechsel, zu Reparaturen, zur Sequenzialität sowie zu Gesprächsbeginn und -ende auf (vgl. Thar 2015: 19). Diese Untersuchung führt Thar (2015: 250–251) zu Erkenntnissen im Hinblick auf den Mensch-Maschine-Dialog, u. a. beobachtet sie Schwierigkeiten im Sprecher*innenwechsel durch *barge-ins*, d. h. die sofortige Unterbrechung der Äußerungsproduktion des VUI, sowie bei der Durchführung von Reparaturen, die, so konstatiert sie, „immer selbstdurchgeführt umgesetzt werden" (Thar 2015: 250) müssen. Diese Befunde werden auch in der Konzeption der Analysen der vorliegenden Arbeit wieder aufgegriffen.[84] Weitere Analysen von Telefonaten führen Thar (2015: 271–273) schließlich zu acht „Optimierungsprinzipien", die u. a. die Bereiche Gesprächsstrategien und -organisation, Rezeptionssignale, Reparaturen und Gesprächssorten umfasst. Sie stellt dabei prosodische Besonderheiten bei den Stimmeingaben fest, die die Nutzer*innen verfolgen (insbesondere eine langsame, überartikulierte Sprechweise), obwohl dies die fehlerfreie Spracherkennung beeinträchtigen kann. Folglich empfiehlt Thar, die Systeme so zu gestalten, dass sie diese Sprechweise zum Vorteil nutzen können. Weitere Empfehlungen betreffen primär die Ausrichtung des VUI-Designs an grundlegenden Prinzipien der Gesprächsorganisation sowie der Gattungsspezifik der Telefonate, in denen die Systeme zum Einsatz kommen. Insgesamt liefert die Arbeit mehrere Anknüpfungspunkte auch für die vorliegenden Analysen und wohl die einzige umfassende sprachwissenschaftliche bzw. gesprächsanalytische Studie zu telefonbasierten VUI-Systemen.

Zu telefonischen Sprachdialogsystemen aus medienwissenschaftlicher bzw -historischer Sicht arbeitete auch Volmar (2019) (siehe auch Waldecker/Volmar 2022). Er zeichnet deren Entwicklung von der Einführung des Tastentelefons über das Aufkommen von Tonwahlverfahren in Call-Centern bis hin zu KI-basierter automatischer Spracherkennung zu Smart Speakern nach. Er argumentiert, dass die Entwicklung von häuslichen Sprachassistenzsystemen eng verbunden ist mit der des Call-Centers, in der Nutzer*innen bereits an den Umgang mit maschinellen, dialogisch und mündlich operierenden Gegenübern gewöhnt wurden. Der Gebrauch von Sprachassistenzsystemen erinnere an Praktiken der Call-Center-Nutzung. Folglich schlägt er vor, Smart Speaker als „call centers for the home" zu konzeptualisieren (Volmar 2019: 72) und verweist auf sich ausdifferenzierende

83 Siehe Kap. 2.1.4.

84 Siehe insbesondere Kap. 6.1.3 und 6.1.4.

Motive für den kommerziellen Einsatz von Sprachdialogsystemen: So sei es Anliegen von Call-Center-Betreibern, die notwendigen Personalressourcen zu senken und die Effizienz zu steigern, demgegenüber könne die Auswertung der erhobenen Sprachdaten zu unterschiedlichen Zwecken eine Motivation der Dienstanbieter von Smart Speakern sein.

3.2 Smart Speaker

Stationäre Sprachassistenzsysteme für den Privathaushalt sind noch nicht lange verfügbar: In Deutschland war der erste Smart Speaker von Amazon (der *Amazon Echo* mit dem Sprachdialogsystem *Alexa*) ab Anfang 2017 ohne Vorbestellung erhältlich, auf dem US-amerikanischen Markt ca. anderthalb Jahre früher. *Google Home* bzw. *Nest* und der *HomePod* der Firma Apple folgten Mitte 2017 bzw. Mitte 2018 (vgl. Hoy 2018: 82). Folglich beziehen sich bisher veröffentlichte Studien zu großen Teilen auf den angloamerikanischen Raum, in dem alle Geräte etwas früher marktüblich waren, und sind insgesamt noch sehr jung, wie auch die Geschichte der Sprachtechnologie zeigt, über die nachfolgend ein Überblick gegeben werden soll.

3.2.1 Historische Entwicklung der Sprachtechnologie

Die Historie des sprachlichen Austauschs von Menschen und Maschinen ist eng verwoben mit der Entwicklung von Rechenmaschinen und Computern im Allgemeinen. Sie ist darüber hinaus mit der großen Frage nach der Künstlichen Intelligenz (KI) verknüpft: Ab wann kann eine Maschine intellektuell einem Menschen gleichwertig sein? Diese Entwicklungen und Erzählungen sind so oft in der einschlägigen Literatur mit unterschiedlichen Schwerpunkten beschrieben worden, dass hier davon abgesehen werden soll, noch einmal die philosophischen Überlegungen zur Unterscheidung von menschlicher und maschineller „Intelligenz“ von René Descartes von 1637 oder Denis Diderot von 1769 darzulegen. Es soll an dieser Stelle ebenso auf eine Erläuterung des von Alan Turing um 1950 ersonnenen Turing-Tests zur Unterscheidung menschlichen und maschinellen Denkens verzichtet werden[85] wie auf eine weitere Erzählung der Erfindung von ELIZA (Weizen-

85 Siehe dazu Turing (1950); eine Anwendung auf zeitgenössische Dialogsysteme findet sich etwa bei Kaerlein (2018: 19–21); siehe auch Nebel (2019) und explizit zur Geschichte der Stimmsynthese Borbach (2018). Turings Beitrag, der auch als ein Grundstein für die Entwicklung von Digitalität betrachtet werden kann, und das darin enthaltene Experiment zur Unterscheidung

baum 1966) und der daraus resultierenden Kritik an der Vorstellung von der ‚starken künstlichen Intelligenz' (vgl. Weizenbaum 1977).[86] Auch der „Schachtürke" des Wolfgang von Kempelen von 1789 und darauf aufbauende Wizard-of-Oz-Experimente, die auf einer Täuschung der Nutzer*innen basieren, sollen hier nicht detailliert beschrieben werden (siehe dazu Felderer/Strouhal 2004).[87]

Damit will die Arbeit keineswegs in einen ahistorischen Blindflug geraten. Wolfgang von Kempelen, Konrad Zuse, Alan Turing, Joseph Weizenbaum und andere haben zweifelsohne bedeutende Grundsteine für eine Erforschung und Weiterentwicklung des Austauschs zwischen Mensch und Maschine gelegt, die auch für die Gegenwartsforschung von höchster Relevanz sind (vgl. etwa Baranovska/Höltgen 2018b: 11). Doch nicht nur sind diese technologischen Entwicklungen mit ihren sozialen Konsequenzen mittlerweile umfangreich dargestellt worden (siehe dazu etwa die Monographien von Heintz 1993 und Wilker 2002),[88] sie sind auch für die hier durchgeführten Analysen nur implizit relevant: Durch Smart Speaker und andere Voice User Interfaces (VUIs) – spätestens mit der Präsentation von Siri 2011 – ist der mündliche und (bedingt) konversationsartige Austausch mit Maschinen Realität geworden. Es lässt sich dabei beobachten: Der Austausch ist keine Täuschung wie bei Wolfgang von Kempelen mehr, bei der ein Mensch die Antworten vorsagt. Die Anwendung von auf Wahrscheinlichkeitsrechnungen beruhenden NLP-Prozessen macht die Funktionsweise der Systeme verlässlicher und er prägt ihre Anwendungsgebiete. Die Systeme sind deutlich komplexer und

von Mensch und Maschine sind wirkmächtig und können implizit für die folgenden Analysen eine Rolle spielen, werden darin aber nicht explizit, sodass eine ausführliche Darstellung hier unterbleiben soll.

86 Zweifellos hat ELIZA großen Einfluss auf die ethnomethodologische bzw. praxeologische Betrachtung von Mensch-Maschine-Dialogen sowie auf deren Erforschung im Allgemeinen gehabt – das zeigt nicht nur das gesteigerte Interesse Harold Garfinkels an ELIZA (vgl. Suchman 1987: 65; Lynch 2011: 932), sondern auch die weitere Beschäftigung damit in rezenten Forschungsarbeiten (Baranovska/Höltgen 2018a; Shrager 2021; Eisenmann et al. 2023).

87 Es unterbleibt hier auch ein in einschlägigen Arbeiten häufig zu findender Abgleich zwischen den jüngeren Entwicklungen der Realität und populären Fiktionen in älteren Filmen, etwa aus Filmen wie *2001: Odyssee im Weltraum* (1968) mit dem darin eingeführten Sprachroboter *HAL 9000* (siehe Baum 2018). Weitere aktuellere Beschäftigungen mit dieser Frage in Film und Literatur finden sich etwa in *Her* (2013), in der sich der Protagonist mit dem „Operating System" Samantha auf eine Liebesbeziehung einlässt, in Romanen wie *Maschinen wie ich* (McEwan 2019), Erzählungen und Kurzgeschichten (Kehlmann 2021; Braslavsky 2019) sowie in der populärwissenschaftlichen Literatur (etwa Ramge 2018; Misselhorn 2021; Lenzen 2024).

88 Zum Turing-Test siehe auch Krummheuer (2010: 78–79) und Lotze (2016: 28); zur Erfindung und zu den Folgen von Joseph Weizenbaums ELIZA siehe ebenfalls Krummheuer (2010: 79–80) und Lotze (2016: 31–34). Für eine zusammenhängende, populärwissenschaftliche Darstellung siehe Drösser (2020: 15–30).

v. a. ‚gehaltvoller' als die Dialoge, die etwa mit ELIZA möglich waren: Über VUIs können Informationen abgefragt, Programme bedient und Haushaltsgeräte gesteuert werden. Der Austausch entspricht dennoch nicht den in der Popkultur gezeichneten Fiktionen vom Mensch-Maschine-Austausch der Zukunft: Sie werden zwar von den Herstellern als *Persona* mit Identitätsmerkmalen ausgestattet, entwickeln aber keinen eigenen Willen oder Gefühle wie Samantha in „Her". Der Austausch mit VUIs ist bei den o. g. Systemen akustisch für kompetente Sprecher*innen gut von zwischenmenschlichen Interaktionen zu unterscheiden.[89] Smart Speaker in ihrer zuvor beschriebenen Erscheinungsform basieren durch ihre Digitalität, ihre im Hintergrund prozessierte Vernetztheit, ihre stochastische Funktionsweise und nicht zuletzt durch ihre Kommerzialisierung insofern auf ganz anderen technologischen und sozialen Grundlagen als ihre Vorläufer.

Die jüngere Geschichte der Sprachtechnologie setzt mit den Erfolgen bei den korpusbezogenen Verfahren der Sprachtechnologie in den 80er-Jahren ein (vgl. Carstensen et al. 2010: 21). Diese bedeuteten eine Abkehr von symbolischen Ansätzen mit Universalitätsanspruch und konzentrierten sich auf das *Training*, d. h. den Einsatz von für die Nutzung optimierten, vorbereiteten Datensätzen zur Modelloptimierung bei der Erkennung von Sprachmaterial. Wie Carstensen et al. (2010: 21) herausstellen, ist zu diesem Zeitpunkt (noch) nicht die stochastische Vorgehensweise, sondern der Einbezug eines Trainingsverfahrens zur Anpassung von Rechenmodellen für verschiedene (u. a. stochastische) Verfahren entscheidend. Die Abkehr vom Universalismus hin zur Anwendung der Modelle und Anpassung der Programme durch die Trainingsdaten auf einen bestimmten Einsatzzweck hin waren die wesentliche Veränderung im Vergleich zu vorherigen Versuchen. Gleichwohl wurden v. a. Ansätze weiterentwickelt, die dem, wie Drösser (2020: 12) es vergleicht, „statistischen Lernen" eines Kindes im Erstspracherwerb ähneln. *Hidden-Markov-Modelle* zur Umwandlung von Laut in Text und *N-Gramme* zur Identifizierung der Laute und Wörter (siehe dazu Carstensen et al. 2010: 130–136; Pfister/Kaufmann 2017: 107–110 sowie Kap. 3.2.2) wurden mit mehr Rechenleistung ausgebaut, konnten über mehr Trainingsdaten verfügen und in den 90er-Jahren konnten die Verfahren für einfache Anwendungsbereiche (z. B. einfache Telefon-Dialogsysteme) kommerziell eingesetzt werden.

Die in den 80er- und 90er-Jahren zur Verfügung stehenden Daten und v. a. die Rechenkapazitäten reichten allerdings keineswegs aus, um damit algorithmi-

89 Eine Ausnahme könnte das von Google 2018 vorgestellte System „Google Duplex" sein, das jedoch nicht in der Breite verfügbar ist. Durch die jüngeren Entwicklungen im Bereich der Large Language Models (LLMs) ist möglicherweise ebenfalls eine zunehmende Einschränkung der Unterscheidbarkeit von zwischenmenschlichen Interaktionen und Interface-Dialogen zu treffen, wobei diese aber zum Zeitpunkt der Datenerhebung für die vorliegende Arbeit noch nicht in der Breite verfügbar waren; zur Relevanz LLM-basierter Systeme siehe Kap. 7.3.

sches *Deep Learning* wie es heute Standard im Natural Language Processing (NLP) ist, zu ermöglichen. Sie scheiterten noch an größeren Erkennungsaufgaben und hatten hohe Fehlerquoten. Spracherkennung beschränkte sich weiterhin auf einzelne Schlüsselwörter oder basale Eingaben und hatte nur unter wenigen Aspekten Ähnlichkeit mit menschlichen Konversationen. Dennoch entwickelten sich durch den zunehmenden Ausbau von Netzwerken und die damit grundlegend veränderte Menge und Verfügbarkeit von Daten Anwendungen im Bereich der KI weiter, wie z. B. siegreiche Schachcomputer (Deep Blue von IBM gewann 1997 gegen den damaligen Schachweltmeister).

Für qualitativ hochwertige Spracherkennung und -produktion reichte dies dennoch noch nicht aus, auch wenn 1990 mit der Spracherkennungssoftware DragonDictate erste Anfänge unternommen worden waren. Erst die deutlich umfangreicheren Rechenkapazitäten und die Weiterentwicklung von *Deep Learning* um die 2010er-Jahre, mit denen sogenannte neuronale Netze der Standard wurden, ermöglichten die Mustererkennung in weitaus größerem Stil, die für Bild-, Text- und Spracherkennung angewendet werden konnte. Hier konnte ein anderes IBM-Produkt, der Computer Watson, in der Quiz-Show *Jeopardy!* gewinnen. Watson war im Unterschied zu Deep Blue ein „aus Daten lernendes System" (Ramge 2018: 40), das viel mehr Rechenkapazität aufwies und die Optimierung der in ihm eingebauten Modelle insofern in viel kürzerer Zeit ermöglichte. Zu einer erhöhten Präzision trug zudem das „Long Short-Term-Memory"-Verfahren bei, das zwar bereits seit Ende der 90er-Jahre vorgestellt wurde (vgl. Hochreiter/Schmidhuber 1997), aber wegen der in der Breite nicht zur Verfügung stehenden Hardware kaum zum Einsatz kam. Es setzte sich durch und führte zu entsprechenden Leistungen, z. B. Anfang der 2010er-Jahre in der Erkennung von Handschriften (vgl. Schmidhuber 2019). Die Forschung großer Tech-Firmen wie IBM, Apple, Facebook oder Microsoft führten ebenfalls zu einem Schub, der zwischen 2000 und 2010 gravierende Veränderungen im Bereich der Bild- und Spracherkennung auslöste. Dieser Schub trug auch zur Entwicklung der Software Siri bei, die 2011 von Apple vorgestellt wurde und den ersten IPA darstellt, der über ein VUI bedienbar und auf marktüblichen Produkten in der Breite verfügbar war. Konkurrenzprodukte von Google, Amazon und Microsoft folgten in den Jahren 2012 bis 2014 und dominieren seitdem den Markt für IPA und insbesondere für IPA auf eigenen Endgeräten (Smart Speaker).

Die Entwicklungen seit der Jahrtausendwende waren also ganz wesentlich davon geprägt, dass große Rechenkapazitäten Modelle Realität werden ließen, die zuvor bereits seit den 70er- und 80er-Jahren entwickelt worden waren. Sie waren außerdem wesentlich davon beeinflusst, dass das private Internet sich ausbreitete und die Wirtschaft ein Marktpotenzial in entsprechenden Anwendungen erkannte und entsprechend investierte. Mathematisch-informatisch-technische Ent-

wicklungen waren ebenfalls an diesen Prozessen beteiligt, doch lag hier bereits viel Pionierarbeit vor, auf die mit dem Aufkommen der technologischen Verbesserungen aufgebaut werden konnte.

