Innovating Alexa amid the Rise of Large Language Models

Sociotechnical Transitions in Algorithmic Development Practices

Niklas Strüver

Abstract *For about a decade, Amazon's Alexa was a pioneer in automatic speech processing; now, however, new Large Language Models (LLMs) are posing challenges for Amazon. One attempt to confront these challenges is by integrating technologies developed for Alexa by university research teams in the Alexa Prize Competitions (APCs). This chapter examines how participants in these contests deal with the conditions set and the resources provided by Amazon for the competition, and offers a snapshot of the practical development processes of the voice assistant at a time of technological transition. It then outlines some of the path dependencies, risks, benefits, and aspects of structuration that are encountered by the participants in their attempts to innovate Alexa.*

1. Introduction

Over the course of the last decade, Amazon has spent a considerable amount of effort making Alexa reliable enough to be desirable for many households[1]. In the last couple of years however, Amazon had been reducing its generosity to its Alexa division (Kim 2022) – that is, until the competing machine learning company OpenAI introduced large language models (LLMs) to the public, most no-

1 Technically speaking, Alexa is the voice interface for Amazon's cloud products Alexa Voice Service (AVS) and Amazon Web Services (AWS), where all requests are processed by various machine learning algorithms (Crawford and Joler 2018), which are constantly optimized based on the incoming usage data. This service is embedded in the Echo devices produced by Amazon.

toriously with their use cases in the form of ChatGPT in November 2022. As the world was familiarizing itself with a proclaimed revolution of artificial intelligence (AI) technologies, technology companies like Amazon found themselves with an apparent need to catch up. Upgrades announced for Alexa (Bensinger 2024; Jassy 2024; Krishnan 2024) indicate that Amazon is working on ways of integrating LLMs into its voice assistant, which until now had primarily relied on more traditional machine learning approaches. This change in coding approaches for Alexa comes with a set of difficulties that need to be navigated in a competitive field of technology development (Kinsella 2023).

To better understand the transition between two different approaches to making Alexa talk to users, and to gain insights into Amazon's development practices for Alexa, a qualitative expert interview study was conducted to investigate how development is practiced in the Alexa ecosystem. As it is difficult to conduct research within Amazon directly, the Alexa Prize Competitions (APCs), in which university research teams compete to build technologies for Alexa, were chosen as a proxy study context that could offer insights into the technological development of Alexa, as well as into Amazon's approach to cooperating with third parties (universities in this case) that wish to interact with Alexa as a platform. By exploring the views of third-party actors who obtain access to Alexa technologies and are closely supervised by Amazon Alexa staff, the study seeks to contribute to research on the sociotechnical analysis of Amazon's technology for Alexa; ultimately to further understanding of the sociotechnical underpinnings of a technology that is present in many homes globally. To achieve these aims, the questionnaire used in the study was developed to elicit details about the inner workings of cooperation with Amazon, making the APC teams a proxy of analysis for Amazon's Alexa team.

On a theoretical level, this study explores the idea of structuration of platform organizations (Dolata and Schrape 2023) and investigates the practices of infrastructuration (Edwards 2019) that the APC teams developed over the course of the competition. These theoretical tools are employed to analyze the perspectives of highly skilled developers who gain access to Amazon's Alexa technology by agreeing to develop solutions to certain problems set by Amazon. It can be shown how the developers navigate the conditions set by Amazon, as well as how certain technological path dependencies clash with new AI innovations taking place outside Amazon. As this transition in coding tradition is largely (at least in the public eye) initiated by the release of ChatGPT, the overarching interest in this article is to inquire into the APC participants' (shifting) perspectives on Alexa during this period of transition towards LLMs,

and to shed light on their development and innovation practices in this matter. Although the APC participants may not be employed by Amazon, they did receive insights into the corporation's development material, tools, and guiding principles for Alexa, informing them of the current state of the art of Alexa. Ultimately this gives insights into Alexa's sociotechnical underpinnings in a world that was at the time seemingly being revolutionized by a competing technology, and how Amazon and the APC participants attempted to merge existing with new technology, while at the same time navigating their relations of cooperation with each other in an ongoing process of platformization and infrastructuration (Plantin et al. 2018).

2. Research Object and State of Research

To introduce the object at hand, a brief outline of the APC and AI competitions in general is followed by a short summary of research on voice assistants (2.1). To further situate the research interest of this paper, a short overview of research on generative AI is then provided (2.2).

2.1 Studying the Alexa Prize Competitions

Many technology companies hold prize competitions and challenges like Amazon's APCs as a way of outsourcing algorithm development work. The cultural impact of these contests has been analyzed and the balancing of platform interests with complex engineering problems has been discussed at length in the case of the Netflix Prize (Hallinan amd Striphas 2016; Seaver 2022, 56–58). As such, the competition concept has served as the *organizing principle for AI* (Hind et al. 2024). Further, the events have been contextualized within the culture of competitiveness that is underlined by the practice of benchmarking (Orr and Kang 2024), as well as a platformized process that favors a few powerful actors (Luitse et al. 2024). The APCs have not yet received specific academic attention beyond the annual competition proceedings that focus on the computer science aspect (see e.g. Agichtein et al. 2023; Johnston et al. 2023; Shi et al. 2023). The APCs are a series of annual competitions that have been organized by Amazon since 2017, starting with the first Socialbot Grand Challenge (SBC). In that competition, Amazon encouraged universities across the world to create teams comprised of PhD students and professors to compete in a contest to develop a conversational bot that would drive Amazon's voice assistant Alexa (Amazon

2024). The challenge of the first SBC was to create bots capable of holding a 20-minute conversation with users talking to the bot (via Alexa) about various topics. In 2022 Amazon added the Task Bot Challenge (TBC) and in 2023 the SimBot Challenge (SIMBC) to its annual competitions. In the former, participants were invited to design bots that could enable Alexa to assist users in complex tasks such as cooking or origami, guiding users verbally through the various steps of a respective task. The latter challenge involved users talking to Alexa to control a robot in a video game environment to achieve small tasks (like retrieving something from a fridge) in said video game that simulates a living space. All of these competitions have a similar structure in time and incentive, running between eight and 18 months and divided into phases for certification (technical requirements of the bots that need to be fulfilled), internal feedback (Amazon employees provide intensive feedback on the bot), and public feedback (the systems go live and users can use the bots). During the last phase of the competition, the prototype bots are available to Alexa users in the United States. It is important to point out that this happens through a dedicated application, clearly separating the competition from the regular Alexa service. When a user invokes the corresponding skill for the competition they are randomly assigned one of the competitors' bots, without knowing which one it is – there is no way for them to target specific bots. After an interaction, users have the option to evaluate the bot with a star rating from one to five and a sentence of feedback. These ratings are used to rank the university teams on a leadership board that is updated daily, determining who advances to the final stage (which is a continuation of the previous stage but with less competitors and more users) and eventually determining the placement of the winning teams and the allocation of the prize money.

Studying this competition contributes to the body of research that undertakes sociotechnical analysis of voice assistants like Alexa, furthering understanding of the sociotechnical underpinnings of a technology that is present in many homes globally. Voice assistants have already been studied from multiple perspectives (Minder et al. 2023)[2]. Some research has addressed the plat-

2 It is important to note two prominent strands of critical inquiry into voice assistants, even though they are beyond the scope of this article. Firstly, there is the issue of the gender roles that voice assistants represent and perpetuate and in what ways this can be problematic; for a comprehensive overview see, e.g., Kennedy and Strengers (2020). Secondly, privacy and data security have received a great deal of attention because the devices can give companies access to data, e.g., from conversations, that

formized nature of voice assistants (e.g. Goulden 2019; Pridmore et al. 2019; Sadowski et al. 2021), but few studies to date have focused on the development process of voice assistants (Strüver 2023a; b). By qualitatively inquiring into the procedures of the APCs and competitors' experiences of working with Alexa technologies, it becomes possible to shed light on the inner workings of the sociotechnical relationships and dependencies that underlie Alexa. This is particularly interesting at a time in which speech technologies are prominent in public perception and critical discussion.

2.2 Large Language Models as a problem for Alexa

For a long time, the development of voice technologies was driven by turning linguistical conversation rules into code that determines how artificial voice agents detect users' intents and then give appropriate answers. This "rule-bound rationality of code-driven determination that animated the formative decades of AI research" (Li 2023, 168) was later enhanced by heuristic programming, which enabled more flexibility and improved performance. While stochastic machine learning models that approximate the most likely meaning of and answers to users' queries are commonly used in modern voice technologies (ibid.), for a long time, voice assistants like Alexa have retained some form of determinable answers and heuristics to ensure that certain actions follow certain queries (Kinsella 2023). This has often obliged developers to compile large sets of manually created answers (and templates) that were heuristically matched to what users approached the assistant with. The increased use of LLMs – achieved by the marketization and popularization of various tools and their integration into well-known and widely used applications – now seems to be set to strongly influence how voice assistants will be further developed in the future. Generative AI models like LLMs are a technological development that has recently risen in popularity in many applications for everyday use, with claims that the technology is revolutionizing the field of AI – in the familiar narrative of heralding *the next big thing* (Vannuccini and Prytkova 2024). As they have gained prominence and popularity, LLMs have been critically scrutinized from multiple perspectives (Fourcade and Healy 2024, 94). Essentially, they operate by a form of machine learning that utilizes vast amounts of data and computational power to perform various tasks that

---

they never had access to before; making security and trust controversial topics (see e.g. Mols et al. 2021; Ochs, this volume; Waldecker et al., this volume).

