Breaking Down Barriers:
Exploring The Motivations and Challenges in Adopting Responsible Data Annotation Practices of AI Developers

Author: Mathieu Koot
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT,
The rise of machine learning has awoken a rising demand for outsourced data work. One of the main drivers behind AI development is the practice of AI annotation, a practice largely being outsourced to online labour platforms. Online work platforms are often already identified by their malpractices, and alternative approaches to outsourcing data annotation do exist. This report seeks to explain why organisations do not switch to data annotation services that, as first referred to in this report as Responsible Data Annotation, do take responsibility for the wellbeing of their employees. Based on the job characteristics model, this method was laid out and then utilizes the value proposition canvas to determine how it manages to achieve the expectations and requirements of AI developers. Through the use of interviews with experts in the field, this paper has gathered the necessary information to form the customer profile. Finally, this study provides an interpretation of the inter relationship between the value of RDA and the different customer profiles identified, thus exploring the motivations and challenges of AI developers in transitioning to RDA.

Graduation Committee members:
Dr. J.G. Meijerink

Dr. R. Harms

Keywords
Data Annotation, Job Characteristics, Job design theory, Value Proposition Canvas, Responsible Data Annotation, AI Development

1. INTRODUCTION

The rise of machine learning has awoken a new demand for outsourced data annotation. From Amazon to OpenAI, from big tech corporations to universities, artificial intelligence (AI) developers outsource data annotation. Data Annotation is essential because it converts raw data into the valuable inputs that AI systems need in order to learn, adapt, and perform effectively. With data annotation being one of the main drivers behind AI development, it is seen being largely outsourced to gig workers through online platforms (Tubaro, Le Ludec, et al., 2020). Gig workers are independent contractors who take on short-term jobs or tasks, often through online labour platforms. As these gig workers play an increasingly significant role in AI development, their often isolating and precarious working conditions are drawing increased scrutiny from researchers and legislators (Le Ludec et al., 2023; Morgan et al., 2023; Newlands, 2021). While gig work can offer flexibility and opportunities for those seeking supplementary income or non-traditional employment (Goods et al., 2019), the lack of a standard employment relationship (Meijerink & Keegan, 2019) and other potential drawbacks of these platforms raise concerns about the sustainable future of this employment method.

In the current trends of research into artificial intelligence, the issues of online labour platform malpractices have been highlighted repeatedly. Online labour platforms do not provide a main stable occupation (Tubaro, Le Ludec, et al., 2020) and the un-challenging tasks of data-annotation still appears to be with it challenges and ethical dilemmas (Casilli et al., 2019; Newlands, 2021). Along with traditional HR practices losing their relevance in the gig work environment, attributed to the absence of traditional employer-employee relationships (Meijerink & Keegan, 2019), it becomes clear that fixing these issues cannot be done following the methodologies of current practices. As organisations will be relying evermore on gig workers for data development tasks, the current use of this platform of instability and regulatory struggles (Vallas & Schor, 2020) pushes us to rethink this method of data annotation.

However, an alternative to this approach for data annotation already exists. By addressing the drawbacks and introducing a more traditional employment relationship, this alternative, which has already been adapted by some organisations, presents a method that from this point will be referred to as Responsible Data Annotation (RDA). The RDA approach relies on the job characteristics model for its basis, which emphasizes designing jobs that enhance employee motivation and satisfaction (Hackman & Oldham, 1975). This model identifies five core job characteristics: skill variety, task identity, task significance, autonomy, and feedback. By incorporating these elements, RDA aims to create more fulfilling and secure employment opportunities, thus improving the overall job quality for data annotators. By offering better working conditions focused on long-term employment of workers with a wider range of tasks, RDA seeks to create jobs that offer more value for workers. This responsibility towards and sought after satisfaction of employees is what makes RDA responsible.

In 2021, a pilot program began through the cooperation of several media companies, Regio Gooi en Vechtstreek, UWV, and Werkgevers-Servicepunt Gooi en Vechtstreek to meet the demand for labelled data for AI (Annotatielab, 2024). This pilot, a case study for RDA, offered unemployed residents a learning-work program. Participants gained work experience at the AI Annotation Lab, with guidance from media organizations and job coaching from the region (Lamers et al., 2023). According to AI AnnotatieLab, this method leads to higher quality datasets, as employees better understand their tasks and can securely handle sensitive information. Despite controversies surrounding online labour platforms for data development and the recognition of RDA services as an alternative, there has been little movement from AI developers toward adopting RDA methods.

