Scaling AI Ventures: How to Navigate Tensions between Automation and Augmentation

Jonathan Zebhauser
Freie Universität Berlin
j.zebhauser@fu-berlin.de

Hannes Rothe
ICN Business School
hannes.rothe@icn-artem.com

Janina Sundermeier
Freie Universität Berlin
janina.sundermeier@fu-berlin.de

Abstract

AI ventures promise to automate and augment ever more human tasks. This provides rich opportunities for growth. Yet, digital and human resources that involve AI are oftentimes task-specific and hard to scale. Furthermore, clients remain skeptical to be fully automated by external services. Thus, it remains unclear how AI ventures achieve growth. We adopt a grounded theory approach on an interview study with founders, product managers and investors to inquire how resources afford or constrain scaling in AI ventures. For this, we blend the notion of (non-)scale free resources with the layered architecture of digital technologies. Our study suggests that AI ventures scale by organizing digital and human resources for replicability in that they keep AI-specific resources distant from clients while simultaneously externalizing human-intensive tasks to their clients. As we inquire the roles of human and digital resources, our study suggests that ventures seek to quickly find an optimal degree on the continuum between augmentation and automation when bundling resources.

Keywords: AI Venture, Scaling, Grounded Theory Method, AI Startup, Scale Free Resources.

1. Introduction

Artificial intelligence (AI) marks a new era of information systems management and implies interacting with an ever-evolving frontier of technological advancements in the context of decision making (Berente et al., 2021). AI seemingly gets more and more integrated into our societies (Rahwan et al., 2019) and greater numbers of ventures consider AI essential for their products and services (Weber et al., 2021). We refer to these types of ventures as AI ventures. AI ventures produce market offerings that change the value is being created and captured (Bughin et al., 2018; Chui et al., 2018; Fontana, 2021; Iansiti & Lakhani, 2020), because they promise to automate ever more tasks done by humans or at least augment humans in making better decisions (Raisch & Krakowski, 2021). Together, automation and augmentation provide ample opportunities for venture growth if being able to scale their available resources. We refer to *growth* as the change in a relevant measure of firm size and *scaling* as the relationship between multiple measures of size, e.g., available financial, human or digital resources (Schulte-Althoff et al., 2021; West, 2017). Similar to other digital ventures, AI ventures draw from the advantages of increased modularity, flexibility, and malleability of digital infrastructures (Henfridsson, 2020; Henfridsson et al., 2014; Yoo et al., 2012) for growth (Huang et al. 2017). However, AI ventures seem not to scale like other digital ventures as they leverage AI-specific resources, that differs how specific their digital resources are or entails people with different expertise (Casado & Bornstein, 2020; Chui & Malhotra, 2018; Giustiziero et al., 2021; Linde et al., 2020; Schulte-Althoff et al., 2021).

To understand the problem that AI ventures face when scaling, we adopt the notion of non scale-free resources, including human resources. That is, each additional unit implies an equal increase in costs (Burström et al., 2021; Khan et al., 2020; Sjödin et al., 2021). In fact, AI ventures are unique in how they combine human with digital resources, e.g., digital technology and data annotated with human input. This dependency on non scale-free resources hamper repeated value creation (Jöhnk et al., 2021) and hence, it has a decisive impact on the venture's scaling ambitions (Levinthal & Wu, 2010) as it poses a limit to growth (Penrose, 2009). For AI ventures it can take much longer to repeatedly capture value from the same bundles of digital resources, because new clients require investment of additional human resources for producing, annotating, integrating new data, or updating machine learning (ML) models (Casado & Bornstein, 2020). In addition, testing and monitoring AI applications requires more human oversight than testing and monitoring rule-based software, as it is difficult to specify data and ML model behavior a priori (Breck et al., 2017). As a consequence, recent findings indicate that the average AI venture shows a similar demand for human resources as service

URI: https://hdl.handle.net/10125/103322
978-0-9981331-6-4
(CC BY-NC-ND 4.0)

ventures when growing (Schulte-Althoff et al., 2021). This is surprising in that it counters the intuition that augmentation and automation with AI should support scaling. We therefore ask: *How do AI ventures organize digital and human resources for scaling? How do they repeatedly and at greater pace create and capture value from these resources?* To answer our research questions and to support building theory around this phenomenon, we conduct a qualitative research study. We use a modified grounded theory method (GTM) approach study following Gioia et al. (2013) to build theory from practice. For this, we draw from experiences of experts in scaling AI ventures. Our qualitative study is based on twelve expert interviews that include diverging views of founders, product and business developers and investors, all in the context of AI ventures. During our analysis, we learned that AI ventures organize digital and human resources differently, depending whether these resources relate to the content, service, network, or device layer of the digital infrastructure (Yoo et al., 2010). We present our findings in three aggregated dimensions and four second-order themes. Our research indicates that AI ventures scale by organizing digital and human resources for replicability and far away from the client while simultaneously externalizing human-intensive tasks to their clients. In order to become better in this repetition, ventures build a supportive digital infrastructure around externalized tasks, such as data annotation. At the same time, we learn that clients shy away from being automated by products and services of AI ventures. We thereby show that AI ventures face the unique problem of finding an optimal degree in the continuum between automation and augmentation. AI ventures seek to maximize the use scale-free resources, which increases automation while also following their client's call for augmentation. The study contributes to theoretical and practical discourse about understanding mechanisms behind scalable AI ventures as well as by blending the notion of scale free resources and non-scale free resources with the layered architecture of digital technology to offer further perspectives on AI ventures and scaling.