3.2.2 Funktionsweise und technischer Hintergrund

Smart Speaker werden im Deutschen auch als Sprachassistenten, „Intelligente Persönliche Assistenten“ oder „Intelligente Lautsprecher“ bezeichnet. Der hier zu beschreibende Gerätetyp zeichnet sich dadurch aus, dass er als stationäres Sprachassistenzsystem fest in der häuslichen Wohnumgebung der Nutzer*innen installiert wird und, durch seine Größe und ein obligatorisches Stromkabel, nicht mobil ist.[90] Sie sind „standalone screenless devices“ (Porcheron et al. 2018: 1) und grenzen sich somit von anderen Sprachassistenzsystemen ab, die auch als „Conversational Agents“, „Voice Assistants“, „Sprachassistenten“ oder „Conversational User Interfaces“ bezeichnet werden. Solche rein softwarebasierten Instanzen umfassen zwar auch einen Großteil der oben beschriebenen technologischen Anwendungen. Sie können auch als Voice User Interface (VUI) stimmlich bedient werden, verfügen aber nicht unbedingt über eine eigene Hardware. Vielmehr sind sie als Software auf mehr oder weniger mobilen Geräten wie Smartphones, Tablets, Computern oder auch im Auto installiert. Sie binden dabei in die Verarbeitung der Anfragen und die Darstellung der Ergebnisse auch die Bildschirme mit ein. Sie signalisieren z. B. durch Symbole, dass die Anfrage verarbeitet wird, zeigen unterstützend zur mündlichen Ausgabe auch Text, Tabellen, Karten oder Bilder und ersetzen teilweise sogar die rein mündliche Ausgabe durch eine visuelle Darstellung. Bedford-Strohm (2017: 486) verweist zur Abgrenzung der Smart Speaker von diesen Geräten auch auf den Begriff der „Voice-First-Geräte“, die ausschließlich über die Stimme bedient werden; ein Eingreifen über den visuellen Kanal ist nur bedingt möglich.[91] Der Begriff Smart Speaker soll im Folgenden beibehalten werden. Er verweist erstens über den Lautsprecher auf dessen physische Materialität (auch wenn diese noch mehr umfasst als den Lautsprecher, s. u.). Der Begriff erscheint außerdem passend, weil er mit dem englischen Begriff „Smart“ den im deutschsprachigen Diskurs umstrittenen Begriff der „Intelligenz“

90 Auch die zugehörigen Smartphone-Apps orientieren sich bei der Zuordnung der Geräte an verschiedenen Wohnbereichen (Esszimmer, Wohnzimmer, Küche o. ä.) und rahmen sie damit als stationär und nicht für den tragbaren Einsatz vorgesehen.

91 Zu den Spezifika der einzelnen Assistenzsysteme und die Anbindung insbesondere an die Smartphone-App, über die eine gewisse visuelle Kontrolle über die Geräte hergestellt wird, siehe Kap. 3.3.

umschifft, und ferner, weil die Konnotation mit *Smart Home* passend erscheint, da Smart Speaker wesentlicher Bestandteil einer Smart-Home-Struktur sein können (vgl. Goulden 2019; Strüver 2023a).

Hardwareseitig bestehen diese Systeme v. a. aus drei Komponenten: mehreren, miteinander zusammenwirkenden Mikrofonen (sog. *Far-Field-Voice-Recognition*, siehe Drösser 2020: 71), einem Lautsprecher sowie einer kleinen Recheneinheit, die die Übertragung ins Internet leistet und einige wenige lokale Prozesse steuert. Teilweise verfügen die Modelle außerdem über Lichtringe oder -punkte, die in unterschiedlichen Farben nichtsprachliche Signale geben können (vgl. Bedford-Strohm 2017: 487; Pearl 2016: 219–221).[92] Die einzelnen Komponenten, aus denen das stationäre Sprachassistenzsystem sich zusammensetzt, werden von verschiedenen Software-Anwendungen in Anspruch genommen (vgl. die Abbildung bei Zarcone/Leschanowsky 2023: 169). Diese ergeben erst im Zusammenspiel eine Einheit und sind im Hintergrund als unterschiedliche Prozesse zu begreifen, die auch verschiedene Teildisziplinen der Informatik und der Computerlinguistik berühren.

Die genannten Prozesse lassen sich im Wesentlichen mit Natale (2020: 5) in drei verschiedene Großbereiche unterteilen: *Speech Processing* (SP), *Natural Language Processing* (NLP) sowie *Information Retrieval* (IR).[93] Damit sind jeweils fachlich tiefgreifende informatisch-computerlinguistische Prozesse angesprochen, die im Rahmen dieser Arbeit nicht ausführlich beschrieben werden können und auch nicht müssen. Allerdings sollen die Funktionen, die diese Prozesse für den Austausch zwischen Mensch und Maschine übernehmen, hier kurz überblicksartig vorgestellt werden, um erstens die an die Smart Speaker adressierten Äußerungen der Nutzer*innen vor diesem Hintergrund besser interpretieren zu können und zweitens die Ausgaben des VUI einzuordnen.[94]

Wesentlich für das Funktionieren aller folgenden Dienste ist die dauerhafte Verbindung zum Internet (vgl. Hoy 2018: 82). Lediglich die für die Aktivierung des Geräts notwendige *wake word*-Suche findet auf den lokalen Recheneinheiten der Smart Speaker statt (vgl. Pearl 2016: 127), was auch den VUI-Design-Empfehlungen etwa bei Pearl (2016: 153) entspricht: „recording everything the user says, even when not engaged with your app, and sending it to the cloud is not ethical". Nach Selbstauskunft von Amazon und Google wird diesen Empfehlungen auch bei den Echo- und Google-Assistant-Geräten gefolgt (vgl. Amazon 2021; Google 2021);[95]

92 Für eine genauere Beschreibung der in dieser Arbeit analysierten, aktuell marktüblichen stationären Sprachassistenzsysteme, siehe Kap. 3.3.

93 Für eine Übersicht über das mit Smart Speakern zusammenhängende Fachvokabular siehe auch Kahle/Meißner (2020: 21).

94 Eine auch für Laien verständliche Übersicht liefert das Sachbuch von Drösser (2020).

95 Siehe dazu die Diskussion in Kap. 3.2.6.

gleichwohl findet ein dauerhaftes akustisches Scannen (ohne weitreichende Auswertung) statt, das die mündlich vollzogene Bedienweise ermöglicht (vgl. Hoy 2018: 82). Sobald das Aktivierungswort erkannt wird, wechselt das Gerät in den Listening-Modus und es findet eine Transkription und Auswertung der Daten über die cloudbasierten Online-Dienste der Plattformen statt. Die Verarbeitung wird erst dadurch möglich, dass das aufgenommene akustische Material mit Hilfe von *Speech Recognition*-Anwendungen analysiert wird, für die große Datenmengen und Rechenleistungen erforderlich sind (vgl. Crawford/Joler 2018; Kohne et al. 2020: 20). Für die Spracherkennung, d. h. die reine Transkription von Gesprochenem in Geschriebenes, werden probabilistische Modelle eingesetzt, die es ermöglichen, die Wahrscheinlichkeit auszurechnen, dass ein bestimmter auditiver Input einer bestimmten Abfolge von Buchstaben entspricht (für mathematisch-technische Grundlagen zu diesem Prozess siehe Kamath/Withaker 2019: 141–201): Der auditive Input wird dabei zunächst in eine digitale Darstellungsform gebracht und dann, grob gesagt, mit einem Ansatz der statistischen Spracherkennung analysiert, den Pfister/Kaufmann (2017: 328) wie folgt beschreiben: „Der Erkenner verfügt für jedes Wort des Vokabulars über eine statistische Beschreibung von dessen akustischer Realisierung. Es gilt dasjenige Wort als erkannt, dessen statistische Beschreibung am besten auf die zu erkennende Äusserung [sic] passt“. Diese Erkennung wird durch eine große Menge von Daten ermöglicht, mit denen eine bei Pfister/Kaufmann zuvor als „Erkenner“ bezeichnete mathematische Funktion „trainiert“ wurde. Durch die fortlaufend hinzukommenden Daten, die laut der Datenschutzvereinbarungen mit den Anwender*innen der Geräte auch entsprechend verwendet dürfen, werden die Funktionen immer weiter angepasst und verändern bzw. ‚verbessern‘ das neuronale Netzwerk durch die neuen Berechnungen (siehe auch Knaut 2018: 80; Kipp 2023). Dieser Prozess vollzieht sich zu Teilen ohne manuelles Eingreifen und wird primär aufgrund der statistischen Angaben in der Funktion realisiert (vgl. Kamath/Withaker 2019: 6–7). Dabei werden über sog. *N-Gramm*-Sprachmodelle nicht nur die Tonhöhen eines einzelnen Wortes ausgewertet, sondern auch die Wahrscheinlichkeit des Auftretens in Nachbarschaft zu anderen Wörtern; bei der Berücksichtigung von zwei Wörtern (in älteren Systemen) spricht man hier von *Bi-Gramm*-Sprachmodellen, doch der Prozess ist prinzipiell auf bis zu *n* Wörter erweiterbar (siehe dazu Pfister/Kaufmann 2017: 417–421). Zu nicht unerheblichen Teilen erfolgt das Training jedoch auch durch menschliches Nachhören der gespeicherten Aufnahmen, deren maschinelle Transkription dann ‚verbessert‘ wird, was die mathematischen Berechnungsverfahren wiederum verzeichnen und in ihre Transkriptionsverfahren einfließen lassen – dies hatten alle Betreiberfirmen nach investigativen Recherchen in öffentlichen Statements bekanntgegeben (vgl. Kremp 2019; siehe auch Waldecker/Volmar 2022: 164; Hector 2025: 71).

Die genauen technischen Details über den Prozess der *Speech Recognition* geben die Hersteller nicht öffentlich außerhalb von, für Laien schwierig nachvollziehbaren, Fachjournalen oder eigenen Webseiten[96] bekannt. In etlichen Publikationen zu Smart Speakern aus den Geistes-, Kultur- und Sozialwissenschaften und in der Presse wird bei der Beschreibung die genaue Funktionsweise ausgespart (und auch die vorliegende Arbeit geht diesbezüglich nicht in die Tiefe). Wie jedoch Drösser (2020: 93–99) bemerkt, kommen die erwähnten probabilistischen Erkennungsprogramme in den derzeit verfügbaren Sprachmodellen generell zum Einsatz; erst der Einsatz von *Deep Learning* und die Entwicklung der dafür notwendigen Rechenleistungen machten eine solche Art des *Speech Processings* überhaupt möglich. Sehr ähnlich funktioniert der Prozess auch bei der Sprachsynthese, also dem Erzeugen sprachlicher Laute durch die Geräte. Insbesondere die Prosodie spielt bei der Sprachsynthese eine entscheidende Rolle (vgl. Pfister/Kaufmann 2017: 262–264) und auch sie konnte durch neuronale Netzwerkfunktionen und Rechenleistungen verändert werden, indem sie die statistischen Wahrscheinlichkeiten, z. B. auch für noch unbekannte Lexeme, für eine bestimmte Sprechgeschwindigkeit oder Tonhöhe aus deren Vorkommen in anderen Zusammenhängen heraus errechnen (siehe auch Drösser 2020: 55–60). Grundlage sind dabei schriftliche Inputs, die aus dem zweiten großen Baustein heraus entstehen, den Smart Speaker vereinen: *Natural Language Processing* (NLP).

NLP bezeichnet den Prozess der *Verarbeitung* natürlicher Sprache, d. h. deren Erkennen und graduelles Verstehen (siehe dazu Goksel-Canbek/Mutlu 2016: 594). Der Prozess kann dabei in zwei Teilbereiche unterteilt werden: Natural Language Understanding (NLU) und Natural Language Generation (NLG) (vgl. Kohne et al. 2020: 42–43). Das Verstehen und Generieren verweist hier nicht auf die Transformation von Gesprochenem zu Geschriebenem: Im Bereich des NLP wird ausschließlich mit schriftlichen Daten gearbeitet, der Wechsel zwischen der mündlichen und der schriftlichen Darstellungsweise liegt ganz im Bereich des zuvor erläuterten *Speech Processing*. Der transkribierte Input wird nun (im Prinzip) genauso auf der Basis schriftlichen Inputs verarbeitet wie bei einem Chatbot oder anderen Conversational User Interfaces auch (vgl. Kabel 2020: 1–2; Bedford-Strohm 2017: 486–487). NLU ‚sucht' nun in verschiedenen Schritten in dieser Äußerung (*utterance*) nach *Intents* und entsprechenden *Entities*: Erstere ist, mit Kohne et al. (2020: 94), die „Absicht des Benutzers" oder, genauer genommen, das, was das Programm als mögliche Absicht bzw. aufzurufende Funktion identifiziert (z. B. „Wetter", „Nachrichten" oder „Bahnverbindung"). Letztere sind Spezifikationen, die in der erkannten Funktion abgefragt

96 Siehe etwa Amazon Science (2023).

werden sollen (z. B. „Wetter *in Siegen*“, „*Sport*nachrichten“, „Bahnverbindung *nach Berlin*“). Komplexere Systeme ‚suchen‘ außerdem nach dem *Skill*, d. h. dem spezifischen Programm, mit dem die angeforderte Funktion ausgeführt wird. Diesen *Skills* ist eine *invocation phrase* zugeordnet, die das Programm öffnet und den *Intent* sowie die *Entity* damit verarbeitet. Innerhalb des *Intents* werden zudem funktionsspezifische *Slots* vergeben, die die *Entities* genauer hinsichtlich der ‚Such-Erwartungen‘ seitens der Software klassifizieren; so würde in dem fiktiven Beispiel „Bahnverbindung *nach Berlin*“ im Bereich *nach Berlin* als *Slot* eine Ortsangabe ‚erwartet‘ und die Suche daraufhin eingegrenzt.

Für die Verarbeitung des Inputs auf diese Weise durchläuft das Programm verschiedene Schritte, die grob in Tokenisierung, Lemmatisierung, Wortartenklassifikation, Satzteilung und Wort-Vektorisierung untergliedert werden können (vgl. Kohne et al. 2020: 44–49). In der Tokenisierung werden, auch hier mit Hilfe algorithmischer bzw. statistischer Verfahren, die erkannten Buchstaben in einzelne Wörter oder Satzzeichen unterteilt (vgl. Kamath/Withaker 2019: 92–93). Die erkannten Wörter werden nun durch sog. *Stemming* bzw. Lemmatisierung auf ihre Grundformen reduziert, um sie um Wortendungen durch Deklination und Konjugation und andere morphologische Prozesse zu bereinigen, sodass die Variabilität reduziert und somit die statistische Treffsicherheit erhöht wird; die genauen syntaktischen Bezüge der Wörter untereinander, die durch diesen Prozess verloren gehen, werden dabei marginalisiert (vgl. Kohne et al. 2020: 45–46; Kamath/Withaker 2019: 92). Durch Wortartenklassifikation und Satzteilung wird der Input weiter für die Verarbeitung zugerichtet. Um schließlich die Inputs den ‚gesuchten‘ Stellen zuzuordnen (*Skills*, *Intents*, *Launch*, *Entities*, *Slots*) ist diese Vorbereitung unumgänglich, aber noch längst nicht ausreichend, um eine hohe Treffsicherheit bei dieser Zuordnung zu erzielen, insbesondere nicht, wenn nicht vorausgesetzt wird, dass die Stimmeingaben in ihrer sprachlichen Gestaltung variieren: NLP kann, nach heutigen Standards, ebenso *Utterances* wie „Wie wird das Wetter in Siegen?“ wie auch „Wird es heute in Siegen regnen?“ verarbeiten. Dazu werden die erkannten Wörter vektorisiert, d. h., die Algorithmusfunktion prüft die gemeinsamen Kontexte und die Häufigkeit, mit der diese in den jeweiligen Kontexten auftreten, als Beziehung zueinander. Dieser Vorgang wird bei Kohne et al. (2020: 49) als räumliche Beziehung oder Vektoren dargestellt: Je größer der Abstand, d. h. je geringer das Vorkommen in einem gemeinsamen Kontext, desto weniger wahrscheinlich ist es, dass die Worte synonym für eine bestimmte Suchstelle verwendet werden können. Die Berechnung der gemeinsamen Kontexte erfolgt dabei auf Grundlage von zuvor durch die neuronale Netzwerk-Funktion verarbeiteten Daten, die die Bedeutungsähnlichkeiten und Auftreten unterschiedlicher Wörter in der gleichen Anfrage bzw. dem gleichen Zusammenhang für die Funktion erkennbar macht, sodass die entsprechenden Parameter darauf einge-

stellt sind. Durch die fortlaufende Nutzung entstehen auch hier immer weiter Neuberechnungen: Einmal erkannte Zusammenhänge bzw. Kontexte werden immer stabiler erkannt, während Abweichungen von den einmal erkannten Mustern immer schneller ausgeschlossen werden.[97] Zudem können diese „Advanced Bots" (Kabel 2020: 2) nicht nur den Kontext, sondern auch die Interaktionshistorie auswerten und stellen Wahrscheinlichkeitsbezüge zu in vorherigen *Utterances* geäußerten *Intents* oder *Entities* her.