were previously complicated to execute with algorithmic tools. The humanities and social sciences have highlighted issues of diversity and discrimination in LLMs (Gillespie 2024), have questioned the agency of LLMs (Floridi 2023), and have contextualized the socio-political dimensions of LLMs on a global scale (Amoore et al. 2024). Further, scholars have criticized how much resources the training and maintenance of these models consume (Rillig et al. 2023) due to the enormous computing power they require. On an infrastructural level, this high consumption means that only a few firms can realistically afford to train these types of models, which has led to a significant oligopoly comprising the three largest Western corporations: Amazon, Microsoft, and Google (Srnicek 2022, van der Vlist et al. 2024). The significant rush in development that was precipitated by OpenAI's launch of ChatGPT has created an environment of hectic innovation. Like other companies, Amazon has sought to adapt products such as Alexa to the new LLM technology (Krishnan 2024), despite having previously reduced its development investment for Alexa due to poor business figures (Kim 2022). This has seen Alexa's development essentially reinvigorated by LLMs, which represent a new avenue for innovation that was previously underexplored for Alexa. Amid this global frenzy, as Tekic and Füller observe, universities are a key collaboration target for companies that wish to expand their access to the development of LLM technologies, as universities "are rare places where AI researchers – an expensive and hard-to-find resource – are grown" (2023, 5). This, and the fact that Alexa has traditionally been built with a heavy reliance on manually-coded heuristics only occasionally enhanced with LLMs (Jassy 2024), lead to the these main questions that motivate this paper:

The overarching purpose of the analysis is to elucidate *APC participants' perspectives on Alexa's position in the ongoing technological transition towards LLMs*, thereby also shedding light on Amazon's attempts to incentivize innovation in that direction. To contextualize those perspectives, the integration of LLMs into Alexa is examined against the backdrop of potential path dependencies in Alexa (5.1). Furthermore, the participants' technology development practices are focused upon in order to study the implementation of LLMs into the Alexa system from a *science and technology studies* perspective (5.2). Finally, sufficient context will have been provided for some conclusions to be drawn regarding the ongoing market competition between Alexa and ChatGPT and the role of the APCs therein (5.3).

3. Theorizing the Vortex between Platforms and their Complementors

In order to investigate the research interest, there will be a theoretical introduction into aspects of platform structuration. This begins by focusing on the platform organization's structuring capacity (3.1), which is then contrasted with the infrastructuration practices of developers (3.2).

3.1 Alexa as a Platform in the Alexa Prize

Sociological perspectives often focus on the companies behind the platforms and their power relations (e.g., Dolata 2019). Building on a combination of these perspectives, Strüver has conceptualized the voice assistant Alexa as a platform with multiple roles and purposes situated within Amazon's platform-ecosystem (2023b). He draws attention to the "unifying role for the smart home", that Alexa seems to hold, where it acts as a "connecting point for many different actors and technologies" (Strüver 2023a, 105) and the position of power in which this puts Amazon in relation to homes and businesses. These observations are guided by the idea that platforms and their complementors (Baldwin and Woodard 2008) can be conceptualized in a center–periphery model, with the platform as the locus of action governed by an organizational core that decides how the actors (e.g., users or third parties) interact with the platform through interface design (Ametowobla and Kirchner 2023). In this sense, it is important to understand the platform in a threefold distinction:

> (1) the platform-operating companies as organizing and structuring cores whose goal is to operate a profitable business; (2) the platforms belonging to them as more or less extensive, strongly technically mediated social action spaces not only for economic but also for genuine social activities; and (3) the institutionalized coordination, control and exploitation mechanisms implemented by the platform operators, linking these two constitutive levels of the platform architecture. (Dolata and Schrape 2023, 4)

This threefold distinction requires some tweaking when applied to the APCs, however, since in this case it is in Amazon's interest to continue to innovate

their technology in order to run a profitable business[3] by enabling and situating Alexa as a platform for innovation not only in the context of the competition but also for internal purposes. Applying the three distinctions to the APC, Amazon appears as a coordinating platform company that develops the platform Alexa and the sociotechnical environment of the competition. Acting as a space for a variety of social actions, Alexa becomes the platformized social environment for the APC, in which university teams develop new features, which are put to the test on users' Alexa devices. However, this social space within the platform environment subjects development activities to the constraints of coordination and control of the competition imposed by Amazon – which harkens back to the idea of periphery and center (Ametowobla and Kirchner 2023). In this sense, platforms coordinate not only economic processes, but also various social relationships, which can include the complementing innovation practices of independent third-party developers (Tiwana 2014, 118). The tools available to Amazon to control the platform environment are forms of "[c]oordination and rule-setting, monitoring and exploitation of data, coupled with the ability of the platform companies to quickly, substantially and largely uncontrollably adapt the social and technical rules they establish" (Dolata and Schrape 2023, 8), which locates the origin of power asymmetries between platform companies and the various groups of actors involved in the act of *platform governance* (Gorwa 2019). By means of the Alexa platform, Amazon has control over the technical development and standardization of third-party Alexa products, decides on the possible interactions with and within the platform, and, finally, sets the (contractual) rules, goals, and boundaries of collaboration between third parties and Amazon (e.g., van Dijck et al. 2018, 11; Gillespie 2018, 45–47). These rules, goals, and limits establish and maintain the *hierarchical orientation* (Dolata and Schrape 2023, 8). On top of those there are softer forms of control and orchestration which can act as action-orienting influences that are optional and malleable. These softer forms of control come as resources granted to the teams by Amazon prior to the competition (Agichtein et al. 2023, 3–13; Johnston et al. 2023, 4–12; Shi et al. 2023, 4–8). Exemplary, a Conversational Bot (CoBot) toolkit was offered, which represented a development tool for conversational AI with numerous pre-configured design presets for natural

3 While Alexa is reportedly not profitable for Amazon (Kim 2022), it can be argued that Alexa serves a greater purpose through cross subsidization, data usage, and algorithm development (Strüver 2023b, 21–25).

language understanding and dialogue management[4]. Amazon updates CoBot annually based on the learnings of the previous competition and to reflect ongoing changes in the industry, such as the recent shift to LLMs: "In addition, we also made significant changes in CoBot to support hosting large language models (LLMs), as much as 640 GB, which is 160 times larger than previously hosted in CoBot" (Johnston et al. 2023, 4). The Amazon scientists' highlighting of this latest adaptation of the CoBot tool alludes to the fact that platform companies have the ability to *re-code* their platforms dynamically to adapt to internal and external influences like regulations, new internal Amazon products, or a new competitor like OpenAI's ChatGPT. This *transformative re-coding capacity* enables platforms to dynamically readjust the sociotechnical structuring and institutionalizing elements of their platforms (Frenken and Fuenfschilling 2020, 103–107). Besides contractual changes, this capacity manifests in forms of *orchestration efforts*, i.e., new development tools, programs, application programming interfaces (APIs), microchips, standards, guidelines, or infrastructures of development (van der Vlist 2022; Strüver 2023a); as can be seen with the CoBot tool that was adapted during the release of ChatGPT, altering the competition: "Large language models (LLMs) have played a significant role in the SocialBot Grand Challenge since early in the challenge, but nothing compared to their front stage role" (Johnston et al. 2023, 3) in SBC5. Fittingly, this incentive to integrate more LLMs is transported via the main support tool of the competition, tying back to the goal to advance the science in conversational AI (Amazon 2022b), as well as to please customers, who are experiencing ChatGPT while rating Alexa skills.

Drawing on the distinction between platform company, platform, and the mechanisms of controlling interaction on the platform reveals the sociotechnical elements that allow Amazon to regulate what happens in the APC, which in turn facilitates conjectures to be made about corporate motives for these measures and an attempt to reveal the "high degree of structuregiving, rule-setting and controlling power" (Dolata and Schrape 2023, 14) that companies like Amazon possess. By giving this context on the power that is wielded by big tech

4 CoBot is a typical example of big tech companies leveraging their R&D facilities to develop products that are supposed to reduce innovation costs (Dolata 2019, 189), which eventually influence the development process when incorporated (Strüver 2023a, 114). CoBot "provides abstractions that enable the teams to focus more on scientific advances and reduce time invested into infrastructure, hosting, and scaling." (Johnston et al. 2023, 3)

companies when they structure their platforms, an important analytical step is enabled. Usually, the workings of such companies are largely *opaque* (Burrel 2016), especially concerning their AI technologies, which makes it difficult to investigate the impacts of platform technologies on users and third parties. By examining the resources that Amazon uses to run the APC challenges, it becomes possible to draw conclusions regarding the ways they act within their B2B collaborations, as well as how they develop technologies internally. Against the backdrop of the boom of LLM-driven technologies – which occurred while Alexa was struggling as a product (Kim 2022) – this approach can reveal how Amazon attempted to create an environment in which ideas could be developed for Alexa in a world of abundant LLMs. But to look into this practice of developing technology, a practice perspective on structuration is necessary, as structuration is not a deterministic effort made by Amazon that cannot encounter contingent resistance. Here, the tools of soft control are especially interesting, as they allow for leeway at the level of practice. In analyzing how tools of orchestration impact the APC, the room for negotiation and the limitations of resources of power which attempt to influence the course of action get revealed (Dolata 2024, 191) under the magnifying glass of practice that eventually reproduces or alters structure (Giddens 1984, 15–28). This shift of perspective allows the accounts of the participants to be read through the lens of the *mangle of practice* of developing Alexa at a time when LLMs were seemingly revolutionizing conversational technology development.