2. Conceptual Background

2.1 AI Ventures

AI can be described as an ever-evolving frontier of computational advancements in the context of decision-making problems (Berente et al., 2021). It enables ventures to create new products, business models and services (Brynjolfsson & Mcafee, 2017; Makridakis, 2017). Industry-agnostic, AI technology could have an impact in all industries and sectors. Therefor it can be understood as a general purpose technology (Brynjolfsson & Mcafee, 2017; Trajtenberg, 2018). Following Penrose' understanding of firms (Penrose, 2009), ventures are a set of resource bundles that differ in their ability to produce scale. AI ventures are built on digital and human resources. Digital resources entail software code and digital data. They can be AI-specific, e.g., annotated data, or software code that enables machine learning algorithms. Also human resources can have AI-specific roles, e.g., data scientists, machine learning engineers, data product managers, machine learning researchers which can be summarized as AI experts. Combining these resources is an important aspect of applying AI. The "*human in the loop*", for instance, is supposed to evaluate, interfere or tune decisions enshrined in software code of AI algorithms. Humans manage date in that they annotate, integrate and maintain digital data that is supposed to feed machine learning. Besides these AI specific resources there are complementary non-AI specific resources necessary such as cloud infrastructure, software engineers, or product designers (Fontana, 2021; Metelskaia et al., 2018). Ventures differ in how they combine these resources (Weber et al., 2021) in that they, for instance, combine pre-trained machine learning models with software code for a mobile application to address demands of multiple customers. Similarly, ventures may also produce custom data and software code to create unique AI algorithms within custom service relationships. These different combinations might have different effects on a venture's ability to scale.

2.2. Scaling AI Ventures

Iansiti and Lakhani (2020) propose the notion that AI allows ventures to create and capture value through three mechanisms: first, through the repetition of value creation within a domain, second, through the repetition of value creation between domains and third, through attaining the resources for repetitive value creation and capture. To understand scaling in our context, we first look at digital ventures, of which AI ventures form a subset. Digital ventures draw on digital infrastructure, which enables them to execute their actions on a given structure (Henfridsson, 2020). This setting facilitates two scaling modes: scale due to design flexibility and scale due to design scalability (Henfridsson et al., 2014). Digital ventures have a flexible design that allows rapid reaction to changing circumstances because they use a digital infrastructure that is not pre-defined (Henfridsson et al., 2014). Following the work done by Yoo et al. (2010), we

understand the architecture of the digital firm as a layered modular architecture. It consists of four layers: contents-, service-, device-, and network layer. The contents layer contains the data, while the service layer represents the application functionality that serves the user. The network layer is divided into a physical transport layer including hardware and a logical transmission layer which includes protocols or network standards. The device layer is also divided into a physical machinery layer consisting of e.g., computer hardware, and a logical capability layer that consists of e.g., an operating system. The layered modular architecture combines the modular architecture and the described layered architecture. Modular architecture is structured through standardized interfaces between components - the highest degree of modularity makes these components product-agnostic (Yoo et al., 2010). Digital ventures have a scalable design because of low costs of replication of digital resources (Giustiziero et al., 2021; Shapiro et al., 1998). As scalability is achievable for most digital ventures, scaling faster than competitors is important, especially when it comes to winner-take-all markets (Cohen & Levinthal, 1990; Schilling, 2002). In conclusion, time moderates scaling activities and their success in respect of competing ventures. Studying a credit rating company that was scaling on a rapid pace, Huang et al. (2017) traced three mechanisms that enable rapid scaling, which are data-*driven operation*, *instant release*, and *swift transformation*. Data-driven operation enables framing, hedging and monitoring of opportunities and activities of innovation with data. Instant release enables fast deployment of innovation ideas, only with a short time-lag. Swift transformation enables the effortless contextualization towards new value-in-use aligned with an updated venturing identity (Huang et al., 2017). Through these mechanisms the speed of scaling operations can be increased, drawing on productive techniques at a high pace (Henfridsson, 2020). While they highlight opportunities, challenges for other ventures adopting these strategies remain unclear.