Für die NLG kommt eine Kombination aus Textbausteinen und Variablen zum Einsatz. Während einige der Dialoge manuell einprogrammiert sind, sodass einer bestimmten Frage eine konkrete, voreingestellte Antwort zugeordnet wird (etwa bei Fragen nach dem Sinn des Lebens oder nach Charaktereigenschaften des VUIs), wird bei anderen, funktionell komplexeren Anwendungen die Antwort mit Informationen aus dem *Information Retrieval*-Prozess (s. u.) gespeist und dann in entsprechende Text-Vorlagen eingebaut.[98] Die beiden Prozesse des NLP (NLU und NLG) rahmen zeitlich gesehen den Prozess des *Information Retrieval* (IR), da die NLU zunächst abgeschlossen sein muss, bevor überhaupt IR betrieben werden kann, gleichzeitig die NLG erst erfolgen kann, wenn der IR-Prozess zu einem Abschluss gekommen ist. Der ganze Vorgang dauert meist nur sehr wenige Sekunden, dennoch verzahnen sich hier wie bereits angedeutet prinzipiell autonome Software-Prozesse miteinander. Der letzte hier zu beschreibende Prozess der IR ist insofern voraussetzungsreich, als er auf mit den Cloud-Systemen der Hersteller verknüpfte Datenbanken und Internet-Suchmaschinen angewiesen ist (vgl. Natale 2020: 12–13). Welche Daten dort an welcher Stelle konkret abgefragt werden, hängt davon ab, welche Stellen mit den jeweiligen *Skills* verknüpft sind und welche *Entities* erkannt wurden. Die Stellen, auf die dabei zugegriffen wird, können sehr unterschiedlich sein, doch im Wesentlichen haben auch hier der sprunghafte Anstieg der Rechenleistungen in den letzten Jahren klassische Datenbankabfragen ersetzt, IR ist so vielmehr definiert als „finding material [...] of an unstructured nature [...] that satisfies an information need from within large collections [...]" (Manning/Raghavan/Schütze 2018: 1).

Bei den Suchanfragen wird nicht zwingend von strukturierten Daten ausgegangen, sondern von nicht für die automatisierte Abfrage vorbereiteten Daten:

97 Siehe dazu auch das von Drösser (2020: 35) angeführte Beispiel einer NLP-Anwendung, die durch die Überschwemmung mit rassistischen und beleidigenden Daten so „trainiert" wurde, dass die Funktion nach kurzer Zeit selbst stabil rassistische und beleidigende Ausdrücke ausgab.
98 Für eine übersichtliche Darstellung der Abhängigkeitsverhältnisse von maschinellem Lernen und algorithmischen Funktionen, Sprachtechnologie und Suchanfragen als Anwendungsfall siehe auch Kabel (2020: 40–41).

„Die Suche [...] berücksichtigt dabei die Vagheit und Unvollständigkeit, die sowohl bei der Formulierung des Informationswunsches als auch bei der – ggf. automatischen – Interpretation des Inhalts der betrachteten Dokumente besteht“ (Henrich 2008: 15). Doch welche Schnittstellen genau mit den einzelnen *Skills* verknüpft sind, ist von Anwendung zu Anwendung und den damit verbundenen Datentypen unterschiedlich (vgl. Waitelonis 2018: 8): So ist etwa der auf Amazon-Geräten vorinstallierte Wetter-*Skill* zum Zeitpunkt der Erhebung mit der Plattform „AccuWeather.com“ verknüpft, von wo aus die benötigten Informationen in strukturierter Form abgefragt werden können. Dabei (und auch bei anderen Funktionen, z. B. Routenplanung oder Kartendiensten) werden auch der Standort des Geräts bzw. Geodaten einbezogen. Auch für andere Wissens- bzw. Informationsabfragen (zur Unterscheidung von Daten, Wissen und Information in diesem Kontext siehe Henrich 2008: 18–19) kann auf durch semantische Datenbanken strukturiertes Wissen z. B. in Form des Google Knowledge Graph oder von Wikidata zugegriffen werden (vgl. Fensel et al. 2020: 8–10), was der IR deutlich umfangreichere Informationen bereitstellt und diese auch deutlich präziser werden lässt. Auch wenn also die Methoden der IR in nicht-strukturierten Wissensbeständen ebenfalls zum Einsatz kommen, treiben die Hersteller die Entwicklung von strukturierten Daten als Grundlage für die Wissensabfragen voran.

Solche Knowledge Graphs sind „*very* [Herv. i. O.] large semantic nets that integrate various and heterogeneous information sources to represent knowledge about certain domains of discourse“ (Fensel et al. 2020: 6); es handelt sich also etwas einfacher gesagt um sehr große, automatisch verarbeitbare und semantisch miteinander vernetzte Wissensbestände, die es ermöglichen, einfache Fragen (z. B. nach dem Alter von Prominenten oder der Höhe von Bauwerken) allein auf Grundlage der strukturierten Daten des *semantic web* zu beantworten, ohne auf in Text- oder Bildform und nicht-annotiert vorliegende Internetdokumente zurückgreifen zu müssen, deren maschinelle Verarbeitung trotz ausgeklügelter IR-Methoden häufig noch an ihre Grenzen stößt. Die Ergebnisse der Knowledge Graph-Suche sind bei Suchmaschinen-Abfragen über Google in sog. „Knowledge Panels“ noch über der eigentlichen Ergebnisliste zu sehen und werden in die Sprachausgaben der Smart Speaker eingebunden. Der Grundstock der Wissensbestände in ihrer heutigen Form speiste sich aus Internetquellen wie der Online-Enzyklopädie Wikipedia bzw. Wikidata (vgl. Vrandečić/Krötzsch 2014),[99] Freebase und der Statistik-Basis des *CIA World Factbook* (vgl. Singhal 2012).

99 Für das Online-Projekt Wikipedia dient mittlerweile die von den Wikipedia-Trägerorganisationen gegründete Plattform Wikidata als *semantic web*-Anwendung und zugrundeliegende Datenbasis. Der Google Knowledge Graph speist sich auch weiterhin aus den dort hinterlegten Eintragungen und Zu-

Diese Internetsuchfunktionsweise über Knowledge Graphs ist nicht nur für den Funktionsbereich des IR von Interesse: Durch ihre Vernetztheit können Knowledge Graphs auch Spracherkennungssoftware und NLP-Methoden unterstützen, indem sie Zusammenhänge und Kontexte, in denen Äußerungen auftreten, nicht nur – wie oben beschrieben – auf Basis von Daten*beispielen* und Neuronalen Netzwerk-Funktionen erkennen, sondern auch auf Basis von semantischen Verknüpfungen. Insofern spielt die weitere Entwicklung von Knowledge Graphs besonders für Smart Speaker und andere VUI-basierte Anwendungen eine wesentliche Rolle und bildet z. B. im Fall von Google auch die Basis für die gesamte VUI-Anwendung *Google Home.*

In der Gesamtschau betrachtet sind die Systeme für die Geistes- und Sozialwissenschaften also durchaus schwer zugänglich, auch wenn jüngst Einführungen und sogar Lehrbücher in diesem Bereich vorbereitet werden (etwa Lotze i. V.). Als Hauptquelle für das Wissen über die Funktionsweise können einerseits Blogs und Dokumentationen der Hersteller identifiziert werden, in denen die Schnittstellen auch für Drittanbieter beschrieben werden, die die Entwicklung von Anwendungen für die Geräte planen (vgl. Strüver 2023a). Andererseits sind Analysen von Application Programming Interfaces (APIs) mit unterschiedlicher Ausrichtung ein emergentes, interdisziplinäres Forschungsfeld, das derzeit aus unterschiedlichen Richtungen her Untersuchungen zu Schnittstellen durchführt – etwa aus den Plattform- und Interface-Studies (vgl. Gerlitz et al. 2019; Helmond/Nieborg/van der Vlist 2019; Hind/Seitz 2024) sowie den eher informationswissenschaftlich orientierten Security Studies (siehe etwa Kumar et al. 2018; Ford/Palmer 2019; Iqbal et al. 2022). Burrell (2016) unterscheidet unterschiedliche Formen von Opazität: erstens durch technisches Unverständnis entstehende Unklarheiten, zweitens durch firmenseitige Geheimhaltung entstehende Opazität und drittens technisch bedingte Intransparenzen, die im Charakter etwa von bestimmten algorithmischen Methoden des Machine Learnings angelegt sind. In ihrem Zusammenspiel erschweren sie die analytische Zugänglichkeit KI-basierter Anwendungen für die Sozial- und Geisteswissenschaften (vgl. Rohlfing et al. 2021: 717).

3.2.3 Smart Speaker als Intermediäre und Plattformen

Smart Speaker ermöglichen die stimmbasierte Nutzung unterschiedlicher Funktionen, die über die Systemanbieter (Amazon, Google und Apple) internetbasiert

sammenhängen und fließt somit auch in die Ausgaben der Smart Speaker und anderer Voice Assistants ein.

bereitgestellt werden. Dazu zählen erstens systeminterne Basisfunktionen, z. B. die Nutzung von Weckruf- und Timer-Funktionen, die Ansage der Uhrzeit und des Datums oder die Durchführung einfacher Rechenoperationen. Zweitens ermöglichen sie über Drittanbieter den Zugriff auf im Internet von Dritten bereitgestellte Angebote, z. B. auf Kommunikationsanwendungen, Wetterinformationen, Wissensdatenbanken, Nachrichten, Streamingplattformen und Kartendienste einschließlich der Verkehrsmeldungen. Teilweise greifen die Systemanbieter dabei direkt auf die Inhalte der Drittanbieter zu, etwa im Fall von ‚einfachen' Informationsabfragen (Wetter, Lottozahlen, Sportergebnisse), teilweise werden die Inhalte aber auch durch die Anwendungen der Drittanbieter vermittelt, etwa bei der Wiedergabe von persönlichen Mitteilungen aus Messenger-Diensten, der tagesaktuellen Meldungen eines bestimmten Nachrichtenanbieters oder der Suche nach Musiktiteln auf spezifischen Streaming-Diensten wie Spotify; dabei bevorzugen die Systemanbieter über die Standard-Einstellungen häufig eigene Dienste (z. B. den Streaming-Dienst „AppleMusic" im Fall des „HomePods" von Apple), sind aber meist nicht darauf limitiert. Drittens erlauben Smart Speaker teilweise den Zugriff auf Online-Shopping-Angebote (z. B. auf Amazon) und bieten die Möglichkeit, Einkäufe einschließlich des Bezahlvorgangs abzuwickeln und zu routinisieren. Viertens sind die Steuerung und Verwaltung von Smart-Home-Anwendungen möglich (u. a. Licht und Heizung, Jalousien, Küchengeräte und Unterhaltungsmedien). Die Geräte sind dabei internetbasiert mit den Benutzer-Konten der Anwender*innen verknüpft – sowohl mit den Accounts bei den Systemanbietern (z. B. Amazon- oder Google-Konto), als auch mit den Accounts bei eventuellen Drittanbietern (z. B. Spotify-Konto). Über die Systemanbieter-Konten sind sie auch verknüpft mit den Smartphones der Anwender*innen, die die stimmliche Bedienung der Geräte ergänzen, sowie mit den Smart-Home-Geräten. Darüber hinaus sind sie mit den Schnittstellen bzw. Anwendungen der Drittanbieter verbunden (vgl. Hector 2025).

In der interdisziplinären Medienforschung – insbesondere in den Medienwissenschaften, der Kommunikationswissenschaft und Publizistik sowie den Science and Technology Studies (STS) – sind Smart Speaker je nach fokussierter Funktion unterschiedlich konzeptualisiert und entsprechend mit verschiedenen theoretischen und methodologischen Zugriffen untersucht worden. Dabei lassen sich erstens Strömungen ausmachen, die Smart Speaker als Intermediäre gefasst haben, während sie zweitens als (Teil) digitale(r) Plattformen beschrieben wurden. Dabei sind unter unterschiedlichen Gesichtspunkten auch die Spezifika und Folgen ihrer Interfaces beleuchtet worden (dazu zählt auch die vorliegende Arbeit) – auf

weitere Veröffentlichungen dazu wird auch im Folgekapitel zur Nutzung und Aneignung der Geräte eingegangen.[100]

Ein eher kommunikationswissenschaftlich-publizistisch orientierter Forschungsstrang bezieht sich auf die Betrachtung von Smart Speakern als *Intermediäre*. Dieser Zugang konzentriert sich auf die Funktion von Smart Speakern als Vermittler zwischen bereitgestellten „Inhalten" auf der einen Seite und den Nutzer*innen auf der anderen Seite, wobei mit „Inhalten" redaktionell mehr oder weniger stark kuratierte, v. a. journalistische Formen gemeint sind, die über (Medien)intermediäre aufbereitet und vermittelt werden (vgl. Weidmüller et al. 2021: 2). Untersuchungen zu Smart Speakern als Intermediäre konzentrieren sich häufig auf deren Rolle bei der Wiedergabe von Nachrichteninhalten (vgl. Weidmüller et al. 2021) und Radioangebote (vgl. Gattringer/Handel 2021) sowie den Einfluss, den der Interface-Typ auf die Wahrnehmung der Informationen haben könnte (vgl. Weidmüller/Etzrodt/Engesser 2022; Mooshammer/Etzrodt 2022; Frehmann 2023). Dabei nehmen Weidmüller et al. (2021: 7) das Verhältnis von Anbietern bzw. Redaktionen journalistischer Inhalte mit den Gerätebetreibern (insbesondere Amazon, Google und Apple) und den Charakteristika der Geräte vor dem Hintergrund des Gebots der Transparenz im Medienstaatsvertrag (MStV) in den Blick und kommen zu dem Schluss, dass in wesentlichen Teilen keine Nachvollziehbarkeit über die Kriterien für Auswahl und persönliche Aufbereitung der Inhalte bestand und nur teilweise kenntlich gemacht wurde, welche Inhalte präsentiert wurden und woher sie stammen (vgl. auch Natale/Cooke 2021). Bei der Frage, inwieweit Nutzer*innen den wiedergegebenen Nachrichteninhalten vertrauen, zeigte sich in einer Interview-Studie von Weidmüller/Etzrodt/Engesser (2022) für bestimmte abgefragte Dimensionen der Vertrauenswürdigkeit ein Zusammenhang zwischen der stimmbasierten Bedienung der Interfaces und dem Grad, dem Nutzer*innen den wiedergegebenen Inhalten vertrauten. Zu ähnlichen Ergebnissen kommt auch Frehmann (2023) in ihrer auf Online-Umfragen basierenden Studie. Auch andere kommunikationswissenschaftlich ausgerichtete Untersuchungen liefern Hinweise darauf, dass die Vermittlung von Inhalten über VUIs einen Einfluss auf die Wahrnehmung dieser haben könnten – der hybride Charakter des Austauschs mit VUIs und der Status der ‚Gesprächsbeteiligung' (insbesondere in Mehrparteienkonstellationen) hatte dabei dem Anschein nach eine Auswirkung auf die Verarbeitung und Bewertung der Inhalte (vgl. Weidmüller 2022; Etzrodt 2022). Dies unterstreicht die Relevanz einer gesprächsanalytischen Untersuchung der Gesprächs‚beteiligung' von VUIs, die hier als Desiderat identifiziert werden kann.