3.2 Platform practices as infrastructuration

As Plantin et al. (2018) argue, platforms can be infrastructuralized when infrastructures are platformized. This has also been shown to apply to voice assistants when users incorporate them into their daily lives as an infrastructure (Strüver 2023b). Infrastructures can be viewed as sociotechnical systems made up of a mixture of routines, artifacts, standards, plans, conventions, technological devices, or organizational institutions (Star and Ruhleder 1996, 113). These infrastructures can become central to everyday life when they are embedded in practices and subtend social, technological, and built worlds, as they do not need to be reconsidered in the moment of invoking them to perform a task (Slota and Bowker 2017, 537). This is true for users who rely on infrastructures, but not for the communities involved in the social, political, and economic work of building, maintaining, and upgrading infrastructures (Bowker and Star 2000, 109). All groups, however, learn to interact with in-

frastructures and their conventions of practice as part of membership in their given communities (Star and Ruhleder 1996, 113). In this respect, they adopt behavioral regularities that become (organizational) routines, which then come to be part of the functioning of infrastructure. Drawing on Giddens' (1984) structuration theory, Edwards describes this process of embedding infrastructural skills in humans' habits and skills as infrastructuration: "infrastructure both shapes and relies upon the continual performances or rehearsals of agents" (2019, 358). When users or engineers acquire the habits and skills to interact with an infrastructure as part of membership, they start playing a vital role in its functioning, thereby reproducing the structural elements. Giddens specifically remarks on actors' capacity for agency to make contingent decisions to be bounded by their perception (1984, 27), rendering these learned habits as a way "of black-boxing action patterns that may once have been deliberately chosen or designed" (Edwards 2019, 359), by providing infrastructuralized action scripts "on which users, maintainers and builders can all tacitly rely" (ibid.). In that sense, infrastructural practices become an embodiment of standards (Slota and Bowker 2017, 537) as they reproduce the (infra-)structures that enable them. When infrastructures are embedded in large sociotechnical systems, most decisions that govern the functioning of the system have been made without the active participation of either users or engineers. However, by adopting norms, routines, and habits and reproducing them in daily practice, these black-boxed standards can become invisible in practice without anyone's need to reflect on their origin, or on the choices that may have led to a particular design. This infrastructuralization of platforms and their logics defines how practices become entrenched in the structures of the platforms that enable them:

> once they [practices] become habitual and routine, these once-cognitive acts become quasi-mechanical. Most of the time, that is a virtue; they contribute to the smooth workings of infrastructure while remaining invisible themselves. Yet by burying choices and creating path dependencies, they can also have negative consequences, sometimes dramatically so. (Edwards 2019, 361)

This draws back to the structuring aspects of said infrastructure, since a well-established infrastructure can lead to path dependencies and sociotechnical lock-in effects due to large user bases that expect a certain functionality or an engineering team that is used to a familiar direction of development. With

such structural inertia, it is uncertain how many collective resources have to be leveraged to change institutionalized structures.

These sociotechnical path dependencies can lead to resistance to change, even in seemingly fluid electronic infrastructures (Star 1999, 389) such as platforms (Strüver 2023b, 24). Habitual and materialized infrastructures are manifested in the form, for example, of certain functions, algorithms, or company goals that have shaped Alexa since its conception and have become familiar to users and developers alike. They may have contributed to a reduction of contingency and made certain development paths more likely than others in structuring the platform Alexa. However, faced with the facts that, on the one hand, Alexa does not seem to be succeeding economically for the company Amazon (Kim 2022), and on the other hand, that competitors seem to revolutionize the fields of Alexa's core technologies, the corporation has incentives to question the viability of some structures that have guided Alexa for years, and to explore new ways of developing Alexa (Jassy 2024; Krishnan 2024). To investigate Amazon's responses to this situation, the idea of infrastructuration can be used to trace how competition participants developed common practices of development during the course of the contest and how they handled the integration of LLMs into their bots while negotiating the existing Alexa infrastructure, its limitations, and Amazon's elements of structuration. This turn towards the routines, forms of resistance, and power resources in practice and practical work can highlight how the new complex technologies being developed for Alexa were still embedded in a social system and an accomplishment of data practices, which "does not just happen on its own, but is manifested through everyday interactions between people, infrastructures, and established conventions" (Burkhardt et al. 2022, 11).

4. Study Design and Material

Studying the big tech companies of Silicon Valley from within is nigh impossible – at least if the study is to conform with the methodological standards and guidelines of sociology. The firms' inaccessibility is one of the reasons for choosing to investigate the APC, as it allows an insight into the inner workings of Amazon's Alexa team – or at least to the parts of it that competitors interact with. The other reason is that Amazon relies heavily on third parties for their core businesses (e.g., Khan 2018; Rowberry 2022, 42–43; Weigel 2023), so studying these can reveal how one of the world's biggest technology com-

panies conducts and manages its power relations. To inquire into the inner workings of Alexa and one part of its third-party ecosystem, a qualitative expert interview study was conducted with participants in the APC. 158 competitors from 2022 and 2023 were invited by email to take part in the study and offered a 25USD/EUR incentive to signify sincerity. This led to twelve one-hour interviews being conducted in early 2024. Nine interviewees were based in the USA, from diverse demographic groups within the population (Starr and Freeland 2023); the other three were in Europe. Overall, participants came from ten different university teams that had taken part in three different competitions. Seven were PhD students, two MSc students, and three professors in faculty and team-leading positions. Final placement in the competition of the teams whose members agreed to participate in the interviews was not skewed in any particular direction. Mirroring the uneven gender representation in the field of computer sciences, there were only two women in the sample of interviewees. An attempt to counter this was not successful, and the imbalance in the field was discussed in some interviews. Online video and voice interviews were chosen as a means of communication due to the global scheduling advantages (Self 2021).

The study was carried out with good intentions and the most academic rigor, but was nonetheless subject to some limitations. First and foremost, the interviews were conducted at the start of 2024 with participants who had competed in the 2022/23 APCs, which ended in August 2023. Considering the extremely fast pace at which LLMs are developing, technical judgements and statements made at the time of the interviews, as well as evaluations of Alexa at the time of the competition, may very well be outdated by now. Nonetheless, some intricacies of the transition between technologies can still be gleaned from this analysis. The guiding questions (Helfferich 2019, 676–677) for the study were designed to elicit details about the inner workings of cooperation with Amazon and to produce narratives by the interviewees reliving their course through the competition as they experienced it. In this sense, the interviews were equal parts qualitative narrative interview (ibid.) and expert interview (Bogner et al. 2014). The narrative component of the interviews aimed to evoke a more personal conversation tracing the participants' experiences, to complement expert knowledge, conducive to evoking statements about the competition that exceed a factual retelling. Participants had signed non-disclosure agreements with Amazon in the course of the competition. However, the chosen methodology seemed to alleviate interviewees' fears of breaking the terms of those contracts, as the conversations were

generally fluent and free in their flow. With participants' signed consent to the storage and usage of their data for scientific purposes, the interviews were locally recorded, transcribed, and anonymized; identifying statements were removed. Interviewees were assigned pseudonyms using a global random name generator (Bogner et al. 2014, 89–90). Analysis was carried out following the procedure of an inductive thematic qualitative data analysis (Kuckartz 2014, 70). In the following, interviewees' quotes are referenced by pseudonyms and the paragraph numbers of statements (Pseudonym, Paragraph number). All interviewees are referred to by the neutral pronoun "they" for inclusivity, and to protect their identities. The data sharing agreement signed by the participants does not allow the full transcripts to be made accessible to the public due to the sensitivity of the material.

5. Analysis: Perspectives on Building AI for Alexa

In order to address the overarching research interest – the APC participants' perspectives on Alexa's position in the ongoing technological transition towards LLMs – three topics are discussed in the following. First, the analysis focuses on the benefits, problems, and risks that come with integrating LLMs (5.1), then it compares two modes of actually integrating LLMs into Alexa (5.2). Lastly, an insight is offered into the role of the APC in developing LLMs in a competitive market (5.3).

5.1 Navigating the implementation of LLMs into Alexa

When investigating how integrating LLMs into the inner workings of Alexa relates to the conditions and structures that Amazon has set for Alexa, a great deal can be gleaned by addressing the benefits and problems perceived by the competitors of the APCs. A large portion of dialogues with Alexa are – or were at the time – determined by a heuristic that chooses from archetypes of manually-coded answers. This works well for easy-to-determine services like asking about the weather, turning on the living room lights, or asking trivia questions. Especially for more sensitive conversation topics, such as health advice, there are entirely preprogrammed responses that have been coded manually by engineers at Amazon, but this cannot feasibly be done for all the potential topics users might approach Alexa with. It can be assumed that when users talk to

Alexa, they do not want to constantly hear 'non-answers' that reveal the assistant's incapacity to engage in a given topic.

When competition participants as developers were preparing answers for the question of *what their bot's favorite sport was*, they might have included a list of dialogue options for popular sports, but probably did not consider every existing type: "We didn't cover everything. For example, for the other part [other sports], we could use the LLM" (Dart, 108). The flexibility of topics that can be handled by an LLM was one of their main perceived advantages, and was highlighted multiple times. Talking about sports is relatively simple, but "if it's something more involved, like: 'Oh, what are your opinions on Taylor Swift?', then the heuristic gets confused and there's no branch that matches it" (Scott, 42). While this comment addresses the same issue – that a heuristic model is unable to cover vast amounts of content – Scott's example concerns Alexa being asked about its opinion in a conversation. The implementation of LLMs could shift the structure of the conversation from a bot asking questions to users to instigate a dialogue and then posing follow-up questions, to a more flexible and reciprocal conversation model (Bardiola, 8; Centis, 29; Dart, 109). While the developers mentioned other advantages of LLMs, such as easier classification of users' responses via LLMs (Longwei, 87), or pre-trained models that can respond to sensitive topics (Gardé, 70), their flexibility was a recurrent theme mentioned throughout the interviews. It was particularly highly appreciated by competitors in the social bot challenge, who emphasized that LLMs can generate answers for any question, regardless of content. This reflects the structuring elements of the competition set by Amazon. The goal specified for the SBC: to achieve a 20-minute coherent and engaging conversation in two thirds of their bot's conversations (Amazon 2022b), clearly incentivized the implementation of a technology that enables flexible conversation. Further, Amazon provided various pre-trained models to facilitate this specific goal of "chitchat" (Centis, 29–32), which some of the participants included in their bots. Lastly, it is easy to imagine that an Alexa capable of sustaining longer conversations would generate more data that in turn can be commodified via the logics of platform capitalism (Srnicek 2022; Strüver 2023b), providing a further incentive for Amazon to pursue this goal. As Johnston et al. (2023, 24) reflect on the goal of the competition, they recognize that LLMs made the 20-minute goal very achievable while also pointing to some drawbacks of using LLMs.