2.3. Scale free and non-scale free resources

Ventures, considered as bundles of resources (Penrose, 2009), depend on characteristics of these resources especially when it comes to scaling. Some resources, like brand names, scale almost without boundaries in that they can be replicated across many domains (scale-free), other resources can not be easily replicated and therefore produce limited scale (non scale-free) (Levinthal & Wu, 2010). The application of such non-scale free resources depends on the opportunity costs of deploying them in another domain (Levinthal & Wu, 2010). Thus, ventures have to assess the utility of deploying resources in one domain compared to another, especially when considering to offer products across market segments (Levinthal & Wu, 2010) or when testing different markets. This is important, because ventures are initially unaware of the market segments in that they are eventually able to scale (Giustiziero et al., 2021). This implies that when a resource bundle is scalable, the opportunity cost of deploying these resources anywhere else than in the ventures focal domain is high, as it contains complementary resources in form of specialized human and managerial resources. Digital resources are supposedly scale-free as they can be replicated almost error-free, being globally distributed at low costs and steadily improved in performance and costs as more they are used (Agrawal et al., 2018; Brynjolfsson & McAfee, 2014; Giustiziero et al., 2021). More specialized resources need to be managed with opportunity costs in mind. Human resources cannot simultaneously develop software code for a generic product while also conducting service for a specific client project. This is especially important the more unique, rare, and highly regarded these human resources are. AI experts fall into that category due to their level of training in software development, data analysis, statistics, and/or management training, which might put a unique burden on AI ventures.

3. Methodology

To understand how AI ventures organize their digital and human resources for scale and replicable value creation and capture, we used a qualitative research approach. Our research design follows a modified grounded theory methodology (GTM) approach (Corbin & Strauss, 2015; Gioia et al., 2013) in the process of analysis, which is suited to understand and explore an IS-related phenomenon in a complex environment (Wiesche et al., 2017). Our approach to data collection differs from GTM: We prepared for an empirical exploration of our research field using extant theory as Goldkuhl and Cronholm (2019) propose.

Due to the evolving frontier of AI and its implications on business we saw the need to inform ourselves to be able to reach an insightful level in the conversation with experts. Authors we draw from regarding the topics AI, constraints of scaling and current forms of value creation and capture in AI ventures are, among others, Anderson and Tushman

ID	Function	Stage of Venture	Business model typology that covers divergent resource bundles	Market	Interview Duration	AI Experience
ID1	Founder	Grown Venture	AI-charged Product/Service Provider	B2B	31:12 min	Over 10 years
ID2	Founder	Grown Venture	Data Analytics Provider	B2B	76:57 min	Over 20 years
ID3	Managing Director	Early Stage Venture	AI Development Facilitator	B2B	49:00 min	Over 5 years
ID4	Founder	Mid Stage Venture	Deep Tech Researcher	B2B	22:55 min	Over 5 years
ID5	Product Manager	Grown Venture	Data Analytics Provider	B2B	73:27 min	Over 10 years
ID6	Founder	Early Stage Venture	AI-charged Product/Service Provider	B2C	39:29 min	Over 2 years
ID7	Investor	Series A Startups	-	All	63:56 min	Over 5 years
ID8	Founder	Mid Stage Venture	Data Analytics Provider	B2B	33:22 min	Over 10 years
ID9	Business Developer	AI Venture Builder	All Patterns	All	46:29 min	Over 2 years
ID10	Product Manager	Mid Stage Venture	AI Development Facilitator	B2B	35:49 min	Over 7 years
ID11	Investor	Series A Startups	-	All	36:58 min	Over 7 years
ID12	Product Manager	Mid Stage Venture	AI Development Facilitator	B2B	18:36 min	Over 7 years

Figure 1. Data Structure.

(1990), Casado and Bornstein (2020), Fontana (2021), Giustiziero et al. (2021), Iansiti and Lakhani (2020), and Schulte-Althoff et al. (2021). For the analysis we adopted the approach of Gioia et al. (2013). In the first-order analysis we identified and used empirical codes and terms that seemed central to the interviewees. In the second-order analysis we identified theoretical concepts related to our empirical observations, before finally turning to further abstraction in aggregate dimensions. Interviews are a common method for data collection in GTM studies (Corbin & Strauss, 2015; Gioia et al., 2013). To capture individuals' experiences and perspectives framed by our research focus, we chose to conduct semi-structured expert interviews. To visualize our findings, we followed the approach of Gioia et al. (2013).

3.1. Data Collection

Our sampling covers different levels of insights into how resources afford or constrain scaling in AI ventures. We expected founders to reflect on the impact of initial resources in their ventures as well as their strategic development. Product and business developers were supposed to provide insights into operating digital resources in grown ventures. Investors active in multiple AI ventures were supposed to provide a perspective of a well-informed third-party. Our sample was supposed to include ventures incorporating value creation and value capture from divergent resource bundles. We used the business model typology from Weber et al. (2021) to operationalize these divergency in our sample. As researching AI ventures in general is a high level view on a field with great differences regarding sector and customer dynamics, we included experiences with different industries, B2B and B2C markets as well as different regional focuses (US, India, EU). To keep the exploration space open-ended we used our first collection of first-order categories. To derive labels for our first-order categories, we used abstraction. Following a first iteration of open coding, we switched to axial coding. Here, we found that a conceptual framework could help us structure the impact of different digital resources on an AI ventures' ability to scale. We chose the concept of a layered modular architecture (Yoo et al., 2010) and its corresponding layers to link our findings to the digital infrastructure of an AI venture, as described in the conceptual background. This iteration helped us sort first-order categories and organize them into second-order themes. While the first-order concepts are still close to our raw data, the second order themes have a strategical quality, inspired by Huang et al. (2022). The aggregated dimensions help us embed our findings in theory and therefor have a theoretical quality. Using layered modular architecture as a conceptual framework, we saw that keeping it as a part in the aggregated dimensions is useful and helps to transmit our theoretical contribution.