100 Zur Interface-Forschung siehe Kap. 2.1.4.

Ein anderer interdisziplinärer Forschungszweig untersucht Smart Speaker als *Plattformen* sowie als Bestandteil digitaler *Plattformsysteme* und nimmt somit die Rolle und Motive der Systemanbieter in den Blick, die den Zugriff auf Smart-Speaker-Funktionen bereitstellen und Verbindungen zu Drittanbietern herstellen (vgl. Strüver 2023b). Digitale Plattformen und ihre jeweiligen Konfigurationen können zwar je nach Zweck sehr unterschiedlich aussehen, sie dienen jedoch immer dem Ziel, verschiedene Teilnehmende oder Gruppen von Teilnehmenden zusammenzubringen und dort einen mehr oder weniger stark präfigurierten und zielgerichteten Austausch zu ermöglichen (vgl. van Dijck 2013: 29; für einen Überblick siehe van der Vlist 2022: 32–40). Plattformen sind dabei keine neutralen Instanzen, sondern durch Daten, Algorithmen, Eigentumsrechte und Geschäftsmodelle beeinflusst (vgl. van Dijck/Poell/Waal 2018: 325; siehe auch Gillespie 2018). Deren Anbieter prästrukturieren nicht nur durch die Interface-Gestaltung die Art des möglichen Austauschs (siehe unten), sondern sie verfolgen auch das Ziel einer Monetarisierung der Teilnahme der Nutzer*innen auf unterschiedliche Weisen (vgl. Srnicek 2016), primär durch die Auswertung von Nutzungsdaten. Smart Speaker sind damit Teil einer erweiterten Wertschöpfungskette, der u. a. auch ihre Nutzer*innen angehören (vgl. Crawford/Joler 2018). Die Quelle für den Gewinn der Plattformen ist dabei zu einem Großteil personalisierte Werbung, ein Geschäftsmodell zahlreicher Internet-Firmen (vgl. van der Vlist 2022: 25). Dies geht von der Annahme aus, dass die Daten Nutzungsmuster offenbaren, die Aufschluss über die private Lebensführung der Nutzer*innen sowie die konsumierten Produkte geben, die wiederum zur Bewerbung anderer Produkte genutzt und an Drittanbieter verkauft werden können (vgl. Sadowski 2020: 116; Khan 2018: 118–119). Ein komplementärer Erklärungsansatz ist, dass die aufgezeichneten Stimmeingaben und Plattformzugriffe, die bei jeder Nutzung des Smart Speakers erfasst werden, außerdem als Bestandteil der Weiterentwicklung von datenintensiven Systemen zur automatischen Spracherkennung verwertet werden (vgl. Crawford/Joler 2018). In diesem Sinne sind Smart Speaker Bestandteil digitaler Plattform-Ökosysteme.[101]

Strüver (2023b) zeigt darüber hinaus, dass es bei Amazon und dem VUI Alexa *insbesondere* die Bedeutung des Smart Speakers im Plattformsystem ist, die das Betreiben der Smart-Speaker-Anwendung für die Systemanbieter attraktiv macht: Über die Nutzung der VUIs kann Amazon als Plattformbetreiber neue Erkenntnisse über die Anwender*innen und deren Nutzungsverhalten gewinnen, mögliche neue Anwendungen oder Angebote identifizieren und

101 Implikationen für die Themenbereiche Datenschutz, -verwertung und Privatsphäre, die auch über die Plattformisierung hinaus reichen, werden in Kap. 3.2.6 diskutiert.

diese zugleich an den Nutzer*innen fortlaufend testen und verbessern (siehe auch Marres/Stark 2020). Wie Strüver (2023b) weiter herleitet, ist es daher auch der Anspruch der Betreiber, eine möglichst positive Nutzungserfahrung zu ermöglichen, um die Test- und Verbesserungsmöglichkeiten verschiedener Dienste und den Einblick in die Nutzungsweisen der Anwender*innen über den Betrieb des Smart Speakers weiter aufrecht zu erhalten (siehe auch Dahlgren et al. 2021). Eine Strategie, die Stellung des VUIs bei den Nutzer*innen zu stabilisieren, ist die obligatorische Einbindung des VUI in Smart-Home-Anwendungen. Die Nutzung des Smart Speakers von Amazon zur Steuerung des Smart Home verspricht nicht nur eine reibungslose Einbindung neuer Smart-Home-Anwendungen in den Haushalt und dessen Smart-Home-Konfiguration (vgl. Strüver 2023a), sondern auch eine fortlaufend komfortable und ‚unsichtbare' Bedienung dieser Geräte (vgl. Strüver 2023b). Amazon stellt dazu eine möglichst große Kompatibilität mit Drittanbietern sicher und kann so als feste Instanz im Smart Home umfassend Informationen über die Nutzung der Dienste und mögliche Verbesserungspotenziale gewinnen. Zugleich werden zahlreiche Alltagspraktiken der Haushaltsmitglieder unter Beteiligung der zentralen Smart-Home-Steuerung vollzogen, was sie anfällig für die Beeinflussung und Normierung durch die Plattformbetreiber macht (vgl. Goulden 2019). Damit sind Smart Speaker nicht nur Bestandteil eines Plattform-Ökosystems, sondern auch selbst nicht-neutrale Plattformen innerhalb der Haushalte ihrer Anwender*innen.

3.2.4 Nutzung und Aneignung von Smart Speakern

Empirische Studien zur Nutzung und zur longitudinalen Perspektive auf die Nutzung und Aneignung von Smart Speakern sind bislang noch rar. In einer repräsentativen Befragung zwischen Oktober 2023 und September 2024 in Deutschland gaben 25 Prozent der Befragten an, einen Smart Speaker zu besitzen (vgl. Statista 2025b); dieser Wert ist in vergleichbaren Umfragen in den vergangenen fünf Jahren stabil. Im Folgenden sollen auf Grundlage empirischer Arbeiten die zuletzt beobachteten Nutzungsweisen, mögliche Einflussfaktoren und Aneignungs- bzw. Domestizierungsprozesse[102] in den Haushalten herausgestellt werden, um dabei die mit der vorliegenden Arbeit zu füllende Forschungslücke noch präziser zu fassen. Dabei wird nicht übersehen, dass etliche Studien sich mit VUIs auseinandersetzen, aber nicht spezifisch Smart Speaker in den Blick nehmen. So beschäftigt sich etwa die Interview-Studie von Luger/Sellen (2016) mit „Conversational

102 Siehe Kap. 2.3.

Agents“. Es wurden dabei aber vornehmlich Nutzer*innen von Siri, Google Now und Cortana befragt, die diese nicht über Smart Speaker, sondern über andere Endgeräte wie Smartphones, Tablets und Computer bedienen (vgl. Luger/Sellen 2016: 5289). Die herausgearbeiteten primären Nutzungsszenarien wie z. B. eine Tastatureingabe zu vermeiden, weil man unterwegs ist (vgl. Luger/Sellen 2016: 5291), sind insofern nicht übertragbar. Gleichwohl könnte der festgestellte „gap between user expectation and system operation“ (Luger/Sellen 2016: 5295) durchaus auch bei Smart Speakern eine Rolle spielen (siehe dazu Strüver 2020).[103]

Etliche Studien zu Smart Speakern haben zunächst Potenzialanalysen vorgenommen, die basalen Funktionen vorgestellt und mögliche Anwendungsfälle diskutiert (vgl. Goksel-Canbek/Mutlu 2016; Bedford-Strohm 2017; Hoy 2018). Eine auf Smart Speaker scharfgestellte Interview-Studie mit 19 Interviews von Nutzer*innen und unter Einbezug von 170 Protokollaufzeichnungen, die in der zugehörigen Smartphone-App hinterlegt werden,[104] legen Ammari et al. (2019) vor. 18 der 19 Befragten hatten einen Smart Speaker zu Hause und Anspruch der Studie war u. a. die Untersuchung der alltäglichen Nutzungskontexte. Diese sind vorhersehbarerweise abhängig von den konkreten Nutzer*innen, es zeichnet sich jedoch hier deutlich ab, dass ‚einfache‘ Anwendungsfälle mit routinisiertem Charakter (z. B. Abspielen von Musik, Timer beim Kochen, Erinnerungen oder Wetter-Abfragen) dominieren. Einen wesentlichen Anteil nimmt auch die Steuerung anderer Smart-Home-Anwendungen ein, die ebenfalls nach festen Mustern erfolgt, so z. B. das Licht ein- bzw. ausschalten oder die Heizung regulieren (vgl. Ammari et al. 2019: 15). Die Dominanz dieser Nutzungstypen bestätigt sich auch einer Befragungsstudie von Lopatovska et al. (2019). In neun verschiedenen Haushalten wurden über einen Zeitraum von jeweils vier Tagen insgesamt 136 Anfragen an einen Amazon *Echo* gestellt, von denen 86 Wetterabfragen, Musikwiedergabe und die Kontrolle anderer Smart-Home-Geräte waren.

Das Spektrum von Anwendungsfällen, das die ebenfalls interviewbasierte, qualitativ ausgerichtete Studie von Brause/Blank (2020) mit dem Konzept der „Use Genres“[105] aus der Domestizierungstheorie (siehe dazu Bakardjieva 2005) un-

103 Zum Aufbau von Nutzungserwartungen an Conversational Agents in der Werbung siehe auch die Analyse von Hennig/Hauptmann (2019).

104 Siehe dazu Kap. 3.3 sowie Habscheid et al. (2021).

105 „Use Genres“ sind ein Konzept aus der Domestizierungsforschung, mit dem Nutzungsweisen von Technologien aus ihrer konkreten, praktischen Situierung heraus beschrieben und verstanden werden (vgl. Bakardjieva 2006: 73). Solche ‚Gattungen‘ der Nutzung sind in diesem Sinne nicht etwas von den Herstellern oder durch die Technologie Angelegtes, sondern entstehen dynamisch und in Relation zu den Handlungen der Nutzer*innen (vgl. Brause/Blank 2020: 753).

terscheidet, ist deutlich weiter aufgefächert. Es gehören dazu „companionship, self-control and productivity, health care support, better sleep, peace of mind and improved accessibility“ (Brause/Blank 2020: 759) sowie Sicherheitsvorkehrungen und sogar Ausspionierungen in einer Ehe. Besondere Beachtung findet eine räumlich verteilte Nutzung, d. h. Anwendungsfälle, die über den direkten Austausch zwischen dem Menschen und der Instanz des Smart Speakers hinaus wirken – so z. B. die Steuerung von Smart Home-Anwendungen, die Folgen für den Status anderer, vernetzter Geräte haben, oder das Erhalten von Nachrichten auf dem Mobiltelefon (vgl. Brause/Blank 2020: 757). Die Autor*innen entwickeln die „Use Genres“ aus den Interviewdaten heraus und liefern insofern sehr wertvolle Hinweise auf mögliche Verwendungsweisen, die auch in den im Analyseteil dieser Arbeit untersuchten Daten eine Rolle spielen. Dass die Nutzer*innen selbst diese Kategorien relevant machen, macht die Bestandsaufnahme im Sinne einer ethnomethodologischen Vorgehensweise besonders wertvoll.[106]

Die Inbetriebnahme sowie die Integration anderer Geräte und Konten sind nicht immer unproblematisch und störanfällig, wie auch anhand der später präsentierten Daten deutlich wird. In Interviews der verschiedenen Studien wird dargestellt, dass die Übersichtlichkeit in den zugehörigen Smartphone-Apps verschiedener Hersteller nicht unbedingt gegeben ist; ferner wird berichtet, dass bei mehreren Smart Speakern die räumliche Zuordnung Probleme verursacht und z. B. das Licht im falschen Raum ausgeschaltet wird (vgl. Ammari et al. 2019: 16). Die fortlaufende, routinisierte Nutzung über einen Zeitraum von drei bis sechs Monaten hinweg scheint jedoch insgesamt bei den genannten, ‚eingespielten‘ und in die Alltagspraktiken der Nutzer*innen übergegangenen Anwendungsfällen zu liegen, die primär eher ‚einfache‘ Nutzungsszenarien umfassen. In Befragungen zeigte sich zudem eine Tendenz, dass die Nutzungshäufigkeit des Smart Speakers über längere Zeit nach Anschaffung eher abnimmt (vgl. Lopatovska et al. 2019: 991; siehe auch Barthel/Helmer/Reineke 2023), andere nutzer*innenzentrierte Faktoren (insbesondere Alter und Technologieaffinität) schienen jedoch eher geringe Auswirkungen auf die Nutzung von Smart Speakern zu haben (vgl. Lopatovska et al. 2019: 990).

Die Tageszeit als Einflussfaktor auf die genutzten Anwendungen diskutieren Bentley et al. (2018) in einer empirisch breit angelegten Studie in 88 verschiedenen Haushalten, in denen über einen durchschnittlichen Beobachtungszeitraum von 110 Tagen insgesamt 65.499 Aktionen mit Smart Speakern ausgewertet wurden. Sie bestätigen, dass Musikwiedergabe, Wissensabfragen und Smart-Home-Anwendungen zu den am häufigsten genutzten Kategorien gehören. Sie bestätigen

106 Siehe dazu Kap. 4.1.

auf dieser Basis allerdings nicht die hohe relative Häufigkeit von Wetterabfragen. Auch wenn die Gesamtverteilung der genutzten Kategorien im Tagesverlauf recht konstant bleibt, beobachten die Autor*innen Höhen für Wetterabfragen am Morgen zwischen sechs und neun Uhr und für die Bedienung von Smart-Home-Elementen (insbesondere wohl Licht und Heizung) in den Abendstunden (vgl. Bentley et al. 2018: 7). Uhrzeitabfragen haben eine Spitze zwischen drei und neun Uhr am Morgen „likely as users lay in bed wondering how many hours they have left before they need to get up“ (Bentley et al. 2018: 7). Am Wochenende findet außerdem sehr viel häufiger eine Nutzung der Smart Speaker statt. Weitere demografische Einflussfaktoren wie Alter oder die Größe des Haushalts werden zwar auch in dieser Studie diskutiert, die Nutzungsszenarien bleiben aber im Wesentlichen konstant (siehe dazu und für die folgenden Angaben Bentley et al. 2018: 16–18). In der longitudinalen Betrachtungsweise zeigte sich, dass die Nutzungsszenarien überwiegend konstant verteilt blieben, dass aber Smart-Home-Anwendungen tendenziell zunehmen und die Musiksteuerung demgegenüber tendenziell abnimmt. Die Länge der Inputs (und somit ein Indikator für die Komplexität derer) nimmt über die Zeit zu, korreliert aber auch mit der Zunahme von Wissensabfragen, für die längere Inputs notwendig sind, insbesondere im Vergleich zu kurzen, routinierten Befehlen wie Uhrzeitabfragen. Die Daten dieser größeren Studie belegen die bereits zuvor geäußerten Vermutungen noch einmal eindrücklich: Zusammenfassend sind routinierte Anwendungsfälle mit kurzen und meist zweizügigen Ein-/Ausgabe-Sequenzen fest in den Alltag integriert und weisen sogar über den Tagesverlauf eine gewisse Stabilität auf. Neue Anwendungen werden selten ausprobiert, einzig die Steuerung von Smart-Home-Anwendungen nimmt nach der Anschaffung tendenziell weiter zu.

Die bisher genannten Studien bezogen sich auf den US-amerikanischen Raum. Zur Nutzung und Aneignung von Smart Speakern forschten für den deutschsprachigen Raum mit einem qualitativ-ethnografischen Untersuchungsdesign Pins et al. (2020), die u. a. Medientagebücher einsetzten, um die Nutzung von Smart Speakern empirisch zu erforschen. Sie bestätigen in kleinerem Rahmen die Ergebnisse zur Nutzung von Smart Speakern von Ammari et al. (2019) dahingehend, dass auch in deren Daten die ‚einfachen' Anwendungsfälle deutlich überwogen und kompliziertere Nutzungskontexte mit mehreren Zügen und neue Skills in vielen Haushalten nicht ausprobiert wurden. In zehn Interviews, die auf Basis der Grounded Theorie (Strauss/Corbin 2010) ausgewertet wurden, zeigte sich, dass die Smart Speaker „Bestandteil des häuslichen Lebensmittelpunktes“ wurden (Pins et al. 2020: 353), indem sie an zentralen Stellen und in gemeinsam von allen Haushaltsmitgliedern genutzten Räumen aufgestellt wurden; dort sind sie „als fester Bestandteil in ihrem Alltag eingebettet“ (Pins et al. 2020: 357). Außerdem zeigte sich eine Anpassungsleistung an die VUIs seitens der Befragten, die

angaben, die Formulierung der ‚richtigen' Stimmeingaben und eine bestimmte Sprechweise über die Zeit erlernt zu haben; in einem befragten Haushalt gibt es eine geteilte Liste mit Befehlen für bestimmte Zwecke (Pins et al. 2020: 354). Smart Speaker lassen sich zudem anpassen: Individuelle Befehle können programmiert und mit bestimmten Funktionen verknüpft werden (z. B. „Schicht im Schacht" schaltet eine bestimmte Lampe aus). Pins et al. (2020: 356) konstatieren, dass „die Interaktion sich häufig auf kurze, geschlossene Sequenzen reduziert – hauptsächlich deshalb, weil Nutzer an komplexeren Interaktionen scheiterten". Sie stellen eine Lücke zwischen einem prinzipiell breiten möglichen Nutzungsspektrum fest, das jedoch nur zu einem Bruchteil ausgeschöpft wird, und benennen als limitierenden Faktor klar die Sprachsteuerung, die erstens zu häufig keine gute Erkennung der Sprache liefere und zweitens zu wenig Freiheiten in der Auswahl der benötigten Inhalte lasse (vgl. Pins et al. 2020: 356; siehe auch Habscheid/Hector/Hrncal 2025).

Zum Potenzial der Protokolldaten der Assistenzsysteme arbeiten Habscheid et al. (2021) und stellen in ihrer qualitativ-explorativen Untersuchung heraus, dass dieser Datentyp zwar einerseits Aufschlüsse darüber geben kann, wie der Austausch zwischen Nutzer*innen von VUIs und den Plattformen konfiguriert ist. Es entsteht ein Nutzungsprotokoll, das bestimmte Datentypen enthält und zur Bearbeitung und Löschung vorsieht, während sich andererseits der Export des Protokolls einschließlich seiner Audio-Mitschnitte umständlich gestaltet. Für eine detaillierte Betrachtung der Einbettung von VUIs in die soziale Praxis greift eine ausschließlich auf diesen Protokollen basierende Analyse allerdings zu kurz: die Erfassung durch das Protokoll ist dafür zu bruchstückhaft (vgl. Habscheid et al. 2021; Hector 2025: 75). Diese Erkenntnisse werfen wichtige Anschlussfragen auf, zu deren Beantwortung die vorliegende Arbeit einen Beitrag leisten will: Mit welchen Praktiken passen sich die Nutzer*innen den Anforderungen der Smart Speaker an, welche übernehmen sie in ihren alltäglichen Gebrauch? Welche Strategien zur Fehlerbehebung sind empirisch beobachtbar? Welcher Status wird den Geräten in laufenden sozialen Interaktionen zugeschrieben? Die Studie von Pins et al. (2020) zeigt auf Basis der Interviews, dass genau diese Punkte bei der Bedienung der Smart Speaker eine Rolle spielen und die Technologie die Alltagspraktiken der Nutzer*innen verändern kann. Als explorative Studie lieferte sie insofern sehr wertvolle Impulse für die folgende Arbeit; der genaue Vollzug und die Einbettung in den Alltag jedoch kann nur auf Basis von Aufnahmen *in situ* untersucht werden.