The most obvious drawback is latency[5]. Multiple developers reported that adding more LLM capabilities to their bots increased the time that it took for the bot to answer, as generative models take longer than a heuristic model with pre-configured answers would (e.g., Breen, 44; Centis, 53; Dart, 10; Raju, 50). One developer elaborated upon the problem with latency by focusing on users' limited attention span and it being better to give a mediocre answer quickly than a good one really slowly (Scott, 43), because:

> Just latency is very, very important. And especially when you're talking to a bot; very, very frequently when our bot was good, but slow, we would see people just getting bored. Because you're sitting there trying to talk to this thing and waiting for like 10 seconds. And so, you just leave and give it a bad rating. … So, a huge focus for me was just trying to reduce those latencies. And to that end, we used other Amazon products and things databases for smart caching and that type of thing. (Scott, 18)

Scott's remarks point to several effects of structuration. For one, using Amazon tools that help in the process reflects a form of orchestrated efficiency. Further, Scott mentions their dependency on the feedback stars of users in the later stage of the contest, which is one metric of success in the competition. As "platform participants", users are "integrated into the monitoring and control systems of the platforms as decentralized co-controllers" (Dolata and Schrape 2023, 13). The resulting pressure to balance quality against latency is part of an infrastructuration process whereby the teams decide to what degree to include LLMs despite their increased latency, and then observe how their decisions are received as reflected in users' ratings. These are contingent decisions that the teams make; another participant described a different prioritization: "There are a lot of constraints on resources and latency using large language models,

5 It has to be noted that eight of the twelve participants emphasized lack of resources while simultaneously mentioning problems with latency. They deplored constraints on computing resources and funding, particularly as running an LLM is costly in both. Put poignantly: "working with machine learning is very expensive at this point, and if you don't have enough computer resources, then you fall behind" (Chidi, 101). Which puts an emphasis on the unequal conditions that generative AI is being developed and distributed in, as there are very few companies that are able to supply the capital and material basis for large-scale LLM usage (Srnicek 2022; Luitse 2024; van der Vlist et al. 2024).

and given the time constraints we got something working fast and then never replaced it" (Breen, 44). The potential for agency in development is thus limited by users' ratings, which teams are obliged to heed if they want to succeed and stay in the competition.

When talking to a voice assistant, users generally expect the assistant to respond to their query in a fairly reliable way. Users can only assume that assistants will perform their various algorithmic language processing steps correctly and give appropriate answers (see, e.g., Strüver 2023b; Hector and Hrncal 2024). However, the developers interviewed indicate that integrating LLMs into their Alexa bots can potentially lead to a reduction in the reliability of answers, as engineers have limited control over the quality of responses: "up to some point, we can control the quality but we cannot guarantee 100 % quality every single time for every topic" (Dart, 107). This can lead to bots sometimes *not giving good or correct answers* (Chidi, 111), especially in comparison to the entirely controllable scripts (Dart, 111) of heuristic models. Some teams decided to incorporate less LLMs specifically for this reason. Dart mentioned that with an increased proportion of LLMs within the bot, it could "hallucinate" (Dart, 16), which was also mentioned in the official recap of the SBC5, alongside contradictory answers (Johnston et al. 2023, 24). Thus, a certain volatility leaks into the system when implementing generative AI into Alexa bots. As the inflexible heuristic scripts are one of the oldest forms of machine learning (Li 2023), the resources to control their outputs are well established and institutionalized by professional education and tools, serving as forms of structure to produce reliable answers from Alexa. Comparably, LLMs are relatively new and seem to show a lack of established practices of control, leaving the teams to deal with the tasks of infrastructuring on the fly. One participant put the importance of controllable answers into perspective as follows:

> You have to work on those safety features. It will be more harmful if it comes out of a voicebot instead of just a chatbot, right? There are cases like that. I think there are much more things to do before they can just use ChatGPT in a voice assistant. And I'm sure there will be legal consequences, too. Because children use the voice assistant because they do not have access to ChatGPT. (Chidi, 141)

Safety features that have yet to be developed for the integration of these types of LLMs could be a way to increase robustness of input and output. On the one hand, Chidi points to the less specifically explicated queries that are expressed

orally; which users would have to adapt in time, as they *learn how to talk* to voice assistants (Habscheid 2023, 185–186), while establishing new routines. On the other hand, the fact that voice interfaces are more accessible to, for example, children, due to their specific characteristics as a medium (Soffer 2020, 932), can cause problems when considering the lack of quality control. At the same time, developing more reliable institutionalized methods of structuring and controlling answers given by generative AI is in the interest of Amazon from a brand perspective, structuring the development of Alexa. Emily West calls the brand of a company the experienceable *face* for consumers to interact and relate with, impacting a company's success. Seemingly, Amazon's branding and advertisement is intentionally innocuous, attempting to achieve familiarity while offering minimum identity. Amazon's brand is defined by the affective convenience and ease of use of their consumer products (West 2022, 25–27). Alexa, too, is supposed to convey exactly these unobtrusive brand points, as it acts in a way of *idealized servitude* (Phan 2019, 29) that does not draw attention to itself but simply functions as a reliable touchpoint for users and enables *frustration free* (Strüver 2023a) service. Amazon "builds an affective relationship with its customers through interaction. And a key part of that interaction is reliable access to and efficient delivery of goods, making the affective relation tangible and touchable on a regular basis" (West 2022, 31). Perceiving Alexa in the light of the importance of this type of convenient, familiar, and reliable branding that is mainly conveyed through interaction highlights how volatile answers of an LLM-driven Alexa could threaten this brand image. Answers that are wrong, contradictory, or offensive, and easily accessible to all household members, could tarnish Amazon's reputation. Which is even more important considering that users' trust in voice assistants has been shown to correlate strongly with their sympathy towards the company behind the assistant (Weidmüller et al. 2022, 644). It is therefore no coincidence that Amazon actively applies internal and external quality control measures and moderation to protect its good reputation from unintended consequences of innovation, and strongly incentivizes high conversation quality during the APC.

While some developers report that the frameworks provided by Amazon struggled with interaction with the real world (Erwin, 96–98; Pak, 101), one participant rounds this discussion off with a succinct contextualization of different programming approaches for voice technologies:

> Because a lot of what makes ChatGPT seem so amazing and so impressive is that there's nothing at stake with the answer being correct.

> And if it works 90% of the time, it's like 'wow this works 90% of the time', but what are the situations where being wrong 10% of the time is okay? (laughs) I think that's something that we don't really have a very good answer about and we don't really have a very good answer about what the real trajectory is for getting kind of more accurate information out of these things …
> Think about the way that Siri was built, or the way that the existing assistant functionality is built on the Alexa devices for example; you know those systems were built in a particular way to make sure that they had predictable accuracy. Where in some sense once the speech recognition could be as bad as you like but if the words got recognized correctly, it would play the song that you asked for. (Breen, 73–74)

This reflects how Alexa was originally built with classical and established machine learning tools. It produces reliable results to specific queries. Which is what Amazon has built its market share on, especially in the domestic internet of things, where Alexa acts as a central hub to coordinate smart home devices (Strüver 2023a). As long as these problems prevail, preserving this functionality and position in the market serves as a strong incentive for Amazon to not completely switch to LLMs. Amazon might not desire to break the institutionalized usage of Alexa in users' homes:

> There are a lot of low stakes and kind of information access applications where ChatGPT is sort of a plausible current tool; but for things like assistants that have to hook up with something that's happening in the world, where the outcome matters, it's a lot further away than it might look. Just because you want to be able to have some guarantees. (Breen, 75)

This emphasizes LLMs' weakness of reliability, especially in interactions with the real world, where they could be implemented into material processes and routines. Assuming that users integrate Alexa as a device to control their smart homes – as intended by Amazon – and have performed a sense of infrastructuration in establishing routines with the device, they have black-boxed certain aspects of those interactions and presumably would not want to reconsider their smart home infrastructure on a daily basis: it would be against the use case to have to ask Alexa three times to turn on the lights or to lock the door. With Alexa already embedded in smart homes across the globe, users have developed certain path dependencies. However, these can be broken if the device

ceases to provide the technical infrastructure that enables the promised convenience and reliability of Amazon's brand. Especially this connection to the smart home leads to questions around the technical implementation of LLMs alongside more traditional ways of developing the assistant, which will be explored through the developers' perspectives next.

5.2 Implementing LLMs into Alexa: Deciding who talks to the user

Against the backdrop of the risks and benefits of LLMs and their implementation into Alexa, the following will look at the practices of infrastructuration that the developers describe when integrating LLMs into their Alexa bots. Corporate interests of staying innovative and profitable during a time of technological innovation seem economically rational, as Alexa and the developers face the repercussions of a competitor releasing a popular new technology: "Suddenly, users were expecting much better conversations than what was achievable by the stupid rule-based systems that we started with" (Centis, 35), and, consequentially, many users tried to tease Alexa (Gardé, 48). Breen compared the Alexa experience prior to the advent of ChatGPT to a call-center-AI that guides users through the functions that it can do effectively and concluded: "that's essentially the opposite of the design patterns that are rewarded in this Amazon competition" (Breen, 66). This presents an assumption on the structure of the competition set by Amazon, which gets reinforced by the fact that Amazon provides an API for detecting when a user found a conversation boring or wanted to terminate it (Bardiola, 115). According to the interviewees, users were essentially expecting Alexa to be more than it used to be, and generative AI was seen as one tool that could achieve that by providing more flexibility to react to different topics, which Amazon structurally incentivizes by the competition design and the resources it offers. If the teams accepted this structuration of their innovation process, they needed to establish when to use an LLM and when to deploy classical heuristics to talk to the users. More often than not, this decision was rather an accomplishment in practice (Burkhardt et al. 2022) that was influenced by means of structure, than a general ruling, as is explored in the following.