1.1 Research Objective

There has been little discussion about what the reasons are for organisation to be switching to which can be described as RDA methods as well as what barriers there are that are stopping them from making the switch. And given that the alternative is indeed better cost-based than the online gig work-based method, what is preventing organisations from opting for the method that is not troubled by controversial working conditions (Le Ludec et al., 2023)? Following this gap in knowledge, the question this paper aims to answer is:

"Why do AI developers (not) transition towards Responsible Data Annotation methods?"

By answering this question, this research aims to better understand the intention of users of data annotation services. More specifically, the research objective is to assess the motives of organisations and individual in need of annotated data in switching to RDA methods for data annotation services. While the job characteristics model of Hackman and Oldham (1975) will be used to explain RDA, it is the value proposition canvas that is used to show whether the characteristics of RDA-based data annotation alone are enough to motivate organisations to move away from online gig work platforms by providing insights in the level of value RDA might bring to AI developers (Osterwalder et al., 2014). By analysing the factors influencing this decision-making process and identifying potential barriers to adoption, light will be shed on the viability of RDA methods as a more ethical and sustainable alternative for data annotation services.

2. LITERATURE REVIEW

2.1 Data Annotation

2.1.1 Machine Learning and Data Annotation

Human contribution to AI development in the future is not certain, and there are already examples of implementation of technology managing to automate parts of the image annotation processes (Frey & Osborne, 2017; Greenwald et al., 2022). Further, as one is tempted to assume that the manual annotation of data is but a transitory phenomenon, as long as there are humans who can perform tasks more cheaply than machines it will be humans doing the work (Gray & Suri, 2017; Tubaro, Casilli, et al., 2020).

Machine learning has progressed from laboratory curiosity to widespread commercial use (Jordan & Mitchell, 2015). For AI development, machine learning has emerged as the method of choice for software development. Instead of manually fine tuning software to work, systems are trained by giving it data of desired input-output behaviour (Jordan & Mitchell, 2015). These machine-learning models are trained with high quality datasets of accurate and relevant information as the quality of the model made is highly dependent on the quality of the training data (Graham & Ferrari, 2022).

The term of data preparation introduced by Tubaro, Casilli, et al. (2020) covers the gig-work related function of data development related to data generation and annotation. Schmidt (2019) notes that there is a major difference between these two tasks. Data generation tasks are actually ideally performed through gig work platforms as they require little qualifications except fluency in the demanded language and masses of providers (Schmidt, 2019; Tubaro & Casilli, 2019).

On the other hand, the task of data annotation is increasingly complex and requires more understanding, making it so that it can be executed by a smaller number of individuals. Data annotation is the process of labelling or tagging data to provide context or meaning to it. This labelling is what enables computers to understand and interpret the data, making it possible for algorithms and models to analyse and derive insight from the data. For the purpose of this study, only data annotation is researched.

Competitiveness in data starved markets has resulted in organisations that require similar sets of annotated data to prepare each their own instead of collectively producing a pool of data, inflating the demand for data annotation services (Schmidt, 2022). While some believe that gig-work is but a temporary input to AI development, seeing a future in which data annotation as well as other tasks will be fully automated (Suleyman, 2023), there will always be a need for human touch (Ekbia & Nardi, 2017). The growing need for data will keep demand for gig work high in the foreseeable future (Tubaro, Casilli, et al., 2020) seen as that the development of AI has created a paradox; "The paradox of automation's last mile", introduced by Gray and Suri (2017). Because as AI progresses further, it also results in the rapid creation and destruction of new types of temporary labour jobs. Since machine learning is a data intensive process, data annotation work is consequently also labour intensive (Tubaro & Casilli, 2019).

2.1.2 Data Annotation Work Through Online Labour Platforms

The growing need for data has created a new demand for outsourcing data development work. Data annotation is, in nature, a menial and demotivating task for workers. Employers choose to avoid putting their valuable developers on this task as to not exhaust and demoralize them. Additionally, the sheer volume of data to be annotated often exceeds the capacity of in-house teams, necessitating the use of external resources, for which online labour platforms are being chosen (Le Ludec et al., 2023). Online Labour Platforms are mediators of employment arrangements in which individual find short-term tasks or projects via websites or mobile apps that connect them to clients and process payments (Goods et al., 2019; Kuhn & Galloway, 2019). Amazon Mechanical Turk is one of these gig worker platforms performing discrete, on-demand tasks that computers are unable to do economically. Mechanical Turk is a cheap way to hire remote workers for "human intelligence tasks" such as photo tagging, and has also proven to be a popular source of data for researchers in the social sciences (Buhrmester et al., 2018). But although it would be a godsend for AI developers to be able to get hold of valuable data for mere pennies, this method comes with its dark side.