4. Results

Following Gioia et al. (2013) we present our first- and second-order Figure 1. As mentioned before, the layered modular architecture (Yoo et al., 2010) revealed itself as a scaffolding to sort our findings as we went back and forth between theory and qualitative data. In the following, we explain each dimension according to his scaffolding.

4.1. Contents layer organized for scale

An important means to create scale free resource bundles for AI ventures involves relationships with

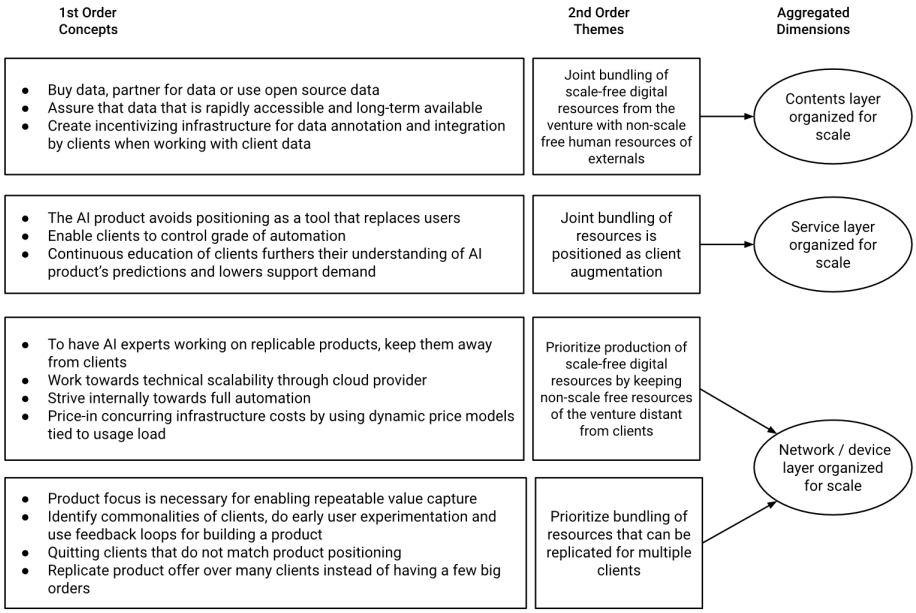

Figure 1. Data Structure.

clients on the contents layer. The corresponding first order concepts are displayed in figure 1.

Joint bundling of scale-free digital resources from the venture with non-scale free human resources of externals: Value creation through AI depends heavily on data. Preparing, integrating and monitoring new data tie resources. Scale at rapid pace can only be achieved if these tasks are whether highly efficient or externalized. Data sources vary depending on the market segment the venture operates in. Our findings show that the source of data, its accessibility and availability play an important role if aiming for scale as a venture. If data sources are diversified and its management cannot be fulfilled internally without loosing resources, an infrastructure for externalizing the corresponding tasks has to be created.

First, to organize such scale free data sources, three ways of acquiring data were highlighted by most interviewed experts: buy data, partner for data or use open-source data. Speaking about an AI venture working in health-sector, expert 10 states: "*We either buy data, create partnerships or find ways that we can collect data. [...] Sometimes the health systems have let us build up a data set based on the product that we were building in pilot with them*". In other market segments, reliable open-source data may be the way to go, as expert 11 explains: "*But most either buy datasets or use existing open-source resources that are already there. We also see this very often, especially for example in the Geo-Spatial area [...]*".

Second, most of the interviewed experts pointed to the importance of rapid and long-term access to the data which the product gets built on. Binding data through contracts is an option, expert 8 recommends: "*You tell them: If you want to use this technology, you have to give us the data […] And then they either do it or they don't, and that's why the contract is concluded or not*". As expert 11 states, that this is also important when selecting clients: "*It must be ensured that this is not just a one-time customer, but someone who can imagine working with us for many years, because this is the only way to ensure financial sustainability and data volume*". Thus, assuring rapid and long-term availability of the data used by the AI venture is one aspect to consider when building a scale free contents layer.

Third, most interviewed ventures enable and incentivize their clients to annotate data themselves in that they produced tools to make these tasks easy, e.g., building intuitive interfaces or creating "*no-code*" applications that required little prerequisite knowledge in data management. Also, this environment should be incentivizing, so that clients annotate the data and integrating it by themselves. Most interviewees stated that simplifying the data integration process for the users is important. Expert 5 states: *"[...] you have to think of getting to a self-serve kind of model [...] that is the only way to scale your company [...] The trick is in figuring out incentivizing your users.*". In B2B cases some client data may already have been annotated, e.g., stocktaking. Therefore, ventures found that data integration could be automatized to a high degree - one

of the ventures (ID8) lowered the data integration of new B2B clients to approx. two hours. Therefore, this venture is able to deploy their own resources for the process. The same venture uses CAD models of clients to create synthetic visual data with an automatized process, resulting in even higher annotation accuracy. One venture (ID2) we spoke with gives clients access to APIs which the clients then have to use. In that way, clients are forced to handle data management on their own - still, this venture offers support which earlier was done by internal experts and now is handled through partnerships with consultancies, another way to keep the contents layer scale free.