5.2.1 Building a pipeline: Classifying criteria that govern when to swap between models

"There's usually a fork in the road. You try and see if there's an easy non-AI response you can give" (Scott, 42). This remark generally applies to *if-statements*

that can be dealt with by simply programmed conversational heuristics that are well established and institutionalized through open-source models, but also through tools like CoBot (Bardiola, 114) that are developed by Amazon based on their experiences with Alexa and therefore come with a certain range of answers and topics. The most prominent examples were conversations about sports, or the types of food liked by users, i.e., contexts where the space for answers was easily categorizable. If the topic is outside the scope of the predetermined heuristics, using an LLM seems evident. But remembering that developers *limited how often they used LLMs* because doing so was expensive and introduced latency, gave an incentive to further complicate this decision process of deciding which models users talk to. The question became about how to combine these different approaches. Developers described how they arrived at a "blend of pre-scripted dialogues and the new answers generated by the new generative models" (Bardiola, 12), by building a *pipeline* (Chen, 100–101; Raju, 52; Chidi, 108; Dart, 18) that used multiple components to create a "hybrid approach" (Dart, 107) between different models that the Alexa bot[6] used to talk to its users. The word pipeline – albeit an industry standard-term – evokes a tangible image of infrastructure that *matters* (Slota and Bowker 2017, 530): it guides data through different checkpoints and permits certain functions while prohibiting others, transporting backgrounded contingent values and decisions. Even before considering the concrete pipeline implementation, developers had to take stock of which available existing heuristics they wanted to continue using. These could range from previous work in the field, open-source resources, or self-made models, to the tools and resources provided by Amazon. One interviewee reported that their university had had a team participating in the competition for several years (it is common for the same team/faculty leader within a university to have a changing team of students that participates annually under a similar name) and had built its own repertoire of manually-coded dialogues, which they liked to keep using:

6 While the analysis here concerns determining which type of technology is used to talk to users when, it is important to remember that there are differences between the regular Alexa and the Alexa skill that users access to talk to the Alexa bots developed in the competition. The latter is not congruent with the regular Alexa. Additionally, users can get confused by the competition skill, having expected that "they [would be] speaking to the same bot, but in the end they got one of the nine." (Bardiola, 113) This introduces another layer of 'who is the user talking to?' that is specific to this competition.

> The previous rounds of Kunkka [anonymized team name], the bot I was working on, they also used LLMs. But now we are focusing a lot on using them and employing them even more. What we did was, we were trying to enrich those [manually-coded] dialogues. So, use the dialogues that we have, because they are good. And, the quality is, I would say, very nice. We didn't want to discard it. It also could make the things a bit tough, because we were not starting freshly. I think some teams did that; they could come up with the whole architecture from scratch. But we are already using something. We were kind of limited in some sense, to what we are able to do. (Dart, 107)

What the member of team Kunkka described here is the process of infrastructuration in situ over the span of several annual competitions as described by Edwards (2019). Situational decisions made by previous teams to develop, use, and expand manually-coded heuristics for their bot (which, in Giddens' sense can be seen as rational, given the bounded temporal perspective of each team's efforts, because LLMs were far less capable in the previous iterations of the competition) become black-boxed, routinized, and materialized in the systems that subsequent teams use for later competitions. With the competition taking place annually, the decisions made by previous teams to use manually-coded methods do not need to be reconsidered in the moment of setting up the infrastructure for the next competition. This infrastructure is learnt as part of their team membership; with usually the faculty or team leader remaining the same to convey practices. Further, this institutionalization of infrastructural practices is reinforced if a team did well in the previous years because their process of infrastructuring has been structurally validated by Amazon and the users. Ironically, this makes teams with a proven infrastructure resistant to Amazon's orchestration measures to a degree – e.g., Dart described their team's active non-use of CoBot, for better or worse: their existing infrastructure enabled certain actions and limited others. In order to reconsider their infrastructuration process and respond to the call of implementing LLMs, they needed to question their routinized decisions, examine what they would like to retain, and eventually find ways to merge the existing base with new models. However, because they had a solid basis before the competition started, they were in the luxurious position of being able to evaluate whether they perceived the extent of power exercised via the means of structure and orchestration to be pervasive enough to warrant changes in their bot and to what degree. In this example, the con-

cepts of *duality of structure and action in a reciprocal reproduction* (Giddens 1984), as well as infrastructurized path dependencies are tangible.

Keeping in mind this perspective of situated practices that get institutionalized through the ongoing (re)production of structures within practices helps one to understand how the developers solved the problem of merging established systems with the new LLMs from a procedural perspective of everyday interactions. The member of Kunkka described the process of *injecting* phrases generated by LLMs into their bot as a phase of *constant experimenting* as they tried to merge the two approaches. In order to do that, they reported having to invent "ending criteria, when to end the dialogue, when we should switch to it" (Dart, 110). This short description hints at the process of decision-making involved in merging the two systems by building a pipeline, that guides data flows: The developers needed to establish rules for the usage of LLMs in a conversation, considering the prevalent action-structuring elements like constraints of resources, latency, and quality control. In all likelihood they switched to an LLM when the conversation topic or prompt was beyond the scope of their manually-coded heuristics. They then needed to find a way to define and *classify* (Bowker and Star 2000) a point in the conversation when it could be transferred back to the heuristics model while adhering to acceptable conversational conventions (as incentivized by the APCs goals). This again represents a case of developing a technical infrastructure that is accomplished by a string of decisions that eventually get black-boxed within a model, representing a switching mechanism to decide which type of machine learning the users talk to. The process of black-boxing makes their decision processes transparent and imperceptible in practice to users, as it has not to be reconsidered in conversation with the Alexa bot. A switching mechanism like this exemplifies how *opaque* conversation with Alexa can be, as it shows how during a single conversation, multiple switches can take place, with users talking to different algorithms that have different strengths and biases and are built in fundamentally different ways. This evokes the previously elaborated topic of suitable application space for LLMs and the question of "what are the situations where being wrong 10% of the time is okay?" (Breen, 73), as developers are obliged to make decisions that have significant consequences for users[7]. Hidden to users remains the decision of how much priority is given

7 This problem is exacerbated by aspects of unintentional events: Complex conversational models that switch between algorithms often need to have another superseding model that can repair the flow of dialogue should the bot fail to keep its

to quality or accuracy in a particular scenario, i.e., whether a human-written heuristic model is answering, or a generative AI with a higher volatility. This is a hyperbolic problematization however, as obfuscation of this kind is structurally incentivized and normalized by the aspirations of Amazon, which sets the goal of fluent conversation with Alexa – unimpeded by drawing attention to precisely these infrastructural technicalities worked out here. Ultimately, users simply talk to Alexa as some form of actor, regardless of the subtending model.

5.2.2 Transitioning between algorithmic approaches through testing

To further understand how LLMs can be integrated into Alexa bots, the previous approach to implementation can be contrasted with the option of prioritizing the implementation of LLMs. During the 2022/23 APC, an abundance of LLM models were getting published at a fast pace, where "papers are literally coming out every single week at this point" (Chidi, 121). This led to a volatile environment of rapidly changing models as the participants tried to implement generative AI into their bots: "Several times during the competition we changed the main model. It was not just [motivated by] Amazon; it was mostly new models appearing on the market. And you're like quickly redoing everything to make sure that it would work better" (Gardé, 76). Furthermore, Bardiola pointed out that finding and implementing suitable LLMs into their bots was not as straight forward as one might imagine (Bardiola, 41). With the perceived need to constantly exchange suitable LLMs, deciding how to introduce LLMs into the bots required developers to consider possible practices and infrastructures of testing algorithms. One of Amazon's central advertising points for the APC is the contact to the Alexa user base and the promise that "the immediate feedback from these customers will help students [the APC developers] improve their algorithms much faster than previously possible" (Amazon 2024). Live testing is a core function of the Alexa platform for Amazon (Strüver 2023b, 15–17) and is reproduced by the

---

outputs oriented towards the goal that the user is trying to achieve in their conversation (Erwin, 36). Further, Bardiola (117) explained that if an LLM malfunctioned on the weekend, or during the night, when their team's support service was offline, they would let the bot refer to Amazon's inferior and less specialized LLM as a backup. Ensuring the uptime of a service is structurally enforced by Amazon's certification standards for technologies that interact with Alexa (Strüver 2023a, 113). Developers' nods to the crucial work of maintenance (Bowker and Star 2000, 160–161) from and on the bot further complicate the question of who is talking to the user.

teams, when they rely on the platformized mechanisms of feedback established by Amazon. While assessing the applicability of LLMs for Alexa, this is highly interesting, as their performance is more complex to measure and goals like fluent conversations or succinct guiding through a task are hard to quantify. *Benchmarking* is a prevalent and highly institutionalized practice among machine learning researchers and involves the constant attempt to outperform previous algorithms within a *competitive computing culture* (Orr and Kang 2024). Usually, algorithms are compared by means of quantifiable measures like how long it takes to execute certain standardized tasks, which can also be applied to LLMs. Quantifying a successful conversation, however, while not impossible, is more complicated and subjective than calculating an algorithm's efficiency at transcribing speech. Against this backdrop, the testing process gains another dimension, as developers reproduce the competitive computing culture of their academic discipline by frequently changing models in the hope of improving performance as well as being incentivized to use the resources provided to them by Amazon – which sets the APC up in a way that also reproduces this culture. Here, motives of constant refinements endorsed by Amazon become conflated with the normative goal of striving for improvement that is inherently cultivated by universities and places of education of this profession and, correspondingly, research field: "Machine learning researchers are always very optimistic [about algorithms] because it's just the way they're hill climbing and of course if you can make the thing one percent better every year, eventually it will be very, very good" (Breen, 74). Recognizing this institutionalized motivation to implement different LLMs contextualizes the process of navigating the intersection between LLMs and heuristics, as described by Scott in their step-by-step account of how their team incrementally replaced heuristics with LLMs in their bot:

> Scott: I mentioned the heuristics and using LLMs earlier. When we started off, a very, very major chunk of our code was just heuristic-based [manually-coded]. And we only really used an LLM if all the heuristics failed and over time our big transition was having fewer and fewer and fewer heuristics and more and more LLM. And quite often we'd run A/B tests where we got rid of a huge chunk of heuristics and check to see if the model still did well, and oftentimes it would fail and not do well. Then we'd have to go in and investigate and debug and figure out why.