2.1.3 Online Labour Platform Malpractices

Research by Le Ludec et al. (2023) on gig-workers who make use of online labour platforms has shown that they are repeatedly suffering from poorly paid, abusive, often automated, managerial practices, such as arbitrary account deactivation and wage theft from platform clients (Casilli et al., 2019; Gray & Suri, 2019). The human labour involved in AI annotation is often rendered invisible by being overlooked, marginalised and under rewarded through the use of online labour platforms and organisational practices. (Newlands, 2021). With their relatively low pay-rate and the repetitive nature of actions performed, online labour platforms for data annotation has earned the befitting title of "digital sweatshops" (Cushing, 2013). And as these online labour platforms are not officially employers of gig workers, things such as benefits like paid time off or health and worker insurance will not be offered (De Stefano, 2016). While the level of HRM responsibilities of online labour platforms can be considered high, the traditional employment relationship is absent (Meijerink & Keegan, 2019).

2.2 Responsible Data Annotation (RDA)

2.2.1 RDA as a Modus Operandi

Responsible Data Annotation (RDA) is a term that this paper introduces to describe a method of data annotation that considers the effects on workers and of human touch to deliver high-quality datasets for machine learning algorithms. This alternative method for data annotation focuses on addressing the drawbacks of current practices by focussing on the wellbeing of workers. By using Hackman & Oldham's Job characteristics model (1975), an explanation of what makes RDA responsible is given. Workers will be offered long-term employment opportunities with a wider range of tasks compared to online gig work platforms. RDA involves holding providers of data annotation services to high standards, ensuring accuracy of data labelling, high levels of data security and privacy by providing workers with deeper understanding of their tasks. RDA also aims at tackling the Corporate Social Responsibilities of organisations when developing machine learning. By providing long-term employment opportunities, RDA promotes fair treatment, job security, and access to benefits for workers, aligning it with the principles of responsible research and innovation giving AI developers an alternative to current practices more aligned with their responsibilities towards society (Jarmai et al., 2020; Le Ludec et al., 2023). RDA provides transparency about how to data is processed as well as who has access and works on it.

2.2.2 Job Characteristics of RDA

To establish how transitioning to RDA practices for data development could be better for both workers and employers alike, this paper will be relying on the Job design theory. This theory gives a model relating critical psychological states via job characteristics to personal and work related outcomes (Hackman & Oldham, 1975). By structuring the tasks of data annotators to encompass the different principles of job design, workers are expected to experience increased satisfaction, motivation, and performance. The low task significance nor identity related to data annotation work through online labour platforms, have both been identified by Blanz (2017) as having the highest relation to job satisfaction referring to Hackman and Oldham (1975) findings. Jobs done through those platforms require low level of skills and do not align with the need for personal growth.

Given that higher job satisfaction is achieved with increased motivation and productivity, employment through RDA methods seek to provide more task significance to workers, resulting in higher-quality data partly due to an increased sense of responsibility and variety of skills required. (Hackman & Oldham, 1975). Job design notes the importance of job rotation, which involves periodically rotating employees through different tasks. Although this approach is supposed to provide workers with opportunities to develop new skills and reduce boredom associated with performing the same task repetitively, changing the menial nature of data annotation is impossible. One way of improving job satisfaction of workers with RDA is by offering training relating to the field of data development, as seen in the example of the AI AnnotatieLab (Lamers et al., 2023), providing not just an opportunity for personal growth, but also increasing the skills related to data annotation work improving the quality of the annotated data.

Task significance is introduced into the data annotation process by RDA through its emphasis on providing workers with deeper understanding of their tasks with training, helping them realize how their contributions fit into the larger picture (Hackman & Oldham, 1976). RDA initiatives would involve the use of training programs for workers. This training goes beyond basic task instructions, aiming to deepen the understanding of the data they are annotating, specific requirements and the desired outcomes of AI developers. This allows RDA methods to perform their tasks with greater accuracy and diligence. Workers understand that their annotations directly influence the quality of the training data, which in turn, affects the performance of machine learning algorithms.