Our conclusion is that ventures facilitate a scale free contents layer in their digital infrastructure that enables clients to annotate, integrate and manage data. Thereby, clients remain inherently involved in the AI application which, even if supposed to be highly automatic, thereby still offers augmentation functions. While in a lot of domains building a scale free resource bundle for full automation may be difficult to achieve, an alternative is figuring out a replicable process for externalizing these tasks to the client. As data quality and data distribution influence the scale and cost of the value creation, our interviewees' experience shows that non-experts working on data had to be educated and monitored by experts in order to ensure that the venture can still replicate their use of digital resource.

4.2. Service layer organized for scale

Joint bundling of resources is positioned as client augmentation: Scaling value creation requires repeatedly serving clients in similar ways. Most of the interviewees learned that clients, however, do not want to be "*automated*", as, e.g., expert 4 describes: *"[...] there were always discrepancies between the management, which had the pressure to become more productive, and the actual researchers [users on client-side], who said, 'Well, if we think this through to the end, then you don't need me anymore.'*". Thus, ventures need to walk the fine line, replicating as many resources as they can in order to scale, while also not losing the client's trust to not automate tasks that are typically done by human users. Most interviewees mention that offering education to foster understanding of the AI technology in use on client side facilitates value capture at greater pace by minimizing support interventions. Through delivering products that clients perceive as "augmenting", the fear of being automated by AI can be addressed. We use the term "*service layer organized for scale*" to denote that the resource bundle involved in the service layer should be organized to be scale free. This might ask for specific strategies, as clients might want to be "in the loop" and not fully automatized.

First, interviewees mention, that the fear of being replaced by AI often accompanies stakeholders. Therefor ventures may position their offer as augmentative instead of replacing, as expert 5 states: „*The biggest challenge with AI-Systems has been that it can replace the people who are potentially going to adopt it.*". Remaining in the loop rather than relying on full automation keeps the decision-making authority on how the AI behaves at least partly in the customer's perceived sphere of influence.

Second, with the goal of reaching a high degree of automation as well as a customer/product interaction that requires as little support and intervention from the venture as possible, some of the companies we interviewed enable their clients to control grade of automation. Expert 2 explains that they design their products interface in a way that the customer has both the ability to intervene in the automated result but also learns over time to interfere less, as it generally leads to less optimal results. Expert 12 reports that they work with a continuum that allows the user to determine the level of automation provided by the application. In their experience, this has always led to decreasing intervention by customers, as they increasingly trust the automation and the motivation to intervene themselves decreases as a result. Enabling clients to control the grade of automation by deciding when and how often to intervene, while educating transparently about the efficiency of the automation and comparing results of decisions made by users and the product, ventures may reduce the resources tied to client intervention and reach a higher degree of automation on the client side.

Third, concepts such as statistical uncertainty are hard to grasp. Lack of knowledge on how AI works can lead to clients escalating more often, as expert 2 describes: "*Escalation occurs and we realize that nothing is going on. They just didn't understand that there are always statistical uncertainties.*". Expert 5 adds: "*So like most of these technical products usually need a good hand-holding with go to market teams and educating your customers on how to use your products.*". Client escalation ties up support resources and reduces the pace of value capture. To allow for uninterrupted value capture ventures educate their clients in concepts like prediction and uncertainty. Client education should be developed as a scale-free resource which scales with its demand. As this can be difficult to achieve within a venture, partnering with external consultants can be a solution. Our theoretical conclusion is that by offering augmentation while minimizing client input long term by education on the service layer can be organized for scale.

4.3. Network / device layer organized for scale

Prioritize production of scale-free digital resources by keeping non-scale free resources of the venture distant from clients: Ventures need an infrastructure that enables the creation of replicable resource bundles. Thus, the resource bundle involved in the network and the device layer and its corresponding activities have to be organized to be scale free. First, we learned that in order to avoid solving unique problems of individual clients with custom solutions, AI experts were kept at distant from the client. This enabled them to fully focus on building scale free digital resources, as expert 1 describes: "*the exciting question is, do you get the people who are, I'll call them AI experts, do you get them decoupled enough from the respective customers that they actually build, I'll say standard products?".* Also, expert 2, the most senior AI expert we spoke to, explains: "*As a scientist, of course, I found it great when we had a lot of different applications based on the same core idea.".* Interestingly, we noticed that this problem was perceived by our informants as both sided: clients were considered to ask for custom solutions that meet their particular needs, while AI experts seem to have a strong emphasis on solving hard and particularly new problems and therefor are keen to develop as much custom solutions as they can. Thus, keeping AI experts at some distance to client conversations and their inquiries, lowered the chances of producing custom solutions.

Second, to scale fast all interviewed ventures rely on cloud technology at the network layer. While this can be costly, it allows for rapid scale and internationalization, even though data privacy preferences might enforce a shift to a localized cloud solution in some cases. Third, some of the experts interviewed clarify that the internal goal is full automation, which is often not equally perceived on client-side, as, e.g., expert 11 states: "*So, internally, automation is very hot. Externally, however, I don't think it's being sold quite as much yet.*". Expert 2 explains: "*We are in favor of fully automating this, of course, but many customers don't want that.*". Designing for a high grade of automation enables strong replicability on the network and device layer.