Dieser Aufgabe gehen nur einige wenige Arbeiten nach, deren Befunde nachfolgend zusammengefasst werden sollen. Porcheron et al. (2018) explorieren in ihrer sehr einschlägigen Studie, auf welche Weise Smart Speaker in die sozialen und interaktionalen Zusammenhänge des Alltags eingebettet werden. Die Unter-

suchung findet auf Basis von ein-monatigen Audio-Aufnahmen in fünf Haushalten und insgesamt ca. sechs Stunden Audiomaterial statt, das mit Hilfe eines sog. Conditional Voice Recorders (CVR) zusammengetragen wurde.[107] Diesem Material nähern sie sich mit einer konversationsanalytischen und an die HCI und CSCW angebundenen ethnomethodologischen Perspektive. Dabei sind die Schlüsselergebnisse eher auf eine Design-Perspektive zugeschnitten und geben konkrete Ratschläge an zukünftige Design-Projekte von VUIs (vgl. Porcheron et al. 2018: 10), die für die vorliegende Studie weniger einschlägig sind; von besonderem Interesse sind aber die Detail-Analysen. In diesen wird unterschieden zwischen der Einbettung von Amazons *Echo*[108] in die häuslichen Alltagsaktivitäten auf der einen und der Einbettung in die sequenzielle Organisation auf der anderen Seite. Die Adressierung und Einbindung der Nutzung des Smart Speakers in den häuslichen Alltag erfolgt, so konstatieren Porcheron et al. (2018: 5), „with relative ease through everyday talk", was v. a. auf der Annahme fußt, dass etwa das gemeinsame Abendessen in einer Familie ohnehin eine „multi-activity"-Situation ist, bei der sich konstant mehrere Handlungsstränge überlappen und insofern die Adressierung und Einbindung eines VUI einen von mehreren dieser Stränge darstellt und nicht heraussticht. Sie beobachten, dass bei der Aushandlung von Zugriffsrechten auf den Smart Speaker kompetitive Momente beobachtet werden können, die konversationellen Methoden der Rederechtsaushandlung gleichen. Das Herstellen von „accountability" (Garfinkel 1967: 34) erfolgt, ohne eine hohe Spezifität für genau diese Art der Darstellungen, der „accounts" (Garfinkel 1967: 2), feststellen zu können, die sich auf Äußerungen zu Adressierungen eines Smart Speakers beziehen: „such utterances are treated in similar kinds of ways to the ways that all social actions are treated: as accountable to the situation they are in" (Porcheron et al. 2018: 6).[109]

Deutlich spezifischere Eigenschaften scheint die Einbindung der Geräte Porcheron et al. (2018) zufolge nicht in die Aktivitäten im Allgemeinen, sondern in die sequenzielle Gestaltung des *turn-by-turn talk* aufzuweisen. Für die Gestaltung der Anfrage an den Smart Speaker (etwa eine Frage oder ein Befehl) ist zunächst auffällig, dass in den Porcheron et al. vorliegenden Daten die Äußerung des Aktivierungsworts als eine selbstständige *turn-constructional unit* betrachtet wird,

107 Ein Conditional Voice Recorder ermöglicht es, Audioaufnahmen von VUI-Dialogen zu erheben, auf denen nicht nur die Nennung des Aktivierungsworts und der unmittelbar folgende Teil, sondern auch eine vordefinierte Zeitspanne vor- und nachher aufgezeichnet wird, siehe dazu Kap. 5.2.

108 Andere Hersteller bzw. Modelle werden nicht berücksichtigt.

109 Zu „acountability" bzw. „accounts" als ethnomethodologische Grundbegriffe siehe Kap. 4.1 sowie Bergmann/Meyer (2021b).

nach der ein Sprecher*innenwechsel stattfinden kann. Die kurze Pause zwischen Aktivierungswort und Anfrage wird mehrfach für andere Äußerungen genutzt, die keinen Input an den Smart Speaker darstellen; diese Beobachtung erinnert an die von Pitsch et al. (2017) vorgelegten Befunde zur interaktionalen Dynamik bei der Bedienung von Robotern, in denen der Interaktionsraum ebenfalls auf Phasen der Interaktion *vor* dem Roboter ausgedehnt wurde, in denen die Nutzer*innen den Roboter selbst aber nicht adressieren. Die Pause kann außerdem genutzt werden, um die Steuerung des Smart Speakers von einer anderen Person zu übernehmen, d. h. selbst einen Input an das VUI zu formulieren, bevor der ursprünglich aktivierende Interaktant dies kann (Porcheron et al. 2018: 7). Darüber hinaus scheint gemeinsames Herstellen von Stille für die Produktion einer Anfrage ein übliches Verfahren zu sein, bei der der *turn-by-turn talk* vorübergehend ausgesetzt und nur der Input formuliert wird. Dies trägt auch zu einer besseren Hörbarkeit der Ausgabe bei. Bei Fehlverarbeitung der Anfrage lässt sich ferner eine Art „sequenzielle Kollaboration" beobachten, in der verschiedene Teilnehmer*innen durch prosodische oder lexikalische Variation versuchen, die gewünschte Funktion aufzurufen (vgl. Porcheron et al. 2018: 8–9).

Im Umgang mit den Outputs konzentrieren sich Porcheron et al. (2018: 7–9) auf Störungen der Geräte. So werden, ähnlich wie in zwischenmenschlichen Interaktionen, zu lange Pausen als Hinweise auf Probleme interpretiert, wobei die Toleranz für Zögerungen im Austausch mit VUIs deutlich höher zu sein scheint. Gibt das VUI eine unerwünschte Ausgabe zurück oder formuliert, dass die Eingabe nicht verstanden wurde, kann ein „mismatch" (Porcheron et al. 2018: 8) zwischen der Fehlermeldung des Geräts und der interaktionalen Bearbeitung dieser beobachtet werden. Letztere konzentrieren sich als Reparaturversuche immer wieder auf prosodische und lexikalische Variationen, wobei die prosodischen Strategien im Rahmen der Arbeit von Porcheron et al. (2018) nicht betrachtet werden. Insgesamt kommen sie zu dem Schluss, der Austausch mit VUIs sei „fundamentally different from human interaction" (Porcheron et al. 2018: 9). Der Unterschied ergebe sich primär daraus, dass sich der sequenzielle Verlauf des Dialogs nicht aus der Beziehung der jeweiligen Sequenzteile aufeinander und der darin liegenden Interpretation der vorherigen Äußerungen entfaltet, sondern vielmehr durch die Eingabe-Ausgabe-Struktur vorgegeben ist (vgl. Porcheron et al. 2018: 9).

Die Arbeit von Porcheron et al. (2018) liefert, auch wenn sie die Ableitung von Design-Empfehlungen aus diesen Analysen fokussiert, entscheidende Anknüpfungspunkte für weitere Arbeiten, die die Nutzung von Smart Speakern im praktischen Vollzug beleuchten. So beschreiben Beneteau et al. (2019: 4–5) auf Basis einer Audio-Erhebung bei 10 Familien sechs verschiedene Reparaturstrategien: prosodische, artikulatorische, syntaktische und semantische Anpassungen sowie gesteigerte Lautstärke und Wiederholungen. Die Autor*innen dieser Unter-

suchung analysieren drei Fallbeispiele mit einer ebenfalls konversationsanalytischen Mentalität, um diese Strategien zu illustrieren, und empfehlen zukünftigen VUI-Designs, Fehler und Reparaturen als Fokus der Entwicklung festzulegen (vgl. Beneteau et al. 2019: 10), sowohl in den Bereichen Code-Switching und diskursiver Ordnung wie auch in der Teilnehmenden-Struktur. So raten die Autor*innen dazu, seitens des VUI eine stärkere Dialogsteuerung und präzisere Nachfragen zu realisieren und dabei das Primat der Natürlichsprachigkeit abzulegen (um z. B. nach Keywords zu fragen, die vom VUI einfacher verarbeitet werden können) (vgl. Beneteau et al. 2019: 11). Auch empfehlen die Autor*innen, dass das VUI aktiv die Teilnehmenden-Struktur der Dialogbeteiligten abfragen und steuern könnte, um die Diversität der angewendeten Reparaturstrategien zu erhöhen; diese Empfehlung leitet sich aus der Beobachtung ab, dass bei Beteiligung mehrerer Personen an Smart-Speaker-Dialogen auch eine größere Diversität im Hinblick auf die Anwendung von Reparaturstrategien bestehe (siehe auch Beneteau et al. 2020a; Beneteau et al. 2020b).

Habscheid (2022) diskutiert auf der Grundlage des auch in der vorliegenden Arbeit verwendeten Datenkorpus aus Video-Aufzeichnungen von Inbetriebnahmesituationen und Audio-Aufzeichnungen von der regelmäßigen Nutzung der Systeme[110] die wechselseitigen Anpassungen von Nutzer*innen und VUI aneinander. Dabei fokussiert er auf Situationen des Wechsels zwischen soziotechnischem Dialog und „Meta-Interaktionsraum“ (Habscheid 2022: 168), wobei in Letzterem die Nutzer*innen über die VUIs und den Gebrauch reflektieren, während im ersten der Austausch mit dem VUI im Vordergrund steht. Er stellt heraus, dass Sequenzialität – über Adjazenzpaare hinaus verstanden als „as a fundamental resource of meaning constitution“ (Habscheid 2022: 191) – eine Herausforderung für die Geräte darstelle. Die indexikalische soziale Interaktion der Nutzer*innen zu erfassen bleibt für VUIs bisher eine Grenze der Leistungsfähigkeit und insofern die Anpassungsleistung der Nutzer*innen elementar für das Funktionieren des soziotechnischen Dialogs – soziale Interaktion bleibe eine Simulation der Teilnehmenden und dies aufzudecken ist eine der wesentlichen Vorteile einer praxeologischen Analyse (siehe auch Alač et al. 2020; Hector 2022; Habscheid/Hector/Hrncal 2025).

Eine praxeologische Perspektive wird auch bei Habscheid/Hector/Hrncal (2023) angewendet, um einen gesprächsanalytisch fundierten Beitrag zur Debatte um Agency menschlicher und nicht-menschlicher Entitäten zu leisten. In den qualitati-

110 Die Daten wurden mit einem von Porcheron et al. (2018) adaptierten CVR erhoben, siehe dazu ausführlicher Hector et al. (2022) sowie Kap. 5.

ven Analysen, die sich wiederum auf Beispiele aus dem bereits zuvor erwähnten Datenkorpus aus Inbetriebnahme- und Nutzungssituationen beziehen, kann gezeigt werden, dass Agency zwischen Nutzer*innen und VUIs auch als „dynamic practical accomplishment" (Habscheid/Hector/Hrncal 2023: 23) betrachtet werden kann. Agency und ihre Aushandlung ist dabei aber kein Nullsummenspiel: Die Steigerung der Agency der Nutzer*innen muss nicht mit einem Verlust an Agency auf Seiten des VUIs einhergehen und umgekehrt (vgl. Habscheid/Hector/Hrncal 2023: 24). Eine einzelne Situation kann sowohl Verlust wie auch Steigerung von Agency auf jeder der beiden Seiten bedeuten – zugleich ist die Handlungsmacht der Nutzer*innen sowohl untereinander als auch mit dem des VUIs verbunden, wenn auch nicht determiniert. Der Blick auf den lokalen, situationalen Kontext und die soziale Konfiguration der Beteiligten untereinander ist aber eine wesentliche Einflussgröße. Herausgefordert wird diese allerdings durch die teilweise intransparente und im Hintergrund stattfindende Verwertung der Daten, wie auch Waldecker/Hector/Hoffmann (2024) herausarbeiten. Die Nutzer*innen sind dann zwar *in der Nutzungssituation* möglicherweise mit Agency ausgestattet und beherrschen etwa den konversationellen Floor (vgl. Waldecker/Hector/Hoffmann 2024: 11). Das Interface suggeriert dies zusätzlich über Möglichkeiten der Löschung einzelner Aufnahmen aus dem Protokoll (vgl. Habscheid et al. 2021: 19). Die Nutzer*innen sind aber zugleich nicht mehr Teil des praktischen Vollzugs der Auswertung – Grenzen zwischen Öffentlichkeit und Privatheit werden in diesem Sinne neu verhandelt (vgl. Waldecker/Hector/Hoffmann 2024: 2). Die Autor*innen konstatieren zwei verschiedene Typen von „Agencies": Einerseits sind die Nutzer*innen „able to retain control or a sense of superiority even in complicated situations" (Waldecker/Hector/Hoffmann 2024: 13). Andererseits ist schon die Untersuchung von Agency deutlich erschwert, wenn es um Datenpraktiken geht, die im Zuge der Nutzung und weit darüber hinaus im Hintergrund stattfinden. Auch hier konnte eine praxeologische Analyse helfen, diese Differenzen aufzudecken und mit empirischem Material zu unterfüttern.

3.2.5 Smart Speaker und Gender

Ein wachsendes Forschungsfeld beschäftigt sich interdisziplinär mit Smart Speakern vor dem Hintergrund der sozialen Konstruktion von Gender. Ausgangspunkt ist in vielen Arbeiten der Umstand, dass die meisten VUIs und insofern auch Smart Speaker per Default eine weiblich klingende Stimme eingestellt haben und die zugehörigen Interface-Personae zudem weiblich gelesen sind (etwa Alexa umd Siri). Dies ist aus verschiedenen Disziplinen kritisch diskutiert und analysiert worden (vgl. Natale/Cooke 2021: 1010; Woods 2024: 98–130). Bereits in einer frühen Arbeit, erst ein

Jahr nach dem Start von Apples Sprachassistent Siri, kommt Both (2012: 130) zu dem Schluss, „dass Siri überwiegend auf weiblich konnotierte Register zurückgreift" (siehe auch Both 2011). Both nimmt die „Geschlechter-Performanz auf der Ebene des Dialogs" (Both 2012: 129) in den Blick und untersucht sprachliche Register.[111] Im Hinblick auf die mit einem „weiblichen Stil" programmierte Dialogizität und die weiblich gestaltete Anthropomorphisierung der VUI scheinen die Ausführungen zunächst überzeugend und liefern wertvolle Hinweise etwa auf einen „passiven Eindruck", eine kooperative Haltung und Bescheidenheit und Gefühlsbetontheit als (vermeintlich) weibliche Merkmale. In einer späteren Arbeit konstatiert Both (2014: 109): „Siri's conversational style draws on a stereotypical female image of altruistic and cooperative behavior". Boths Darstellung zu den weiblichen sprachlichen Registern oder dem ‚konversationellen Stil' entbehrt jedoch einer soliden theoretischen Basis zu femininen Stereotypen. Es mangelt ferner an einer dialogischen bzw. interaktionistischen Perspektive, in der ‚Doing Gender' als gemeinsam hergestellte Bezugskategorie der Beteiligten betrachtet wird (West/Zimmermann 1987): Zwar argumentiert Both, dass die sprachlichen Register „eingeschrieben" sind, doch ob diese tatsächlich von den Programmierer*innen auf der einen und von den Nutzer*innen auf der anderen Seite als Gender-Merkmal relevant gemacht werden, kann auf diese Weise nicht ermittelt werden und muss Gegenstand weiterer Forschungen sein.

Gleichwohl scheint nicht nur auf Basis dieser Studie einiges darauf hinzudeuten, dass durch die Auswahl der weiblich klingenden Stimme Geschlechterstereotype manifestiert und evtl. sogar ausgeweitet werden, wie frühe Arbeiten aus einer feministischen Perspektive (Draude 2006; Gustavsson 2005; Weber/Bath 2007) schon für ähnliche digitale Systeme feststellen. Es besteht das Potenzial, dass genderspezifische Stereotype zur Aufteilung von Arbeiten im Haushalt und Care-Arbeit sich durch die Zuordnung einer weiblichen Stimme, aber auch durch die Gestaltung von Werbung reetablieren (vgl. Hennig/Hauptmann 2019; Strengers/Kennedy 2020; Sadowski/Strengers/Kennedy 2021): „the 21st-century smart home maintains current gendered stereotypes by promising wife or housekeeper functionality" (Strengers/Nicholls 2018: 75). Diese auch von Both (2012; 2014) untersuchte „Einschreibung einer geschlechtsspezifischen Arbeitsteilung" (Both 2012: 132) verweist darauf, dass die weiblich gelesene Personifikation der Geräte mit den übernommenen Aufgaben im Service-, Haushalts- und Care-Bereich

111 Auch wenn Both (2012) diesen Begriff hier verwendet, kann hier nicht von sprachlichen Registern im Sinne der linguistischen Anthropologie gesprochen werden, wie sie etwa Agha (2004) versteht; diese werden auch nicht methodisch untersucht (siehe dazu Kap. 7). Both selbst verwendet mit Verweis auf Lübke (2005) auch den Stilbegriff, ohne sprachwissenschaftliche Implikationen hier zu reflektieren.

korrespondiert, die traditionell von weiblichen Personen ausgeführt werden (vgl. Gustavsson 2005). Sie sind insofern als servile, hilfsbereite und zurückhaltende Assistentinnen inszeniert, die sich nur zu den ihr übertragenen Aufgaben äußern, sich aber darüber hinaus nicht ins Geschehen involvieren und insofern an Hauspersonal erinnern (siehe auch Strüver 2020: 1–10; Habscheid et al. 2021; Dickel/Schmidt-Jüngst 2021; Waldecker/Volmar 2022: 172). So manifestieren und erweitern sich, so die Argumentation, durch die Nähe zu menschlichen, weiblichen Stimmen, Gender-Stereotype (vgl. Natale/Cooke 2021: 11) und gegenderte Normen und Phantasien (vgl. Strengers/Sofoulis 2024; Woods 2024: 98–130).