> Interviewer: When you investigated, how did you do that?
>
> Scott: We looked at our ratings. We looked at the average latency in a response [the pause between turns]. We looked at what the actual response was and what it was in response to; what the user said and what the bot said, and we looked at whether it made coherent sense. Oftentimes it wouldn't. And we just investigated by looking at common failure modes. And then you try and reproduce the failure modes, once you put in your supposed fix and if it still fails, then clearly your fix hasn't worked. In that sense, it was very specific in that you look at specific examples and try and fix those.
>
> Interviewer: Sounds like looking through a lot of conversational logs, right?
>
> Scott: Yeah, that part is a lot more tedious to do and for sure you can [do that]; but it's a lot easier to just look at… Over time, as we started to have thousands of conversations, it's easier to just look at conversations that perform poorly and see what specifically might've failed. (Scott, 50–54)

This account highlights how integrating an LLM into the Alexa bot is a highly contingent task that requires extensive testing and verification. Starting with a major portion of their code being heuristics-based, this team transitioned incrementally to utilizing more LLMs by replacing functionalities and constantly validating if each new functionality performed according to expectations, adjusting accordingly, and then reevaluating. To test their changes they employed A/B tests, which *continuously and seamlessly change* (Marres and Stark 2020, 434) the version of the bot that different users interact with at a particular moment in time. The A/B tests described here presumably compared the largely heuristic model with a new version of the bot that had some parts of its conversational heuristic model – e.g., labelling a user's intention through natural language understanding (Longwei, 94) – replaced by an LLM. In such a scenario, one user would talk to the baseline bot as version A and another user would talk to a version B of the bot that has a new LLM element added. The developers can then compare the conversations held by the two versions of the bots, either directly or through metrics. Due to the large volume of conversations, Scott described surveying the metrics' latency in the new version and low user ratings in order to identify outliers. In turn, these metrics helped to locate problems in spe-

cific conversations for closer investigation. Moving from abstract to concrete, the subsequent analysis of the actual conversations, which sought to ascertain problems in the LLM – such as a generative model producing random characters, as reported by a different team: "instead of saying a normal sentence it started generating stars and hashtags" (Bardiola, 87) – served as the basis from which to fix the model and repeat the testing process. As explained by Scott, this procedure for testing the integration of LLMs enabled specific undesired conversations to be targeted.

At this point it is important to recall the characterization of the developers' relationship with the platform organization that develops the platform and establishes institutional rules for how third parties and users can access the social space of the platform (see 3.1). In describing and analyzing the need for extensive testing when implementing LLMs into Alexa bots, two points emerged clearly: on the one hand, users are implemented into the competition as a development tool; they serve as agents of moderating and testing the bots and provide feedback to the APC teams as they navigate the process of integrating LLMs into Alexa. As mentioned earlier, this is a typical aspect of platform companies that involves users in a very calculated way as "decentralized co-controllers" (Dolata and Schrape 2023, 13) to shape, moderate and develop platforms and to re-code them if necessary. This is especially interesting for Amazon considering the lack of established ways to benchmark conversational AI models. Users function as an evaluation instance that does not need to be given specified classifications or criteria to define the diffuse goal of better conversation quality, which makes user interaction via Alexa an even more valuable resource for Amazon. On the other hand, the APC teams get feedback in a form that is determined by Amazon, as every interaction (ratings, comments, and text logs) that they have with the users is structured by the boundaries and conditions of the infrastructure set up by Amazon. Further, Amazon's choice to represent all the contestants' bots as a single Alexa skill that is specific for the APC (which can create confusion among users), instead of making them available as part of Alexa's general service is an act of moderation. This measure protects the brand of Alexa from potentially being associated with faulty bots, while it also opens space for experimentation within the competition, allowing different standards to apply within this dedicated test environment. Generally, while curating a data set is difficult in the APCs' test environment, this is definitely a caveat to the competition. The data set that provides the basis for testing algorithms is absolutely biased to users in the USA, as the Alexa skill for the competition is only available there. Furthermore, it could over-

represent certain demographics, who choose to interact with the APC skills (Centis, 76). Otherwise the data set is seemingly uncontrolled in terms of diversity, which could lead to cultural as well as linguistic biases in the testing of algorithms that eventually might be rolled out onto Echo devices globally. Unlike other AI competitions, in which efforts are made to provide a suitably representative data set for testing, which need to be sufficiently diverse for a technology to be applicable globally (Luitse et al. 2024, 17), such issues are not addressed in the APC. This examination of the ways in which developers test their algorithms when transitioning between heuristics and LLMs thus reveals how Amazon leverages the interaction of the university teams with users of the Alexa platform to develop technologies and institutions for Alexa. Knowledge production on the transition between heuristics and LLMs in the competitions is (unsurprisingly) inherently colored by Amazon's platformized structuration measures and values. The two quoted interview excerpts about development practices at the intersection of LLMs and heuristics can be read as an analogy to the predicament of Amazon's Alexa team: It can only be assumed that the situation that Amazon's Alexa team found itself in during the first year of ChatGPT was shaped by similar reconsiderations of path dependencies and of structuration, as Amazon came to face an external influence that led it to question the viability of maintaining its long established reliance on heuristics. The different ways of navigating the transition between the two machine learning approaches that were being developed in the APC will most likely find their way into the main Alexa system in some form, as they represent somewhat established practices of merging, switching, and testing. Moreover, Amazon's own methods of testing for Alexa are not restricted by the limitations on information that are imposed in the competition; Amazon-employed developers have access to far more comprehensive interactional data (Strüver 2023b). This background can now be contrasted with the competition against ChatGPT and its influence on the APCs.

5.3 Catching Up with Innovation: The APCs as a Testing Ground for Alexa-LLMs

Following these insights into LLM development practices for Alexa, the APC can now be situated within the larger scope of the competitive market of LLM products, especially the popular ChatGPT. During the runtime of the 2022/23 competitions, users across the globe were being introduced to the capabilities of ChatGPT and began to expect similar functions from Alexa. With users

slowly re-institutionalizing what AI agents were expected to do, OpenAI and ChatGPT entered the equation of Amazon's platform structuration. According to some interviewees, the reason for banning use of ChatGPT in the APC was "Because then it would be just easier to go: 'OpenAI, generate a response, be a social bot'" (Dart, 19). While it may seem fairly unremarkable that the use of a competitor's product would be prohibited in an innovation challenge that is intended to proprietarily advance Alexa, the motivation behind this ban is further contextualized by the APC developers' descriptions of the technological status quo of the Alexa system that they came to know during the competition. The LLMs provided by Amazon were, according to participants, along the lines of robustly processing text to find similarities (Breen, 73), and far from reliable or satisfactory to generate coherent utterances (Longwei, 93). Longwei predicted that Amazon's template-based heuristics system would not be used in future APCs, but concluded nonetheless that it "would be kind of hard for Alexa to switch from their previous path to really open for large language models" (Longwei, 95). While exemplary, these sentiments convey the state of Alexa technology at the time that ChatGPT was unveiled. Although it is possible and probable that the APC developers did not get a comprehensive overview of all the ongoing developments at Amazon, their accounts certainly reflect the state of technology that was being offered to third parties wishing to work on the Alexa platform. Assuming that these statements do indeed offer a reasonably accurate estimate of the state of technology of Alexa at the time, it does not surprise that Amazon was undergoing a comprehensive restructuring of organizational resources in the Alexa team (Kim 2022) and announced new plans for Alexa and generative AI in general (Bensinger 2024; Krishnan 2024). In this light, banning the use of ChatGPT in the APC should be seen as part of the measures of restructuring development of the platform Alexa. As a platform organization, Amazon is intent on leveraging a multitude of resources for the further development of Alexa as a technology and platform. This includes the APCs, as Gardé put it: "everything that we developed basically would be owned by Amazon. So, it's a good way for them to get lots of input on different areas of generative computational AI" (Gardé, 142). Allowing the use of ChatGPT could forego the development of possible technological approaches to solutions for problems that Alexa faces. The APCs that took place at this juncture of conversational technology development need to be seen from the perspective of being one of the tools of innovation – at the *periphery of the platform* (Ametowobla and Kirchner 2023) – that Amazon was utilizing in its efforts to orchestrate the development of Alexa.

As is standard practice for big tech companies, Amazon also complements their in-house R&D by buying existing start-ups (Dolata and Schrape 2023, 7). However, compared to such corporate takeovers, the universities involved in the APC represent a looser form of cooperation that is absolved of the need to be economically viable, which enables a distinct room for innovation but also involves different resources of structuration for Amazon. In the APC, Alexa is specifically not an industry platform for innovation on an equal footing (Dolata 2024), but rather a platform that enables Alexa-centric cooperation with university teams. These teams are more malleable and susceptible to Amazon's orchestration efforts in particular ways – the interviewees mentioned gaining industry experience and recognition alongside potential future job offers in the field as motivations for participating in the APC, as well as sought-after funding for their labs and PhDs. Such involvement in the education system can eventually play a structuring role in shaping the field's values and aligning them with the interests of companies that end up employing the – highly sought after (Tekic and Füller 2023, 5) – graduates. In that way Amazon can attempt to let the participants adjust to Alexa's infrastructural path dependencies and let them experiment in developing approaches to transitioning between heuristics and LLMs in ways that comply with Alexa's brand: "Sometimes you can't just replace everything with the new technology. You have to kind of find the right balance between using the new tools and previous tools" (Pak, 100). These observations echo what Luitse et al. conclude from their research on medical AI platform competitions: "the configuration of platforms, competition organisers, and participants concentrates power toward a small number of actors" (2024, 16). In the case of the APCs, this effect is compounded as both the actors of platforms and the competition organizers are represented in unison by Amazon, who can therefore direct the goals of knowledge production towards certain problems, e.g., the transition of a heuristic Alexa towards LLM integration, as is evidenced in the papers published in the proceedings of the SBC[8]. It still remains to be seen whether the models that were developed in the competitions will ultimately find use in Alexa (Longwei, 89), or whether, like the Netflix competition's winning algorithm, they will never be implemented (Seaver 2022, 58). In any case, the APC represents an R&D resource that can be utilized in attempts to re-code Alexa as a platform, but it is a resource that nonetheless remains hard to control due to the contingent development practices of university teams.