Fourth, to assure scalable linkages to resources provided by partners such as cloud providers, a scalable price model has to be established. Expert 2 explains their use of: "*[...] a model where we define a price ladder or a range and also a scaling law for a certain product based on our experience and value and effort estimation[...]*". Also, expert 5 underlines the importance of measuring costs continuously to find the right margins to charge. Expert 11 explains the two price models she observes in their portfolio of AI ventures: Software as a Service, or pricing by consumption, e.g., data processing fees. Our conclusion is that ventures should create an infrastructure that bundles non-scale free resources, especially AI experts, with digital resources on the network and device layer. This enables focal work on replicable products which are then scale free.

Prioritize bundling of resources that can be replicated for multiple clients: First, repeatedly capturing value is important for ventures who seek to grow quickly, which is why producing a resource bundle on that can be replicated easily becomes important for AI ventures. Producing replicable bundles of digital resources from the network and device layer becomes important for AI ventures to avoid becoming an agency for clients that creates one-time payments and project-based revenue. Consulting other AI ventures, expert 1 explains this as a common hurdle that has to be avoided: *"[...] I'm really just sort of always asking: What are you doing to not become an agency?*". Expert 2 explains this with their own story: "*At the very beginning [...] this was still different for customer A from customer B, but we really got to the point where they were all exactly the same [...] before [...] we could essentially only scale by hiring even more data scientists*".

Second, in order to prioritize resource bundles that can be replicated, the AI venture needs to find commonalities between clients. Expert 7, an investor, highlighted that she uses such commonalities between clients as an important indicator for assessing potentials for growth in AI ventures. In an early stage, finding commonalities involves experimentation. These experiments involve engagements with individual clients that help produce digital resources that also serve other clients. This may translate to less current revenue for the venture, as interviewees underline.

Third, quitting clients that require digital resources that cannot be replicated in other client engagements is vital, this involves new sets of data, other expertise of AI experts, or diverging ML models. Thus, keeping various clients at the same time seems important. Interviewees mention decisions might be tough at an early stage, but is the only way to organize the venture for scale in the long term. Expert 2 states: "*We also had to force a few customers: You have to take our product now; we won't develop the other one further*". Expert 8 explains how they solve inquiries for specifications that do not match their offer: *"[Sometimes] we say: You guys, we can't offer that, but we have a development partner and he does it at good conditions*".

Fourth, while big orders by clients are an opportunity for capturing value short, they may lead

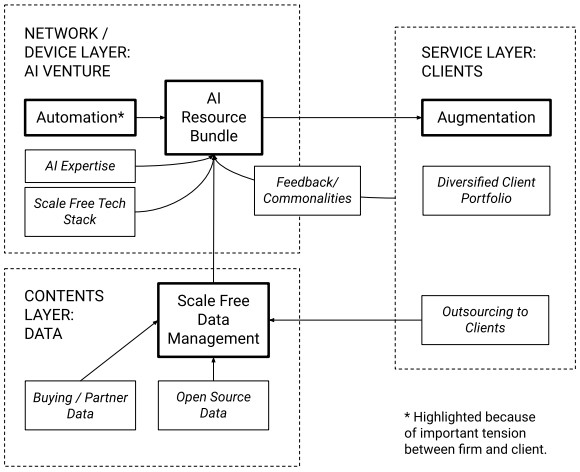

Figure 2. Layers Built for Scale in an AI Venture.

into the "*agency trap*": being dependent on only one or a few clients positions the venture as an agency rather than a product provider. Both expert 1 and 2 emphasizes how a product focus protected them from getting individual major customers for which they would have become "*an internal IT shop*".

5. Discussion

Our study identified three key dimensions that AI ventures may address when striving for the creation of scale free resource bundles, which is important because these bundles allow AI ventures to attain growth. Our work contributes to the ongoing discourse on AI and scaling in three major ways.

First, our qualitative inquiry reveals how ventures tackle a unique tension between automation and augmentation, i.e., deciding to what extent they include human and digital resources into their resources bundles. Second, we found that scaling depends largely on a venture's ability to replicate resource bundling which requires important decisions on securing rare (human) resources on the service and network layer.

Third, we find that resources on the contents layer is an important bottleneck for scaling which ventures solve by introducing means to externalize data. In order to scale, AI ventures are forced to organize their digital and human resources in bundles. At the same time human and managerial resources, which are generally not scale free, are deployed on occasion only and not as permanent actors in these processes. Figure 2 summarizes our core findings.