Natale/Cooke (2021: 1010) führen aus, dass diese Anthropomorphisierung im Dienst von kapitalistischen Interessen steht (siehe auch Natale 2023). Um Sprachassistenten marktfähig zu machen, muss die neue Technologie in bekannte Repräsentationen und Stereotypen eingebettet werden. Dabei greifen die Designer*innen auf diese Kategorien zur Aktivierung von Identitätsmarkern und sozialen Skripts zurück, die einen Roboter „abbilden“ können (vgl. Clark/Fischer 2022), auch wenn sie diesen Prozess selbst nicht unbedingt benennen können (vgl. Sweeney 2021: 153–155; siehe auch Young 2019: 117). Er entfaltet seine Wirkung auch erst im Zusammenspiel zwischen den Nutzer*innen und den stereotypisch weiblichen Charakteristika (Name, Aufgabenbereich, Konversationsstrategien, Stimme), die als eine Art weiblich gelesene Verkörperung eines Algorithmus betrachtet werden können (vgl. Phan 2017: 31). Dass die Interessen des Marktes mit diesen Verfahren durchaus vertreten werden könnten, zeigen Experimente, in denen Nutzungspräferenzen für männlich oder weiblich klingende Stimmen im Dialog mit VUIs getestet wurden; dabei wiesen die Proband*innen geschlechtsunabhängig eine Präferenz für weibliche Stimmen auf (vgl. die Übersicht von Sweeney 2021: 154). Solche Studienergebnisse, die bereits seit den 90er-Jahren bekannt sind, finden Eingang in die Design-Praktiken, tradieren sich dort und führen zu einer Art „cultural ‚common sense‘ design practice, obscuring their linkages to historically specific and socially-produced systems of oppression“ (Sweeney 2021: 155). Die Folge kann somit eine Dekontextualisierung und Depolitisierung der historisch-kulturellen Realität von Haushaltsdiensten sein (vgl. Phan 2019: 4).[112] Auf die Gefahren eines so möglicherweise verursachten gesellschaftlichen Schadens weisen Loideain/Adams (2020) hin und diskutieren potenziell anwendbare

112 Für eine historische Verlängerung der Geschichte von Dienern und Dienstboten bis ins Digitale siehe Krajewski (2010), der in seiner Kultur- bzw. Mediengeschichte des Dieners argumentiert, dass – nach dem Verschwinden menschlicher Diener und der vermehrten Übertragung von Aufgaben an Maschinen im 20. Jahrhundert – derzeit ein Transformationsprozess stattfindet, in dem „den von ‚unbeseelten‘ Dingen verrichteten Diensten derzeit wiederum ein Subjektcharakter zurückerstattet wird“ (Krajewski 2010: 19).

Regulierungsverfahren zur Begrenzung dessen durch EU-Recht und konkret durch die DSGVO (siehe auch Zarcone/Leschanowsky 2023: 172–173).

Betrachtungen aus einer solchen kritischen Perspektive zum Design von VUIs bzw. Smart Speakern bleiben nicht bei Fragen von Gender stehen, sondern nehmen auch andere Domänen wie Rasse, sozialen Status, Bildungsniveau u. a. in den Blick (vgl. Strengers/Nicholls 2018; Phan 2019: 21–23; Schiller/McMahon 2019; Natale/Cooke 2021). Außerdem ist der Phänomenbereich etwa von den Feminist Technology Studies auch auf die Marketing- und Werbekonzepte für Smart-Home-Anwendungen und Technologie-Einsatz in häuslichen Umgebungen und deren Wechselspiel mit Gender Scripts in den Blick genommen worden (vgl. Chambers 2020; Sadowski/Strengers/Kennedy 2021; Woods 2024: 102–105). Eine Inhaltsanalyse von Werbematerial und Marktberichten zu Smart-Home-Anwendungen kam zu dem Ergebnis, dass häufig Frauen nicht als kompetente Anwenderinnen dieser Geräte dargestellt bzw. inszeniert werden, aber als diejenigen, die durch ihre Integration in den Alltag entlastet werden (vgl. Chambers 2020: 314). Tatsächlich steigen vielmehr die gesellschaftlichen, kulturell geprägten Erwartungen an Care-Arbeit und das häusliche Umfeld (vgl. Woods 2024: 119).

Mit Blick auf den linguistischen Fokus der Untersuchung können nicht alle diese Aspekte detaillierter behandelt werden. Als Anschlussfrage an diesen Diskurs ergibt sich jedoch unbedingt diejenige nach der kommunikativen Konstruktion von Gender im Austausch mit den Smart Speakern. „Voice is a key marker for identity and can be inflected with assumed racial, gendered, and other meanings“ (Woods 2024: 129). Auf die Stimme soll insofern ganz besonders da eingegangen werden, wo sie sprachlich hervortritt, etwa bei der Auswahl der Klangfarbe im Zuge der Inbetriebnahme. Dabei sind sowohl die direkten Dialoge mit dem VUI wie auch die vorbereitende und anschließende zwischenmenschliche Interaktion zwischen ko-präsenten Personen relevant. Es könnten unter diesen Gesichtspunkten sprachliche Verfahren identifiziert werden, mit denen die durch das Design eingeschriebenen Gender-Stereotype bestätigt, verworfen oder anderweitig thematisiert werden. So soll der bisher skizzierte Diskurs auf einer empirischen Basis mit einer praxeologischen Perspektive angereichert werden, um so die Gender-Konstruktionen *in situ* zu beschreiben.

3.2.6 Datenschutz, Datenverwertung und Privatsphäre

Ein weiterer Diskurs dreht sich um Fragen nach dem Datenschutz, der Datenverwertung durch die Anbieter und der Privatsphäre der Nutzer*innen von Smart Speakern. Dabei sind verschiedene Aspekte zu unterscheiden: Einerseits wird die Auswertung und weitere Verarbeitung der von den Nutzer*innen bereitgestellten

Daten durch die Hersteller kritisch betrachtet und die (u. a. rechtliche) Zulässigkeit der Verwertung diskutiert (siehe etwa Turow 2021). Andererseits wird hinterfragt, inwieweit in die Privatsphäre der Nutzer*innen von Smart Speakern durch die Geräte eingegriffen wird und inwieweit diese auch nutzer*innseitig (wissentlich oder unwissentlich) zugunsten einer komfortablen Nutzung der Dienste aufgegeben wird (siehe etwa Lutz 2023). Es entsteht somit ein Spannungsfeld zwischen einem erbetenen Mithören auf der einen und einem Eingriff in die Privats- oder sogar Intimsphäre der Haushalte auf der anderen Seite (vgl. Habscheid et al. 2025b).

Zur rechtlichen Zulässigkeit der Verwertung der Daten kommt auf Basis der aktuellen Rechtslage, insbesondere im Hinblick auf die Datenschutzgrundverordnung (DSGVO), ein Gutachten der Wissenschaftlichen Dienste des Deutschen Bundestags (2019) jedenfalls für den Smart Speaker von Amazon zu dem Schluss, dass die Verarbeitung zu den angegebenen Zwecken zulässig ist. Die Gutachter*innen prüften dabei die Voraussetzung der Einwilligung in die Verarbeitung, die notwendige Bereitstellung von Pflichtinformationen und die mögliche Weitergabe an staatliche Stellen. Sie bezogen sich in ihrem Gutachten insbesondere auf die Zulässigkeit der Transkribierung der mündlichen Eingaben, nachdem diese von verschiedener Seite in Zweifel gezogen worden war (vgl. Wissenschaftliche Dienste des Deutschen Bundestages 2019: 5). Die Autor*innen stellen zwar fest: „Amazon dürfte seiner Pflicht zur Informationsvermittlung [...] in hinreichendem Maß nachgekommen sein." (Wissenschaftliche Dienste des Deutschen Bundestages 2019: 9), weisen aber auch darauf hin, dass die Daten potenziell zu sehr verschiedenen Zwecken, auch außerhalb der von Amazon selbst bereitgestellten Plattformen, genutzt werden könnten. Dies würde weitere Informationspflichten nach sich ziehen. Als problematisch im Blick auf die geltende Rechtslage wird in dem Papier hervorgehoben, dass für die Nutzer*innen erkennbar sein muss, wann überhaupt Daten zur weiteren Verarbeitung gesammelt werden. Stimmdaten gelten dabei im Sinne der DSGVO als besonders schutzwürdig und werden in der Kategorie der biometrischen Daten behandelt (vgl. Zarcone/Leschanowsky 2023: 172). Dies sei, so das Gutachten der Wissenschaftlichen Dienste weiter, insbesondere für Dritte relevant, die sich nicht im Klaren darüber sind, dass ein Smart Speaker potenziell das Gesprochene aufzeichnet; ferner sei der Minderjährigenschutz möglicherweise nicht gewährleistet. Beide Nutzer*innengruppen können nicht ohne Weiteres von der Datenerhebung ausgenommen werden (vgl. Wissenschaftliche Dienste des Deutschen Bundestages 2019: 8–9).

Zum letztgenannten Punkt äußerte sich auch die Datenethikkommission der Bundesregierung (2019) in ihrem Bericht. Dieser hebt insbesondere darauf ab, dass nicht nur intransparent ist, wie die erhobenen Daten verarbeitet werden,

sondern auch, wann Daten erhoben werden und ob die Anwesenden darum wissen. Sie empfehlen insofern „[b]indende technische Vorgaben zur Abschaltbarkeit von Mikrofon und Internetverbindung sowie eine Sichtbarmachung, ob das Mikrofon an- oder ausgeschaltet ist“ (Datenethikkommission der Bundesregierung 2019: 101). Dazu gehöre eine dem Medium und der Situation angemessene Transparenzpflicht, z. B. über ein akustisches Signal. Ferner wird eine Reduktion der Online-Verarbeitung der Daten sowie eine Regulierung dieser Vorgänge empfohlen (vgl. Datenethikkommission der Bundesregierung 2019: 101; 118).

Von verschiedener Seite wird hinterfragt, ob tatsächlich – wie es den Angaben der Hersteller entspräche – eine Speicherung und Verwertung in der Cloud nur dann erfolgt, wenn das Aktivierungswort erkannt wird. Diese Aufzeichnungen werden in der Smartphone-App als ‚Protokolldaten‘ hinterlegt, sodass sie für die Nutzer*innen über ein zweites, mit dem Smart Speaker verbundenes Gerät nachvollziehbar werden (vgl. Habscheid et al. 2021). Auf Basis einer Netzwerktraffic-Analyse zeigen etwa Ford/Palmer (2019), dass zwar die Übertragung definitiv unterbrochen wird, wenn das Mikrofon geräteseitig abgeschaltet ist. Zugleich deuten die Daten durchaus darauf hin, dass bei eingeschaltetem Mikrofon auch Daten transportiert werden, deren Übertragung an die Cloud nicht dokumentiert wird. Die beständige Internetverbindung, der für die Inanspruchnahme der Funktionalitäten benötigte „Always-On“-Status und die unklare Weiterverarbeitung der Daten in den Cloud-Servern erschweren eine Kontrolle erheblich (vgl. Gray 2016; Apthorpe et al. 2017). Da auch eine Verifizierung der tatsächlichen Aufnahme des Aktivierungsworts in den Datencentern der Hersteller stattfindet, kann nicht einmal sicher ausgeschlossen werden, dass nicht auch ein Teil der sog. Hotword-Detection ebendort stattfindet. Wird eine Löschung vorgenommen, bestreitet Amazon zwar, dass die Datensätze weiterhin verfügbar sind, gibt jedoch auch an, andere Datensätze teilweise länger zu speichern, um einen reibungslosen Ablauf der genutzten Dienste trotz der Löschung gewährleisten zu können. Dabei wird nicht genauer spezifiziert, welche Datensätze hier gemeint sind (vgl. Amazon 2021). Für Drittanbieter-Daten gibt Amazon an, dass diese nicht gelöscht werden können (vgl. Amazon 2021). Die Frage nach der Datenspeicherung ist bei anderen Herstellern ebenso unklar.

Opazität besteht insofern nicht nur beim *Ob*, sondern auch hinsichtlich der Frage, *wie* die Daten ausgewertet werden. Über die genauen Prozesse der algorithmischen Auswertung der Aufnahmen werden seitens der Hersteller keine Informationen bereitgestellt, Amazon beschreibt aber in knapper Form, dass die Daten für eine Verbesserung genutzt werden, und zwar mit Hilfe von „maschinellem Lernen, einem branchentypischem Prozess, bei dem Menschen einen sehr kleinen Anteil von Anfragen überprüfen, um Alexa zu helfen, die zutreffende In-

terpretation der Anfragen zu verstehen“ (vgl. Amazon 2021).[113] Dieser Passus scheint die Reaktion auf einen medial geführten Diskurs zur Datensicherheit zu sein, bis zu dessen Aufkommen diese Information nicht an die Besitzer*innnen der Geräte weitergegeben wurde. Mit Blick auf Daten*schutz* lässt sich somit auch die Frage stellen, wer Zugriff auf die Daten und welche Bestandteile davon hat: Die Angaben der Hersteller, dass es sich dabei um vollständig anonymisierte Stimmbeispiele handelt, können wohl nicht ganz unbegründet in Zweifel gezogen werden (vgl. Zuboff 2018: 302). Auch der Schutz vor unbefugtem Zugriff auf die Daten scheint nicht durchgängig gewährleistet zu sein und die Hersteller könnten hierbei ihre Sorgfaltspflichten vernachlässigt haben; zu diesem Schluss kommen jedenfalls Berichte von Investigativjournalist*innen und Whistleblowern (vgl. Petereit 2021a; 2021b), denen zufolge der unautorisierte Zugriff auf Sprachaufnahmen über Jahre hinweg für eine vierstellige Anzahl von nicht befugten Amazon-Mitarbeiter*innen gegeben war:. Teilweise herrsche ein „Datenchaos“, Angaben über Speicherort und -art könnten nicht rekonstruiert werden und insofern seien auch die Grundsätze der DSGVO nicht einzuhalten.

Mit einer größeren gesellschaftlichen Perspektive scheint es jedoch zu einfach zu sein, die großen Tech-Firmen ‚lediglich‘ für Missstände im Hinblick auf den Datenschutz und die Intransparenz zu kritisieren. Der Umgang mit *Big Data* und Sprachaufnahmen als Bestandteil davon kann auch als gesamtgesellschaftliche Herausforderung betrachtet werden, bei der sich wirtschaftlich betrachtet die klassischen Rollen von Produzent*in und Konsument*in verschieben und die bisherigen Marktmechanismen und Verfahren zu dessen Regulation versagen. In diesem Prozess sind Amazon, Google und Apple mehr als Dienstleister, die sich nicht an rechtliche Vorgaben halten: Sie sind, um es mit der Wirtschaftswissenschaftlerin Shoshana Zuboff zu sagen, Treiber des „Überwachungskapitalismus“. Mit diesem totalitaristischen Wirtschaftskonzept zeichnet Zuboff (2018) die Folgen der Digitalisierung und der damit einhergehenden Verdatung des privaten Alltags als eine Art real werdende Dystopie. Menschliches Verhalten werde zu „Verhaltensdaten“ (siehe hierfür und für die folgenden Begriffe Zuboff 2018: 22), die nicht alle zur Verbesserung des maschinellen Lernens und somit der angebotenen Dienstleistungen dienen, sondern die als sog. „Verhaltensüberschuss“ in die Entwicklung von „Vorhersageprodukten“ einfließen, die das menschliche Verhalten prädiktiv „erahnen“. Diese Vorhersageprodukte können in einem von Zuboff als „Verhaltensterminkontraktmarkt“ bezeichneten Handelsverkehr verkauft werden

113 Beim Hersteller Google handelt es sich dabei eigenen Angaben zufolge um 0,2 Prozent, bei Amazon und Apple weniger als ein Prozent (siehe ZEIT Online 2019); diese Angaben dürften allerdings angesichts der nicht unerheblichen Bedeutung menschlicher Beteiligung am Prozess des maschinellen Lernens zweifelhaft sein (siehe auch Kremp 2019).

und werden wiederum dadurch besonders wertvoll, dass sie treffsicher sind. Die Erhöhung der Treffsicherheit gebe letztlich Anreiz zur *Beeinflussung* des menschlichen Verhaltens: „Ergebnis dieses Wandels ist, dass automatisierte Maschinenprozesse unser Verhalten nicht nur *kennen*, sondern auch in einer wirtschaftlichen Größenordnung *auszuformen* vermögen [Herv. i. O.]“ (Zuboff 2018: 33). Die Art der Verhaltensdaten und des daraus generierten Datenüberschusses können vielfältig sein (etwa Such- und Einkaufsverhalten im Internet, Kommunikationsverhalten, Nutzungszeiten und -orte usw.). Darunter fallen auch sprachliche Daten bzw. Konversationen, die die Smart Speaker mitschneiden und zu Sprachüberschuss werden. Dieser Sprachüberschuss und der Wettbewerb, in dem die Anbieter aktuell stehen, führen dazu – so die Argumentation Zuboffs (2018: 307–309) –, dass die Art und Qualität der angebotenen Dienstleistungen hinter die „Jagd nach Verhaltensüberschuss“ zurücktritt und primär Potenziale erschlossen werden, um noch weiteren Sprachüberschuss generieren zu können.