8 See https://www.amazon.science/alexa-prize/socialbot-grand-challenge/2022.

6 Conclusion

While the actual workings within Amazon remain opaque, the study did its best to fairly portray the experiences of the interviewed developers. The analysis presented here contributes to the understanding of how Amazon cooperates with third parties that work on the Alexa platform and shows the effects of hierarchical structuring while also highlighting the practical decisions and opportunities for resistance (e.g., not using the CoBot tool offered) that arose during the competition. This helps to critically understand the sociotechnical underpinnings and environments of the development of a technology that is used by many users on a daily basis. This is conducive to the understanding of how modern AI systems are developed and the risks that accompany ongoing changes in technology development. Insights such as these can contribute to shifting the academic discourse in the social sciences and humanities away from a focus on data to concentrate on deepening understanding of the sociotechnical circumstances and means that shape AI development (Srnicek 2022). In the study reported on here, a sociological perspective has been taken to investigate Alexa as a platform and infrastructure and to examine the practical accomplishment of development under structuration. This contributes a genuinely sociological understanding of platforms by empirically scrutinizing Amazon's structuration efforts and the infrastructuring acts that can be found when third party actors such as universities interact with a big tech company like Amazon.

Future studies could expand on this work by building on the arguments presented here and investigating the extent to which they can be applied to different AI technologies like other voice assistants, or using them to inform studies of Alexa usage in the home, or to look into whether LLMs have actually been incorporated into Alexa since the transition described here. As the famous Netflix competition shows, these types of (AI) technologies tend to be ephemeral and even a solution that emerges victorious from a competition might be too complicated to be implemented, or the organizing platform might change its business model, making the solution obsolete (Seaver 2022, 58). What remains, however, are the insights into how technology development is undertaken at the cutting edge of competition, and into the conduct in cooperation of one of the biggest tech companies of the present moment; a corporation that impacts the lives of millions of users globally every day.

Acknowledgements

The author has no conflicts of interest to report. I would like to thank student assistant Aileen Halbe for the data wrangling of the contact emails. Further, I would like to thank the editors for the feedback provided on the article. Lastly, I want to thank all interviewees for the immensely insightful conversations. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 262513311 – SFB 1187. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 262513311 – SFB 1187.

References

Agichtein, Eugene, Michael Johnston, Anna Gottardi, Lavina Vaz, Cris Flagg, Yao Lu, Shaohua Liu, et al. 2023. "Advancing Conversational Task Assistance: The Second Alexa Prize TaskBot Challenge." In *Alexa Prize TaskBot Challenge 2 Proceedings*. https://www.amazon.science/alexa-prize/proceedings/alexa-lets-work-together-introducing-the-second-alexa-prize-taskbot-challenge.

Amazon. 2022a. "SimBot Challenge Rules." Amazon Science. 2022. https://www.amazon.science/alexa-prize/simbot-challenge/rules.

Amazon. 2022b. "SocialBot Grand Challenge Rules." Amazon Science. 2022. https://www.amazon.science/alexa-prize/socialbot-grand-challenge/rules.

Amazon. 2022c. "TaskBot Challenge Rules." Amazon Science. 2022. https://www.amazon.science/alexa-prize/taskbot-challenge/rules.

Amazon. 2024. "Alexa Prize." Amazon Science. 2024. https://www.amazon.science/alexa-prize.

Ametowobla, Dzifa, and Stefan Kirchner. 2023. "The Organization of Digital Platforms: The Role of Digital Technology and Architecture for Social Order." *Zeitschrift Für Soziologie* 52 (2): 143–56. https://doi.org/10.1515/zfsoz-2023-2012.

Amoore, Louise, Alexander Campolo, Benjamin Jacobsen, and Ludovico Rella. 2024. "A World Model: On the Political Logics of Generative AI." *Political Geography* 113 (August): 103134. https://doi.org/10.1016/j.polgeo.2024.103134.

Baldwin, Carliss Y, and C Jason Woodard. 2008. "The Architecture of Platforms: A Unified View." http://hbswk.hbs.edu/item/6025.html.

Bensinger, Greg. 2024. "Exclusive: Amazon Mulls $5 to $10 Monthly Price Tag for Unprofitable Alexa Service, AI Revamp." Reuters, 2024, sec. Tech-

nology. https://www.reuters.com/technology/amazon-mulls-5-10-monthly-price-tag-unprofitable-alexa-service-ai-revamp-2024-06-21/.

Bogner, Alexander, Beate Littig, and Wolfgang Menz. 2014. *Interviews mit Experten: Eine praxisorientierte Einführung*. Wiesbaden: Springer Fachmedien. https://doi.org/10.1007/978-3-531-19416-5.

Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things out: Classification and Its Consequences. Inside Technology.* Cambridge, Mass: MIT Press.

Burkhardt, Marcus, Daniela Van Geenen, Carolin Gerlitz, Sam Hind, Timo Kaerlein, Danny Lämmerhirt, and Axel Volmar. 2022. "Introduction." In *Media in Action*, edited by Marcus Burkhardt, Daniela Van Geenen, Carolin Gerlitz, Sam Hind, Timo Kaerlein, Danny Lämmerhirt, and Axel Volmar, 9–36. Bielefeld: transcript. https://doi.org/10.14361/9783839455616-001.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 1–12. https://doi.org/10.1177/2053951715622512.

Crawford, Kate, and Vladan Joler. 2018. "Anatomy of an AI System." 2018. http://www.anatomyof.ai.

Dijck, Jose van, Thomas Poell, and Martijn de Waal. 2018. *The Platform Society.* New York: Oxford University Press.

Dolata, Ulrich. 2019. "Privatization, Curation, Commodification: Commercial Platforms on the Internet." *Österreichische Zeitschrift für Soziologie* 44 (S1): 181–97. https://doi.org/10.1007/s11614-019-00353-4.

Dolata, Ulrich. 2024. "Industrieplattformen als Markt-, Produktions- und Innovationsflächen. Feldvermessungen und theoretisch-konzeptionelle Überlegungen." *Berliner Journal f ür Soziologie* 34 (2): 171–96. https://doi.org/10.1007/s11609-024-00526-3.

Dolata, Ulrich, and Jan-Felix Schrape. 2023. "Platform Companies on the Internet as a New Organizational Form. A Sociological Perspective." *Innovation: The European Journal of Social Science Research* March: 1–20. https://doi.org/10.1080/13511610.2023.2182217.

Edwards, Paul N. 2019. "Infrastructuration: On Habits, Norms and Routines as Elements of Infrastructure." In *Research in the Sociology of Organizations*, edited by Martin Kornberger, Geoffrey C. Bowker, Julia Elyachar, Andrea Mennicken, Peter Miller, Joanne Randa Nucho, and Neil Pollock, 355–66. Emerald Publishing Limited. https://doi.org/10.1108/S0733-558X20190000062022.

Edwards, Paul N. 2021. "Platforms Are Infrastructures on Fire." In *Your Computer Is on Fire*, edited by Thomas S. Mullaney, Benjamin Peters, Mar Hicks,

and Kavita Philip, 313–36. Cambridge and London: The MIT Press. https://doi.org/10.7551/mitpress/10993.003.0021.

Floridi, Luciano. 2023. "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models." *Philosophy & Technology* 36 (15). https://doi.org/10.1007/s13347-023-00621-y.

Fourcade, Marion, and Kieran Joseph Healy. 2024. *The Ordinal Society.* Cambridge and London: Harvard University Press.

Frenken, Koen, and Lea Fuenfschilling. 2020. "The Rise of Online Platforms and the Triumph of the Corporation." *Sociologica* 14 (3): 101–13. https://doi.org/10.6092/ISSN.1971-8853/11715.

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Cambridge: Polity Press.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Gillespie, Tarleton. 2024. "Generative AI and the Politics of Visibility." *Big Data & Society* 11 (2). https://doi.org/10.1177/20539517241252131.

Gorwa, Robert. 2019. "What Is Platform Governance?" *Information, Communication & Society* 22 (6): 854–71. https://doi.org/10.1080/1369118X.2019.1573914.

Goulden, Murray. 2019. "Delete the Family: Platform Families and the Colonisation of the Smart Home." *Information, Communication & Society* 24 (4): 1–18. https://doi.org/10.1080/1369118X.2019.1668454.

Habscheid, Stephan. 2022. "Socio-Technical Dialogue and Linguistic Interaction. Intelligent Personal Assistants (IPA) in the Private Home." *Sprache und Literatur* 51 (2): 167–96. https://doi.org/10.30965/25890859-05002020.

Hallinan, Blake, and Ted Striphas. 2016. "Recommended for You: The Netflix Prize and the Production of Algorithmic Culture." *New Media & Society* 18 (1): 117–37. https://doi.org/10.1177/1461444814538646.

Hector, Tim Moritz, and Christine Hrncal. 2024. "Sprachassistenzsysteme in der Interaktion." In *Handbuch Sprache und digitale Kommunikation*, edited by Jannis Androutsopoulos and Friedemann Vogel, 309–28. Berlin and Boston: de Gruyter. https://doi.org/10.1515/9783110744163-015.

Helfferich, Cornelia. 2019. "Leitfaden- und Experteninterviews." In *Handbuch Methoden der empirischen Sozialforschung*, edited by Nina Baur and Jörg Blasius, 669–86. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-21308-4_44.

Hind, Sam, Fernando N. van der Vlist, and Max Kanderske. 2024. "Challenges as Catalysts: How Waymo's Open Dataset Challenges Shape AI Development." *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-01927-x.

Jassy, Andy. 2024. "2023 Letter to Shareholders." https://s2.q4cdn.com/299287126/files/doc_financials/2024/ar/Amazon-com-Inc-2023-Shareholder-Letter.pdf.