Tension between Automation and Augmentation: Ventures aim to maximize replication of resource bundles that could be sold to clients. Producing AI-specific digital resources, such as ML models or data, that allow to ventures to increase automation is a corner stone of this endeavor. Clients, however, demand continuous adjustments of resources bundles to their specific needs and therefore would like to retain "humans in the loop". This reveals a tension between automation and augmentation that is generally acknowledged (Raisch & Krakowski, 2021), but seems unique for AI ventures in that these firms need to scale. AI ventures have to find an optimal point in the continuum between automation and augmentation. Larger companies can afford to develop AI produce AI-based resource bundles that involve humans (i.e., augmentation) order to slowly produce more and more data, expertise in creating algorithms, and build client trust so that they later introduce resource bundles that do not include human resources (i.e., automation) (Raisch & Krakowski, 2021). Digital ventures, however, are supposed to scale rapidly and early on (Huang et al., 2017). Thus, while larger firms can gradually produce these bundles (first augmenting, later automating), AI ventures strive for replication early on. Our study exemplifies that ventures learned how include clients "in the loop" who effectively ask for "augmentation", while not falling into an "agency trap". This shows, how AI ventures face the inherent struggle of creating AI-specific resource bundles that need to be adjusted to client demands while avoiding to suffer from reduced scale. One way how AI ventures solve that issue is by creating tools that allow for the use of scale-free resources on the contents layer, i.e., externalizing data annotation. At the same time, venture keep their own non-scale free resources, especially AI experts, distant from the client in order to avoid falling into an "agency trap", i.e., becoming an AI agency that produces only custom AI applications. The tension is depicted in Figure 2.

Replication of Resource Bundling: AI ventures deploy their human resources to produce a core resource bundle that serves a broad market segment and enforce its adaption by denying adaptation of these resources that would serve a diversified portfolio of clients. This way, AI ventures avoid becoming agencies for the client. Instead, they let feedback and user research iteratively shape a replicable product based on commonalities, as depicted in figure 2, while not sharing their AI expertise with the clients. By focusing on a particular resource bundle, the AI ventures more quickly produces resources, e.g., algorithms, human knowledge, or data, that improves the quality of this resource bundle. Human and managerial resources are constantly being made available for this purpose. A scale free tech stack based on state-of-the-art methods such as modularity and cloud infrastructure help to build a replicable resource bundle. While these findings

reiterate how digital ventures learn from client data in order to scale (Huang et al., 2017), it also exemplifies how AI ventures need to balance their willingness to create a scalable resource bundle with the necessity to hide its inner workings from the client. It might therefore explain, why AI ventures scale more comparable to service ventures than digital platforms (Schulte-Althoff et al., 2021).

Contents Layer organized for Scale: We learned that clients were hesitant to be fully automated by an AI venture and rather ask for being augmented, because they can retain some control over the impact of the AI. Enabling human resources of a client to be bundled with digital resources of the AI venture therefore plays two parts. It allows augmentation but also plays into the hands of AI ventures who seek to delimit the use of their own human resources to avoid opportunity costs (Penrose, 2009). This is especially important in cases where ventures work with client data, because data requires much human labor, e.g., for maintenance and labeling. Data is important for many AI-based resources, e.g., ML models (Fontana, 2021). Open data oftentimes does not suffice for producing a purchasable resource bundle. Instead, AI ventures may need to create proprietary data that would be unique for every client relationship. In order to avoid engaging in forming unique resource bundles for every client relationship, AI ventures strive to produce replicability on the contents layer by producing software and standardizing (meta) data, e.g., for labeling. If a venture, that offers an API to their AI for their clients adds hundred clients overnight, data annotation and integration might pose a bottleneck. If the infrastructure allows clients, however, to annotate and integrate the data themselves and with little human support, hundreds of new clients can be on-boarded overnight. This creates synergies between the clients demand for being augmented, their ability to monitor effects of the AI-based resource bundle, and the ventures demand for externalizing parts of the data management: both activities call for educational measures and for harmonization of content and service layer that integrates well with a broad set of clients.

6. Conclusion

Following Gioia et al. (2013) we derived four strategic implications for scaling AI ventures in form of second order themes. Thus, our results provide valuable information for organizing resources in AI ventures for scale. Considering the limitations of our study, opportunities for further research arise: Our findings are not statistically generalizable, as a qualitative approach as ours only aims to gain deep insight into phenomenons (Flick, 2013). AI itself is not one technology but consists of an ever-evolving frontier (Berente et al., 2021), such research only grasps a tiny part of the whole situated in a specific point of this frontier's evolution, which naturally poses a limitation to our research. While we carefully sampled our experts, future research could use a bigger sample size and further validate our findings.

Acknowledgement

We kindly thank K.I.E.Z. (Künstliche Intelligenz Entrepreurship Zentrum), funded by BMWK and EXIST, as well as the Digital Entrepreneurship Hub at Freie Universität Berlin, for enabling us to investigate this topic.

7. References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Press.

Anderson, P., & Tushman, M. L. (1990). Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative science quarterly*, 604–633.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Q*, 45(3), 1433–1450.

Bogner, A., Littig, B., & Menz, W. (2014). *Interviews mit experten: Eine praxisorientierte Einführung*. Springer-Verlag.

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. 2017 *IEEE International Conference on Big Data (Big Data)*, 1123–1132.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Brynjolfsson, E., & Mcafee, A. (2017). Artificial intelligence, for real. *Harvard Business Review*, 1, 1–31.

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. *McKinsey Global Institute*.

Burström, T., Parida, V., Lahti, T., & Wincent, J. (2021). AI-enabled business-model innovation and transformation in industrial ecosystems: A framework, model and outline for further research. *Journal of Business Research*, 127, 85–95.