Dieser Argumentation und dem Konzept des Überwachungskapitalismus muss man nicht folgen; eine ausführliche Diskussion der Szenarien von Zuboff ist an dieser Stelle untunlich (siehe aber Morozov 2019). Unbestreitbar ist allerdings, dass die gesellschaftlichen und ökonomischen Verhältnisse durch die Dominanz der hier untersuchten Anbieter und auch vermittels der hier untersuchten stationären Sprachassistenzsysteme in Bewegung geraten sind (siehe etwa Sadowski 2020). Wie Woods (2018) herleitet, begünstigen sich die im Kapitel zuvor ausgeführten Gender-Stereotype und das Anliegen der Datengewinnung seitens der Großkonzerne gegenseitig. Die Dominanz der Datafizierung sozialer Praktiken und die Anwendung von Künstlicher Intelligenz in allen Lebensbereichen prägen und normieren wirtschaftliche und gesellschaftliche Strukturen des 21. Jahrhunderts. Die Konzentration dieser Prozesse auf kommerziell betriebene Plattformen wie Amazon, die zu Infrastrukturen des täglichen Lebens werden (Dolata/Schrape 2018; Dolata 2019; Goulden 2019; Plantin/Punathambekar 2019; Strüver 2023a), führt zu Veränderungen in der Wirtschaftsordnung (vgl. Srnicek 2016; Montalban/Frigant/Jullien 2019). Diese Entwicklungen werden unterschiedlich bewertet: Während etwa Hepp (2020) die hier angesprochenen Prozesse mit dem Konzept der Mediatisierung in einer durchaus als optimistisch zu bezeichnenden Diskussion von Vor- und Nachteilen bespricht, fürchten andere Autor*innen die Manifestierung von gesellschaftlichen Ungleichheiten, die Verschlechterung von Arbeits- und Lebensverhältnissen und eine erodierende Staatlichkeit angesichts der Macht durch die gesammelten und für verschiedene Zwecke auswertbaren Daten (siehe etwa Crawford/Joler 2018; Couldry/Mejias 2019).

Vor diesem Hintergrund ist auch der Schutz der Privatsphäre nicht mehr nur eine Frage danach, ob sich die Tech-Firmen an geltendes Recht halten, sondern auch, welche Implikationen es für die Nutzer*innen hat, wenn Smart Speaker Au-

dioaufzeichnungen intimer Details aus dem Privatleben an einen proprietären und opak agierenden Cloud-Dienst weiterleiten (vgl. Waldecker/Volmar 2022). Konkret für Smart Speaker wurden bereits Resignation (vgl. Lau/Zimmerman/Schaub 2018: 18), Überforderung sowie die Entwicklung eines „pragmatischen Fatalismus“ (Waldecker/Martin/Hoffmann 2025) oder Zynismus (Lutz/Newlands 2021) seitens der Nutzer*innen beobachtet, der auch in anderen digitalen Zusammenhängen an der Schnittstelle von Intimität, Privatheit und Öffentlichkeit festgestellt werden konnte (siehe etwa Schmidtke/Englert/Waldecker 2019). Dies zeigt an, dass ein Großteil der Gesellschaftsmitglieder den Entwicklungen möglicherweise kritisch, aber nicht aktiv widerständig gegenübersteht. Nutzer*innen, die generell größere Bedenken hinsichtlich der Privatsphäre haben, neigen auch zu größeren Bedenken im Hinblick auf die Überwachung durch Smart Speaker und Sicherheitsschwächen von Smart Speakern, z. B. Datenmissbrauch (vgl. Mols 2021: 163–165). Nutzer*innen, die mehr über die Geräte wussten und im Umgang mit ihnen geübter waren, waren im Hinblick auf Sicherheitsaspekte und die Datenverwertung innerhalb der Plattformlogiken ebenfalls besorgter (vgl. Mols 2021: 163–165). Im Einklang damit halten Lutz/Newlands (2021: 154) auf Grundlage ihrer Befragung britischer Smart-Speaker-Anwender*innen fest, dass diese die größte Sorge im Hinblick auf eine Hörbarkeit der Daten für Dritte haben (z. B. Vertragsarbeiter*innen der Systemanbieter). Zwar scheinen Nutzer*innen als Reaktion darauf unterschiedliche Strategien für „Privacy Work“ zu entwickeln (vgl. Brause/Blank 2023; siehe auch Chalhoub/Flechais 2020), darunter auch die Unterbrechung laufender Gespräche bei einer versehentlichen Aktivierung des Smart Speaker, was wiederum ein Hinweis auf die Relevanz einer empirischen Untersuchung sprachlicher Praktiken in VUI-Dialogen ist. Zugleich zeigen Lutz/Newlands (2021: 155), dass die von ihnen befragten Nutzer*innen einen hohen Grad an Pragmatismus bzw. Zynismus erkennen lassen und an einer intensiveren Auseinandersetzung mit Datenschutzeinstellungen nicht interessiert sind bzw. keinen Aufwand dafür betreiben wollen. Waldecker (2022: 156) konstatiert in einer kritischen Diskussion des Konzepts der *Datensouveränität*: „De facto haben sich die Nutzenden von Smart Speakern in gewissem Rahmen mit der Auswertung ihrer Daten abgefunden“,[114] und nimmt an, dass die Nutzer*innen

114 Die Studie von Waldecker (2022) fasst Teilergebnisse des Projekts „Un/erbetene Beobachtung in Interaktion: ‚Intelligente Persönliche Assistenten‘ (IPA)“ im Sonderforschungsbereich „Medien der Kooperation“ zusammen. In diesem Projekt ist auch die vorliegende Arbeit entstanden. Die genannte Untersuchung stützt sich auf Interviews in denselben Haushalten, in denen auch für die vorliegende Arbeit gemeinsame Datenerhebungen und interdisziplinäre -auswertungen durchgeführt wurden.

eine umfängliche Souveränität weder erlangen können, noch danach zu streben scheinen.

Notwendig für eine weitere Exploration der hier aufgezeigten brisanten Spannungsfelder ist eine empirische Grundlage, die die *Praxis* ins Zentrum der Untersuchungen rückt. Es muss betrachtet werden, wie *in situ* Smart Speaker eingesetzt werden, wie dabei Datenschutz und -verarbeitung sowie Kategorien von Intimsphäre, Privatheit und Öffentlichkeit sprachlich-kommunikativ reflektiert werden und welche Veränderungen im Alltag der Nutzer*innen sich beobachten lassen. Die linguistische Betrachtungsweise, in der der soziale Status der Smart Speaker in den Blick rückt, trägt somit dazu bei, die soziale Wirklichkeit im Umgang mit den skizzierten gesellschaftlichen Diskursen analysierbar zu machen. Dabei wird über die sprachlichen ein Zugriff auf soziale Praktiken ermöglicht und es können die Annahmen der Anwender*innen über die Funktionsweise der Geräte, ihre Einstellungen zu Datenschutz, -verwertung und Privatsphäre so beobachtet werden, wie sie selbst von den Sprecher*innen relevant gesetzt werden. Die mit dieser Brille opak bleibenden Datenpraktiken im „Back-End“ von Smart Speakern, die Auswertung und mittel- bis langfristigen Folgen für die Weiterentwicklung von Algorithmen werden dabei als Domäne betrachtet, zu deren weiterer Untersuchung die Analysen des „Front-Ends“ einen Beitrag leisten können.

3.3 Untersuchte Smart Speaker-Modelle

Die drei am Markt gängigsten Smart Speaker – Amazon Echo, Google Home und Apple HomePod – sollen nachfolgend kurz vorgestellt werden. Dabei soll die äußere Gestaltung, die Bedienweise und Funktionalität der Geräte im Mittelpunkt stehen. Zu dieser gehören notwendigerweise auch stimmliche Eigenschaften des Voice User Interfaces (VUI). Ferner werden die zugehörigen Smartphone-Anwendungen beschrieben. Amazon Echo ist mit weitem Abstand der bekannteste Smart Speaker: 2024 hatten 72 Prozent der Haushalte, die einen Smart Speaker besitzen, ein solches Modell (hierfür und für die folgenden Angaben vgl. Statista 2025a). 22 Prozent der Nutzer*innen verwendeten ein Produkt von Google (Google Home und Google Nest mit jeweils 15 bzw. 8 Prozent). 15 Prozent verwendeten den HomePod von Apple. Die Beschreibung der Smart Speaker-Modelle geht zunächst vom Amazon Echo aus. Anschließend werden Google Home und Apple HomePod v. a. in ihrer Differenz zum Amazon-Produkt vorgestellt.

3.3.1 Amazon Echo

Der Amazon Echo ist in verschiedenen Produktvarianten erschienen. Seit dem Marktstart waren bis zum Zeitpunkt der Datenerhebung vier Generationen mit jeweils zwei verschiedenen Ausführungen erhältlich: Amazon Echo sowie eine weitere Ausführung als „Echo Dot“, die stets mit etwas Verzögerung zur Veröffentlichung der neuen Generation als alternative Produktvariante erschien und insgesamt etwas kleiner und im Hinblick auf den Lautsprecher etwas weniger leistungsstark ist, in der Funktionalität jedoch keine Einschränkungen hat. Die Abb. 1–3 zeigen die ersten drei Generationen des Amazon Echo, wobei die erste Generation (Abb. 1) in der größeren Echo-Variante dargestellt ist, während Generation 2 und 3 in der Echo Dot-Produktvariante abgebildet sind:

Abb. 1: Amazon Echo, 1. Generation; Bild: „Frmorrison“ / CC BY-SA 3.0.

Abb. 2: Amazon Echo Dot, 2. Generation; Bild: Raimond Spekking / CC BY-SA 4.0.

Abb. 3: Amazon Echo Dot 3. Generation mit aktiviertem Lichtring; Bild: „hamburgfinn“ / Pixabay.

Hardwareseitig besteht das System v. a. aus drei Komponenten: mehreren miteinander zusammenwirkenden Mikrofonen (sog. *Far-Field-Voice-Recognition*, siehe Drösser 2020: 71), einem Lautsprecher sowie einer kleinen Recheneinheit, die die

Übertragung ins Internet leistet und einige wenige lokale Prozesse steuert. Über die eingebauten LED-Leuchten sendet das Gerät in unterschiedlichen Farben nichtsprachliche Signale (vgl. Pearl 2016: 219–221; Bedford-Strohm 2017: 487). Die Abb. 4–6 zeigen die LED- und Mikrofon-Einheit eines Echo Dot sowie die zugehörigen Recheneinheiten und den Lautsprecher.

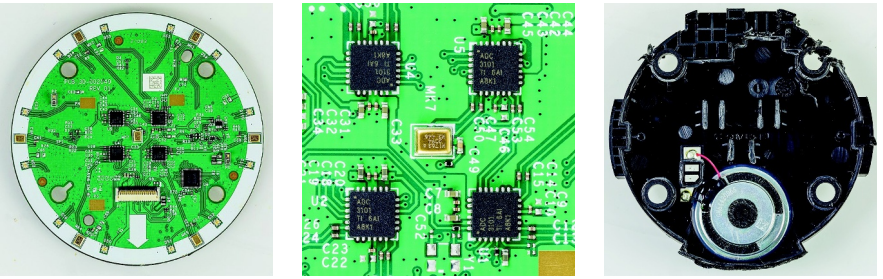

Abb. 4–6: Sechs Lautsprecher und LED-Einheiten, Recheneinheiten und die Unterseite eines Echo Dot mit Lautsprecher-Einheit. Bilder: Raimond Spekking / CC BY-SA 4.0.

Diese einzelnen Komponenten, aus denen sich der Amazon Echo zusammensetzt, werden von verschiedenen Softwareanwendungen in Anspruch genommen. Diese ergeben erst im Zusammenspiel eine Einheit und sind im Hintergrund als unterschiedliche Prozesse zu begreifen, die auch verschiedene Teildisziplinen der Informatik und der Computerlinguistik berühren. Auf Abb. 3 ist der aktivierte Lichtring zu erkennen, der rund um die obere bzw. untere Kante des mit Stoff verkleideten, zylinder- bzw. kugelförmigen Korpus des Geräts gespannt ist. Dieser signalisiert in blauer Farbe, dass die Aufzeichnung aktiv ist, die aufgenommenen Audiodateien also zur Auswertung in die Cloud übertragen werden (Listening-Modus). Ist der Lichtring ausgeschaltet (und passt sich dadurch optisch der Farbe des Smart Speakers an), findet zwar bei hergestellter Internetverbindung und Stromversorgung die Suche nach dem Aktivierungswort statt, aber es erfolgt keine regelmäßige Übertragung in die Cloud.[115] Der Lichtring kann bei der 3. Generation des Geräts noch andere Farben annehmen (siehe dazu Albrecht 2020: 35): Leuchtet der Lichtring etwa orange, wird eine Internetverbindung gesucht, ein „pulsierender gelber Lichtring“ steht für bereitstehende Benachrichtigungen. Ein roter Lichtring zeigt an, dass das Mikrofon abgestellt und somit die Suche nach dem Aktivierungswort unterbrochen wurde. Dies kann über eine der vier

115 Siehe Kap. 3.2.6 für weitere Auswertungen zu diesem Sachverhalt.

Tasten auf der Oberseite des Geräts erfolgen (siehe Abb. 2 und 3). Zwei weitere Tasten können zur Lautstärkeregulierung (mit Plus- und Minus-Symbol) genutzt werden. Eine vierte „Aktionstaste“ dient der Aktivierung des Geräts ohne verbale Nennung des Aktivierungsworts (vgl. Albrecht 2020: 35–36).

Die Positionierung der Tasten und des Lichtrings variieren etwas in der kugelförmigen 4. Generation des Amazon Echo (siehe Abb. 7). Außerdem ist das Gerät mit einem anderen Lautsprecher ausgestattet und hat einen geringfügig anderen Klang, der sich jedoch Reviewer*innen wie Kawalkowski (2020) zufolge kaum vom Vorläufermodell unterscheidet. Die Recheneinheit wurde geringfügig beschleunigt, aber die Funktionen sind sonst nahezu identisch, was sich durch die Verarbeitung der Daten in der Cloud erklärt: Die Hardware ist für das Leistungsspektrum nur noch ein nachgelagerter Bestandteil (siehe auch Stiftung Warentest 2021: 28; Freeman-Mills 2021).

Abb. 7: Amazon Echo Dot, 4. Generation; Bild: Public Domain / Unsplash, Bearbeitung: Sina van Oostrum.

Alle Amazon Echo-Geräte sind mit dem VUI Alexa ausgestattet, das die akustische Steuerung des Geräts über Stimmein- und ausgaben ermöglicht. Die Stimmfarbe des Geräts lässt sich nicht in den werkseitig verbauten Funktionen anpassen, das Aktivierungswort hingegen kann geändert werden (möglich sind auch „Amazon“, „Computer“ oder „Echo“). Auch wenn Alexa im alltäglichen Sprachgebrauch häufig synonym mit dem Smart Speaker verwendet wird (vgl. Albrecht 2020: 7), ist der Sprachassistent auch auf anderen Endgeräten verwendbar, z. B. auf Smartphones, Smart Speakern von anderen Produzenten oder Tablets (vgl. Albrecht 2020: 11). Außerdem lassen sich Drittanbieter-Programme, *Skills*, auf den Geräten installieren, die dann teilweise auch andere VUIs einprogrammiert haben, sodass sich die Stimmfarbe innerhalb der Programme verändern lässt. Die synonyme Verwendung von Alexa als Bezeichnung für den Smart Speaker ist also in doppel-

ter Hinsicht nicht akkurat: Einerseits kann die Software Alexa auf anderen Geräten verwendet werden, andererseits können auf dem Smart Speaker andere Anwendungen genutzt werden.[116]

Weitere Produktvarianten der Amazon Echo-Reihe sind der Amazon Echo Show und der Echo Spot, die jeweils einen Monitor verbaut haben, auf dem Uhrzeit, Wetter, Text- und Videomaterial und andere Anwendungen in Kombination mit der stimmlichen Schnittstelle auch visuell präsentiert werden (hierfür und für das Folgende siehe Albrecht 2020: 18–19). Ermöglicht werden auch Videotelefonate. Geräte mit Bildschirm sind seit 2017 auf dem US-Markt erhältlich und machten global gesehen 21 Prozent der verkauften Amazon-Echo-Geräte aus (vgl. Canalys 2019). Die Geräte werden in die weiteren Ausführungen und in die Analyse nicht einbezogen. Es ist davon auszugehen, dass die Möglichkeit der Wahrnehmung visueller Inhalte, der Steuerung über den Bildschirm (d. h. nicht über den Smartphone-Bildschirm, der ein externes Gerät erforderlich macht) und das gänzlich anders gestaltete Zusammenspiel optischer und akustischer Signale eine Vergleichbarkeit mit Voice-First-Devices (Bedford-Strohm 2017) sehr erschwert. Sicherlich wäre ein Vergleich, der in zukünftigen Arbeiten angestellt werden könnte, aufschlussreich im Hinblick auf die Spezifika, die die stimmliche Benutzerschnittstelle mit sich bringt; dieser soll jedoch im Rahmen dieser Arbeit nicht geleistet werden (siehe aber Hector et al. 2025).