Johnston, Michael, Cris Flagg, Anna Gottardi, Sattvik Sahai, Yao Lu, Samyuth Sagi, Luke Dai, et al. 2023. "Advancing Open Domain Dialog: The Fifth Alexa Prize SocialBot Grand Challenge." In *Alexa Prize SocialBot Grand Challenge 5 Proceedings*. https://www.amazon.science/alexa-prize/proceedings/advancing-open-domain-dialog-the-fifth-alexa-prize-socialbot-grand-challenge.

Khan, Lina, M. 2018. "Amazon—An Infrastructure Service and Its Challenge to Current Antitrust Law." In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, edited by Martin Moore and Damian Tambini, 98–129. New York, NY: Oxford University Press.

Kim, Eugene. 2022. "Amazon Is Gutting Its Voice Assistant, Alexa. Employees Describe a Division in Crisis and Huge Losses on 'a Wasted Opportunity.'" *Business Insider*, November 19, 2022. https://web.archive.org/web/20230106123404/https://www.businessinsider.com/amazon-alexa-job-layoffs-rise-and-fall-2022-11.

Kinsella, Bret. 2023. "Google Assistant and Alexa Are Both Getting Generative AI Makeovers." *Substack newsletter. Synthedia (blog)*, August 2, 2023. https://web.archive.org/web/20230803220610/https://synthedia.substack.com/p/google-assistant-and-alexa-are-both.

Krishnan, Arun. 2024. "Amazon showcases new customer experiences powered by generative AI at CES 2024." *alexa-blog*, January 9, 2024. https://developer.amazon.com/en-US/blogs/alexa/device-makers/2024/01/alexa-characteraiandsplash-ces-2024.html.

Kuckartz, Udo. 2014. *Qualitative Text Analysis: A Guide to Methods, Practice & Using Software*. Translated by Anne McWhertor. Los Angeles: SAGE.

Li, Xiaochang. 2023. "'There's No Data Like More Data': Automatic Speech Recognition and the Making of Algorithmic Culture." *Osiris* 38:165–82. https://doi.org/10.1086/725132.

Luitse, Dieuwertje. 2024. "Platform Power in AI: The Evolution of Cloud Infrastructures in the Political Economy of Artificial Intelligence." *Internet Policy Review* 13 (2). https://doi.org/10.14763/2024.2.1768.

Luitse, Dieuwertje, Tobias Blanke, and Thomas Poell. 2024. "AI Competitions as Infrastructures of Power in Medical Imaging." *Information, Communication & Society* https://doi.org/10.1080/1369118X.2024.2334393.

Marres, Noortje, and David Stark. 2020. "Put to the Test: For a New Sociology of Testing." *The British Journal of Sociology* 71 (3): 423–43. https://doi.org/10.1111/1468-4446.12746.

Minder, Bettina, Patricia Wolf, Matthias Baldauf, and Surabhi Verma. 2023. "Voice Assistants in Private Households: A Conceptual Framework for Future Research in an Interdisciplinary Field." *Humanities and Social Sciences Communications* 10 (1): 173. https://doi.org/10.1057/s41599-023-01615-z.

Mols, Anouk, Yijing Wang, and Jason Pridmore. 2021. "Household Intelligent Personal Assistants in the Netherlands: Exploring Privacy Concerns around Surveillance, Security, and Platforms." *Convergence* 28 (6): 1841–60. https://doi.org/10.1177/13548565211042234.

Orr, Will, and Edward B. Kang. 2024. "AI as a Sport: On the Competitive Epistemologies of Benchmarking." In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1875–84. Rio de Janeiro Brazil: ACM. https://doi.org/10.1145/3630106.3659012.

Phan, Thao. 2019. "Amazon Echo and the Aesthetics of Whiteness." *Catalyst: Feminism, Theory, Technoscience* 5 (1): 1–38. https://doi.org/10.28968/cftt.v5i1.29586.

Plantin, Jean-Christophe, Carl Lagoze, Paul N Edwards, and Christian Sandvig. 2018. "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook." *New Media & Society* 20 (1): 293–310. https://doi.org/10.1177/1461444816661553.

Pridmore, Jason, Michael Zimmer, Jessica Vitak, Anouk Mols, Daniel Trottier, Priya C. Kumar, and Yuting Liao. 2019. "Intelligent Personal Assistants and the Intercultural Negotiations of Dataveillance in Platformed Households." *Surveillance & Society* 17 (1/2): 125–31. https://doi.org/10.24908/ss.v17i1/2.12936.

Reeves, Stuart, Joel E Fischer, Martin Porcheron, and Rein Sikveland. 2019. "Learning How to Talk: Co-Producing Action with and around Voice Agents." In *Proceedings of the "Mensch und Computer" 2019 Workshop on Interacting with Robots and Virtual Agents*, 362–63. München. https://doi.org/10.18420/muc2019-ws-654.

Rillig, Matthias C., Marlene Ågerstrand, Mohan Bi, Kenneth A. Gould, and Uli Sauerland. 2023. "Risks and Benefits of Large Language Models for the En-

vironment." *Environmental Science & Technology* 57 (9): 3464–66. https://doi.org/10.1021/acs.est.3c01106.

Rowberry, Simon Peter. 2022. *Four Shades of Gray: The Amazon Kindle Platform. Platform Studies.* Cambridge, Massachusetts: The MIT Press.

Sadowski, Jathan, Yolande Strengers, and Jenny Kennedy. 2021. "More Work for Big Mother: Revaluing Care and Control in Smart Homes." *Environment and Planning A: Economy and Space* 56 (1): 330–45. https://doi.org/10.1177/0308518X211022366.

Seaver, Nick. 2022. *Computing Taste: Algorithms and the Makers of Music Recommendation.* Chicago: University of Chicago Press.

Self, Becky. 2021. "Conducting Interviews during the COVID-19 Pandemic and Beyond." Translated by Alexander Ryazantsev. *Inter* 13 (4): 9–27. https://doi.org/10.19181/inter.2021.13.4.1.

Shi, Hangjie, Leslie Ball, Govind Thattai, Desheng Zhang, Lucy Hu, Qiaozi (QZ) Gao, Suhaila Shakiah, et al. 2023. "Alexa, Play with Robot: Introducing the First Alexa Prize SimBot Challenge on Embodied AI." In *Alexa Prize SimBot Challenge Proceedings*. https://www.amazon.science/alexa-prize/proceedings/alexa-play-with-robot-introducing-the-first-alexa-prize-simbot-challenge-on-embodied-ai.

Slota, Steven C, and Geoffrey C Bowker. 2017. "How Infrastructures Matter." In *The Handbook of Science and Technology Studies*, edited by Ulrike Felt, Rayvon Fouché, Clark A. Miller, and Laurel Smith-Doerr, Fourth edition, 529–54. Cambridge, Massachusetts: The MIT Press.

Soffer, Oren. 2020. "From Textual Orality to Oral Textuality: The Case of Voice Queries." *Convergence* 26 (4): 927–41. https://doi.org/10.1177/1354856519825773.

Srnicek, Nick. 2022. "Data, Compute, Labor." In *Digital Work in the Planetary Market*, edited by Mark Graham and Fabian Ferrari, 241–61. The MIT Press. https://doi.org/10.7551/mitpress/13835.001.0001.

Star, Susan Leigh, and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces." *Information Systems Research* 7 (1): 111–34. https://doi.org/10.1287/isre.7.1.111.

Starr, Paul, and Edward P. Freeland. 2024. "'People of Color' as a Category and Identity in the United States." *Journal of Ethnic and Migration Studies* 50 (1): 47–67. https://doi.org/10.1080/1369183X.2023.2183929.

Strengers, Yolande, and Jenny Kennedy. 2020. *The smart wife: why Siri, Alexa, and other smart home devices need a feminist reboot.* Cambridge, Massachusetts: The MIT Press.

Strüver, Niklas. 2023a. "Frustration Free: How Alexa Orchestrates the Development of the Smart Home." *Digital Culture & Society* 9 (1): 99–124. https://doi.org/10.14361/dcs-2023-090106.

Strüver, Niklas. 2023b. "Wieso eigentlich Alexa? Konzeptualisierung eines Sprachassistenten als Infrastruktur und Plattform im soziotechnischen Ökosystem Amazons." *kommunikation@gesellschaft* 24 (1): 1–33. https://doi.org/10.15460/kommges.2023.24.1.1194.

Tekic, Zeljko, and Johann Füller. 2023. "Managing Innovation in the Era of AI." *Technology in Society* 73 (May):102254. https://doi.org/10.1016/j.techsoc.2023.102254.

Tiwana, Amrit. 2014. *Platform Ecosystems: Aligning Architecture, Governance, and Strategy.* Amsterdam Waltham, MA: MK.

Van der Vlist, Fernando Nathaniël. 2022. *The Platform as Ecosystem: Configurations and Dynamics of Governance and Power.* Utrecht University. https://doi.org/10.33540/1284.

Van der Vlist, Fernando Nathaniël, Anne Helmond, and Fabian Ferrari. 2024. "Big AI: Cloud Infrastructure Dependence and the Industrialisation of Artificial Intelligence." *Big Data & Society* 11 (1). https://doi.org/10.1177/20539517241232630.

Vannuccini, Simone, and Ekaterina Prytkova. 2024. "Artificial Intelligence's New Clothes? A System Technology Perspective." *Journal of Information Technology* 39 (2): 317–38. https://doi.org/10.1177/02683962231197824.

Weidmüller, Lisa, Katrin Etzrodt, and Sven Engesser. 2022. "Trustworthiness of Voice-Based Assistants: Integrating Interlocutor and Intermediary Predictors." *Publizistik* 67 (4): 625–51. https://doi.org/10.1007/s11616-022-00763-7.

Weigel, Moira. 2023. "Amazon's Trickle-Down Monopoly: Third-Party Sellers and the Transformation of Small Business." Data & Society Research Institute. https://doi.org/10.2139/ssrn.4317167.

West, Emily. 2022. *Buy Now: How Amazon Branded Convenience and Normalized Monopoly. Distribution Matters.* Cambridge: The MIT Press.