Casado, M., & Bornstein, M. (2020). The New Business of AI (and How It's Different From Traditional Software). https://venturebeat. com/2020/02/22/why-ai-ompanies-dont-always-scale-like-traditional-software-startups/

Chui, M., & Malhotra, S. (2018). AI adoption advances, but foundational barriers remain. *Mckinsey and Company*.

Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., & Malhotra, S. (2018). Notes from the AI frontier: Insights from hundreds of use cases. *McKinsey Global Institute*.

Cohen,W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 128–152.

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Flanagan, J. C. (1954). The critical incident technique. *Psychological bulletin*, 51(4), 327.

Flick, U. (2013). *The Sage Handbook of Qualitative Data Analysis*. Sage.

Fontana, A. (2021). *The AI-first company: How to compete and win with artificial intelligence*. Portfolio/Penguin.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia Methodology. *Organizational research methods*, 16(1), 15–31.

Giustiziero, G., Kretschmer, T., Somaya, D., & Wu, B. (2021). Hyperspecialization and hyperscaling: A resource-based theory of the digital firm. *Strategic Management Journal*.

Goldkuhl, G., & Cronholm, S. (2019). Grounded Theory in Information Systems Research-from Themes in IS Discourse to Possible Developments. In *ICIS*.

Henfridsson, O. (2020). Scaling digital enterprises. *Handbook of Digital Innovation*. Edward Elgar Publishing.

Henfridsson, O., Mathiassen, L., & Svahn, F. (2014). Managing technological change in the digital age: The role of architectural frames. *Journal of Information Technology*, 29(1), 27–43.

Huang, J., Henfridsson, O., & Liu, M. J. (2022). Extending digital ventures through templating. *Information Systems Research*, 33(1), 285–310.

Huang, J., Henfridsson, O., Liu, M. J., & Newell, S. (2017). Growing on steroids: Rapidly scaling the user base of digital ventures through digital innovaton. *MIS Quarterly*, 41(1).

Iansiti, M., & Lakhani, K. R. (2020). *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world*. Harvard Business Press.

Jöhnk, J., Weißert, M., & Wyrtki, K. (2021). Ready or not, ai comes—an interview study of organizational ai readiness factors. *Business & Information Systems Engineering*, 63(1), 5–20.

Khan, N., McCarthy, B., & Pradhan, A. (2020). Executive's guide to developing AI at scale. https://www.mckinsey.com/business-functions/quantumblack/our-insights/executives-guide-to-developing-ai-at-scale#intro

Langley, A., & Meziani, N. (2020). Making interviews meaningful. *The Journal of Applied Behavioral Science*, 56(3), 370–391.

Levinthal, D. A., & Wu, B. (2010). Opportunity costs and non-scale free capabilities: Profit maximization, corporate scope, and profit margins. *Strategic Management Journal*, 31(7), 780–801.

Linde, L., Sjödin, D., Parida, V., & Gebauer, H. (2020). Evaluation of digital business model opportunities: A framework for avoiding digitalization traps. *Research-Technology Management*, 64(1), 43–53.

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Futures*, 90, 46–60.

Metelskaia, I., Ignatyeva, O., Denef, S., & Samsonowa, T. (2018). A business model template for ai solutions. *Proceedings of the International Conference on Intelligent Science and Technology*, 35–41.

Penrose, E. (2009). *The theory of the growth of the firm*. Oxford university press.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature*, 568(7753), 477–486.

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1), 192–210.

Schilling, M. A. (2002). Technology success and failure in winner-take-all markets: The impact of learning orientation, timing, and network externalities. *Academy of Management Journal*, 45(2), 387–398.

Schulte-Althoff, M., Fürstenau, D., & Lee, G. M. (2021). A scaling perspective on ai startups. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 6515.

Shapiro, C., Varian, H. R., Carl, S., et al. (1998). *Information rules: A strategic guide to the network economy*. Harvard Business Press.

Sjödin, D., Parida, V., Palmié, M., & Wincent, J. (2021). How ai capabilities enable business model innovation: Scaling ai through co-evolutionary processes and feedback loops. *Journal of Business Research*, 134, 574–587.

Trajtenberg, M. (2018). *AI as the next GPT: A political-economy perspective* (tech. rep.). National Bureau of Economic Research.

Weber, M., Beutter, M., Weking, J., Böhm, M., & Krcmar, H. (2021). AI Startup Business Models. *Business & Information Systems Engineering*, 1–19.

West, G. (2017). *Scale: The universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies*. Penguin Press.

Wiesche, M., Jurisch, M. C., Yetton, P. W., & Krcmar, H. (2017). Grounded theory methodology in information systems research. *MIS quarterly*, 41(3), 685–701.

Yoo, Y., Boland Jr, R. J., Lyytinen, K., & Majchrzak, A. (2012). Organizing for innovation in the digitized world. *Organization Science*, 23(5), 1398–1408.

Yoo, Y., Henfridsson, O., & Lyytinen, K. (2010). Research commentary: The new organizing logic of digital innovation: An agenda for information systems research. *Information Systems Research*, 21(4), 724–735.