Bei den hier im Fokus stehenden Geräten erfolgt wie erwähnt die Einrichtung und ein Großteil der Einstellungen über die Alexa-Smartphone-Applikation, die ebenfalls als Alexa bezeichnet wird und über die gängigen App-Stores auf Android- und Apple-Geräten (Smartphones und Tablets) installiert werden kann. Über die App kann das Gerät bedient und, selbst ohne einen Smart Speaker, der volle Funktionsumfang in Anspruch genommen werden – Stimmeingaben können auch über das im Endgerät verbaute Mikrofon erteilt werden. Die Anwender*innen konzentrieren sich, gerade in den Einrichtungssituationen, immer wieder auf die App, außerdem verweist das VUI bei bestimmten verbal geäußerten Stimmeingaben (etwa zu Smart Speaker-Einstellungen) selbst verbal auf die Anwendung, wenn die Anfrage nicht über das VUI verarbeitet werden kann. Die App ist, abhängig von der verwendeten Version und dem Endgerät, jeweils geringfügig unterschiedlich aufgebaut (insbesondere zwischen Android- und Apple-Geräten treten Unter-

116 Bis zur 3. Generation war als weitere Produktvariante der Echo Plus erhältlich. Er ähnelte den Geräten der Echo-Reihe, hatte also keinen Bildschirm verbaut, war dafür aber hardwareseitig für die Smart Home-Steuerung optimiert und mit einem Temperatursensor ausgestattet (vgl. Albrecht 2020: 17). In der 4. Generation ist die „Plus"-Linie eingestellt, dafür werden mittlerweile Echo Show Geräte mit integriertem Bildschirm vertrieben.

schiede auf). Sie umfasst aber im Wesentlichen die gleichen Funktionen, die exemplarisch in den Abb. 8–10 in der Android-Version von 2021[117] dargestellt sind (für eine detaillierte Funktionsbeschreibung der App siehe Albrecht 2020: 36–49; Habscheid et al. 2021).

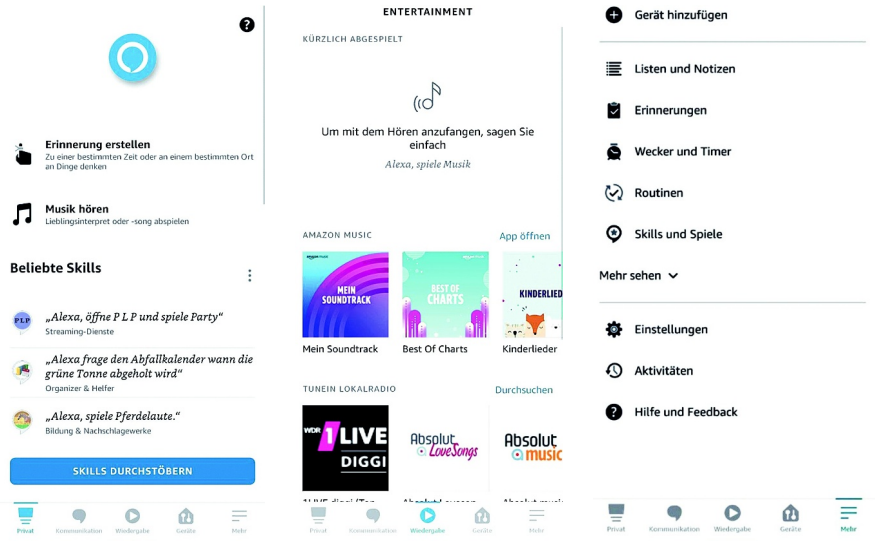

Abb. 8–10: Alexa-Smartphone-App, Funktionsbereiche „Privat", „Wiedergabe" und „Einstellungen"; Bilder: eigene Aufnahme / T.H.

Im Bereich „Privat" (in anderen App-Versionen auch „Startseite") findet sich die Möglichkeit zur Sprachsteuerung über die App im oberen Bereich, der blaue Button mit dem Alexa-Logo in Form einer angedeuteten Sprechblase aktiviert die Sprachsteuerung (Abb. 8). Darunter finden sich personalisierte Vorschläge für weitere Anwendungen, Skills oder neue Funktionsweisen. Über die Toolbar im ganz unteren Teil der App können nun andere Funktionsbereiche der App angewählt werden. Im Menüpunkt „Kommunikation" (hier nicht abgebildet) finden sich Steuerungselemente für Anrufe und Textnachrichten, die über die Amazon-Geräte vermittelt werden. Diese werden hier protokolliert und der Zugriff auf die

117 Dieser Zeitpunkt wurde gewählt, damit die Darstellung hier möglichst ähnlich zur Darstellung auf den Endgeräten der im Rahmen der Studie aufgezeichneten Nutzer*innen ist (die Aufzeichnungen erfolgten zwischen Ende 2020 und Mitte 2022, siehe Kap. 5).

jeweiligen Kontakte hergestellt. Im Bereich „Wiedergabe“ (Abb. 9) werden Musik, Hörbücher, Radio und ähnliche akustische Medieninhalte gesteuert. Playlisten und Radiosender stehen voreingestellt zur Verfügung, können aber auch manuell angelegt werden. Die Verwaltung von Streaming-Diensten wie Spotify oder Deezer kann über dieses Menü erfolgen. Im Bereich „Geräte“ (hier nicht abgebildet) erfolgt die Steuerung der Smart Speaker, die der App hinzugefügt wurden. Außerdem können hier auch Smart Home-Elemente verwaltet werden. Der Funktionsbereich „Mehr“ (Abb. 10), der bei einigen anderen App-Versionen oben links angeordnet ist, enthält weitere Funktionen zur Steuerung des Smart Speakers über die App, z. B. Erinnerungen, Wecker, das Anlegen von Routinen (die vordefinierte Abfolge von bestimmten Funktionen, ggf. zu bestimmten Zeiten) und die Verwaltung weiterer Skills. Im Bereich „Aktivitätsverlauf“ bzw. im Bereich „Datenschutz“ sind als Protokolldaten die ausgeführten Aktionen der Nutzer*innen hinterlegt (siehe Abb. 11).

Abb. 11: Alexa-Smartphone-App, Funktionsbereich „Aktivitätsverlauf“; Bild: eigene Aufnahme / T.H.

Bei mündlicher Steuerung und entsprechender Einstellung werden auch die aufgezeichneten Audio-Daten hier abgespeichert (vgl. Habscheid et al. 2021: 19) und können über den blauen Wiedergabebutton angehört werden, außerdem wird der verstandene Wortlaut transkribiert und die Stimmausgabe (ausschließlich schriftlich) dokumentiert. Die App ermöglicht also die Nutzung wesentlicher Funktionen auch über die App (teilweise mit integrierter stimmlicher Bedienung). Das Funktionsspek-

trum geht über das des VUIs hinaus: Für bestimmte Einstellungen und Anwendungen, gerade, wenn sie eher prinzipiellen Charakter haben, kann nur die App verwendet werden. So können z. B. die Adresse des Standorts, die WLAN-Verbindung, Zahldaten für Online-Shopping und andere Daten, die für das Prozessieren bestimmter Anwendungen erforderlich sind, nicht über das VUI, sondern nur über die App geändert werden. Wie die Anwendung der App in die stimmliche Nutzung eingebunden ist, soll in den Analysen mitbetrachtet werden (insbesondere in den Inbetriebnahme-Situationen), wobei trotz der hohen Relevanz der App weiterhin ein Fokus auf der Anwendung der stimmlichen Benutzerschnittstelle liegen soll.

3.3.2 Google Home und Google Nest

Die Smart Speaker von Google sind im Wesentlichen sehr ähnlich aufgebaut wie die Produktreihe von Amazon. Firmierte die Serie zunächst unter dem Markennamen „Google Home“ wurde sie 2019 in „Google Nest“ umbenannt, ohne dass daraus wesentliche Änderungen am Funktionsumfang erfolgt wären.

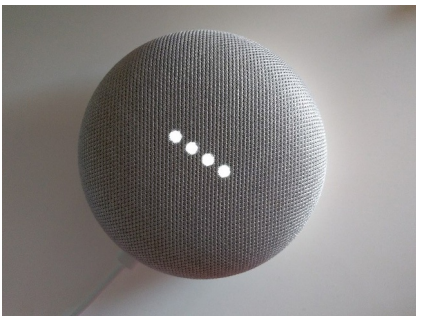

Abb. 12: Google Nest-Reihe: Aktivierter Google Home Mini; Bild: Andrea Marchitelli, Wikimedia Commons / CC BY-SA 4.0.

Abb. 13: Google Nest-Reihe: Google Home Smart Speaker, Google Home Hub und Google Home Mini; Bild: Y2kcrazyjoker4, Wikimedia Commons / CC BY-SA 4.0.

Die drei Geräte (siehe Abb. 13) – Google Home bzw. Google Nest Audio, Google Home Hub bzw. Google Nest Hub und Google Home Mini bzw. Google Nest Mini – sind jeweils die erste bzw. zweite Generation, wobei sich im Leistungsspektrum nur wenig verändert hat: Ähnlich wie beim Amazon Echo findet die Hauptanpassung der Funktionen, vermittelt über eine App, in der Cloud statt. Die Google-Reihe zeigt, anders als die Echo-Reihe, die Aktivität nicht durch einen Lichtring, sondern durch eine Leiste von Lichtpunkten auf dem Display an (siehe Abb. 12). Diese können weiß und orange leuchten, nacheinander aufblinken und pulsieren, wobei Anzahl und Bewegung der Punkte codiert sind – so bedeuten etwa vier orange Punkte, dass das Mikrofon ausgeschaltet ist, während nacheinander aufleuchtende orange Punkte einen Software-Prozess (z. B. ein Update) anzeigen (vgl. Google Support 2021). Auch während der Ersteinrichtung werden über die Punkte Statusmeldungen des Geräts übermittelt. Ferner ist, in Abgrenzung zu Amazon Echo, die Oberfläche des Geräts teilweise über Touch-Technologie funktionalisiert. So kann durch Berührungen an den Rändern des Oberflächengewebes die Lautstärke eingestellt werden. Auch hierbei geben die weißen Punkte Hinweise auf die Höhe der eingestellten Wiedergabelautstärke. Die „Google Home"-App (siehe Abb. 14 für die Startseite) umfasst ebenfalls die Steuerung der wesentlichen Funktionen des Smart Speakers, wobei auffällig ist, dass dieser stärker an das Smartphone und den Google-Account der Nutzer*innen angebunden ist. So verweisen mehrere Links innerhalb der App auf das Benutzer*innen-Konto bei Google, in dem Einstellungen zum „Google Assistant" vorgenommen werden können, das sowohl über die Smart Speaker als auch über das Smartphone oder andere Endgeräte verwendet wird. Dies wird durch das Profilbild des*der Google-Nutzer*in oben rechts deutlich, das zum Aufrufen wesentlicher Einstellungen (z. B. Stimmprofil des VUI[118] oder Eingabe persönlicher Präferenzen bei der Internetsuche) angetippt werden muss.

Außerdem zeigt sich, in Abgrenzung zur Alexa-App, eine stärkere Fokussierung auf einen Haushalt, dem verschiedene Geräte hinzugefügt werden können. Die in Abb. 14 prominent platzierte Überschrift „Neustadt" ist der Name des Haushalts. Dieser Haushalt kann teilweise losgelöst vom Google-Konto verwaltet werden (über die Schaltfläche „Einstellungen", die in der Mitte der Startseite platziert ist, ist er jedoch letztlich ebenfalls immer wieder damit verknüpft). Im Vergleich mit der „Alexa"-App besteht bei dieser zwar auch eine Verknüpfung mit einem Amazon-Konto und die Repräsentation eines ‚Haushalts' mit mehreren Geräten. Gleichwohl ist aber eine deutlich we-

118 Anders als bei Amazon-Produkten kann beim Stimmprofil für verbale Ausgaben des Smart Speakers von Google zwischen zwei verschiedenen Profilen ausgewählt werden.

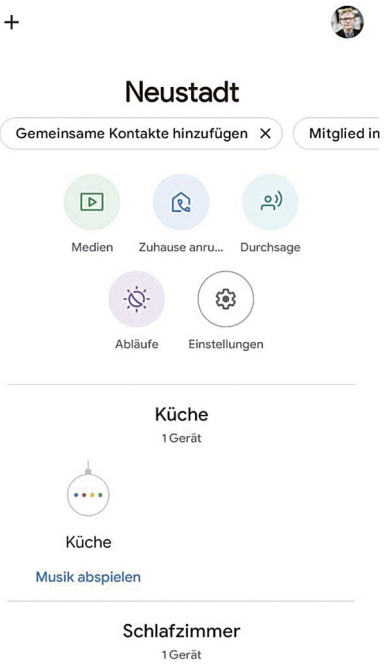

Abb. 14: Startseite der Google-Home-App; Bild: Screenshot / T.H.

niger starke Anbindung an das Amazon-Konto innerhalb der Anwendung sichtbar, im Vordergrund stehen die Funktionen des Smart Speakers. Dies ist auch dadurch erklärbar, dass die Amazon-Reihe nicht mit mehreren strukturell unterschiedlichen Instanzen umgehen muss: Während das VUI von Google sowohl auf dem Smartphone wie auch auf dem Smart Speaker ansprechbar ist und hier in einem integrierten Ansatz ermöglichen will, die gleichen Präferenzen beizubehalten, fokussiert sich Amazon auf die einheitliche Steuerung des Smart Speakers selbst bzw. der App-Instanz des Smart Speakers. Als letzte Unterscheidung zum Amazon-Produkt lässt sich bei Google ferner eine weniger starke Personalisierung feststellen. Das Aktivierungswort von Google Home- bzw. Google Nest-Produkten lautet „Okay Google“ und „Hey Google“ (ist also nicht mit einem onymischen Ausdruck personalisiert) und es zeigt sich auch in den Werbestrategien der Hersteller weniger stark der Aufbau einer „Persona“ (vgl. Dickel/Schmidt-Jüngst 2021).[119]

119 Siehe auch Kap. 6.1.1.

3.3.3 Apple HomePod

Der HomePod von Apple ist seit 2017 auf dem Markt. 2020 erweiterte Apple – analog zu den Serien von Amazon und Google – das Produktangebot um einen HomePod Mini (siehe Abb. 15). Der HomePod ist demgegenüber 2021 eingestellt worden. Das Produkt von Apple ist, ähnlich wie die Google-Reihe, mit dem auch auf dem Smartphone oder Tablet verfügbaren Apple-VUI („Siri“) ausgestattet. Über dieses VUI findet, wie auch bei Google, eine Verknüpfung der Einstellungen von Smartphone und Smart Speaker statt. Die Steuerung erfolgt ebenfalls über eine App, die „Apple Home“-App, die auf dem Smartphone oder Tablet installiert ist. Über diese können auch Smart Home-Elemente gesteuert werden. Die Verknüpfung erfolgt hier mit der Apple-ID, dem Pendant zum Google- oder Amazon-Account.

Abb. 15: Apple HomePod Mini; Bild: Arne Müseler / CC BY-SA 3.0 de, Bearbeitung: Sina van Oostrum.

Der Smart Speaker von Apple ist Stiftung Warentest (2021: 28) zufolge mit einem leistungsfähigeren Lautsprecher ausgestattet als die Produkte der anderen Hersteller, während das VUI (Reaktion auf das Aktivierungswort, „Sprachmelodie, Betonung und Phrasierung“) schlechter als das von Amazon bewertet wurde. Der HomePod lässt keine Drittanbieter-Skills zu. Verknüpfen lässt er sich lediglich mit dem Musik-Streamingdienst von Apple selbst und darüber hinaus nur mit Apple-Anwendungen. Dabei ist die Funktionalität der Anwendungen im Vergleich zur Bedienung über das Smartphone – ähnlich wie bei der Nutzung von Google-Anwendungen über den Smart Speaker – teilweise eingeschränkt.

Wichtig ist zu betonen, dass die technologischen Entwicklungen im Bereich der Smart Speaker sehr schnell voranschreiten. Während der ca. sechs Jahre, die die Geräte jetzt auf dem deutschen Markt erhältlich sind, wurden mehrere Generationen, Produktvarianten und Hardware- wie Software-Updates entwickelt. Zum Zeitpunkt der Veröffentlichung dieser Arbeit werden die gängigen Produkte möglicherweise

bereits anders aussehen, neue Funktionalitäten ermöglichen, während andere wegfallen, neue Marken ausbilden und andere einstellen. Durch die Volatilität dieses Sektors und die noch fehlende Etablierung einheitlicher Standards ist es unmöglich, hier den Anspruch einer vollständig aktuellen Beschreibung des Smart Speaker-Felds zu erfüllen. Insofern hat sich die Beschreibung auf die Geräte konzentriert, die zum Zeitpunkt der Erhebung marktgängig und v. a. in den untersuchten Haushalten verfügbar waren (auf diese wurde auch bei den Abbildungen ein Schwerpunkt gelegt). Neue Versionen und Software- wie Hardware-Updates größerer Art, die nach der Datenerhebung erfolgten, werden hier nicht besprochen, da sie für die Nutzer*innen nicht relevant waren und während der Analyse keinen Einfluss mehr auf die erhobenen Daten hatten.[120]

120 Siehe aber Kap. 7.3.