An RDA approach to data annotation allows for a feedback loop to be created between customers and data annotation workers. This open channel of communication between workers and AI developers allow for improved frequency and quality of feedback (Hackman & Oldham, 1976). This will not only provide the need for an individual to grow, as they will be able to know the actual results of their work, they are also expected to experience more responsibility for the outcomes of their work (Hackman & Oldham, 1976). It is this feedback between AI developers and data annotation services that is lacking for gig workers, limiting the quality of data that they can provide compared to RDA. And as gig workers have no clue what the activities of their work are used for, resulting from only completing a small part of a bigger job, they lack task identity. RDA workers on the other hand are able to finish their data annotation work from start to finish, generally providing them with more task identity (Hackman & Oldham, 1976). And as these RDA workers have increased data annotation skills, it is possible to provide them with more discretion regarding their job. Though data annotation tasks requiring strict adherence to predefined standards, limiting the little autonomy of workers in how they perform their tasks and clients having specific requirements in regards to the handling of their data, having provided the workers with training and having established communication channels can empower them to exercise autonomy within defined boundaries and allows them to express their ideas and generally collaborate more with the AI developers.

2.2.3 How RDA Seeks to Amend Shortcomings

While some of the characteristics of RDA have been addressed, not all its characteristics are better than those of services provided by online labour platforms. By providing traditional employment relationships to workers, RDA finds itself taking away the most powerful ability of nearly endless scalability of online labour platforms that can quickly be ramped up or down depending on the demand. RDA has less flexibility compared to this, needing more time for the recruitment and training of workers. This also results in the lower speed of RDA practices, as an AI developer can no longer upload their list of data and get it back annotated in a matter of days. RDA processes are more structured, not having the large pool of workers that online labour platforms can provide. This smaller pool of workers does though guarantee consistency of labelling through training and overseeing by managers.

RDA methods are expected, as a consequence of better working conditions, to be more expensive upfront for less data compared to others, which it seeks to compensate for by providing higher quality data. It seeks to achieve this by having more accurately annotated data, giving AI developers certainty about the contents of the data they receive. The higher quality data keeps itself from having to be corrected by the clients themselves or from negatively affecting the results of machine learning that would add additional overhead to solve the problems.

Table 1 Comparison Matrix

	Online Labour Platforms	**Responsible Data Annotation**
Efficiency and Speed	Typically, fast due to large pool of workers, but may lack consistency	slower due to more structured processes, but more consistent
Cost Structure	Low cost per task, hidden costs due to errors	High upfront costs, lower overall cost
Quality and Accuracy	Quality is uncertain and variable.	Higher quality and accuracy
Scalability	Endless scalability, can quickly be ramped up or down	Low flexibility, high lead time for scaling
Data Security and Privacy	Concerning lack of control over who has access to data	High level of data security and privacy
Skill Requirements and Training	Low skill requirements, little to no training	Higher skill level and continuous training
Worker impact	Poor working conditions, low pay, and lack of benefits. No positive impact on workers	Better working conditions, benefits from employment. Opportunities for growth and development

2.2.4 The Value Proposition Canvas

The value proposition canvas is a tool used to understand and visualize the relationships between a product or service and its customers introduced by Osterwalder et al. (2014). This canvas will be used to understand the value of RDA services for AI developers in order to answer the research question. By doing so, the canvas helps in understanding how RDA services effectively address the needs and desires of AI developers, likely giving insight into why they may or may not be transitioning towards RDA methods of data annotation. The canvas consists of two mains sections, the customer profile, and the value proposition.

The customer profile exists out of the three segments starting off with customer jobs. This section identifies the tasks, problems or needs that customers, in this case the developers of machine learning applications, are trying to address. These jobs will have to be ranked by importance as AI developers might deem some jobs more crucial to their work than others, because of them occurring more frequently or it heavily impacting their results (Osterwalder et al., 2014). Another section of the Customer profile are pains. The pains represent the negative aspects related to the experiences of customers regarding data annotation. These could include challenges, frustrations, or obstacles that they face as they are trying to annotate data themselves or have it outsourced. The last part of the customer profile are the gains. Gains represent the positive outcomes that AI developers are seeking. These could include desires outcomes, benefits, or improvements that AI developers hope to achieve by using a product or service.

The value proposition on the other hand focusses on the side and characteristics of RDA. The first section of this are the products & services. This section outlines the features of what RDA offers. It describes how the product or service alleviates the pains and creates gains identified in the customer profile.

These characteristics have already been outlined in earlier parts of the report, but there still is the interactions between the customer profile and value proposition which needs further exploration. There are also pain relievers in the value proposition. Pain relievers describe how the product or service addresses the pains identified in the customer profile. They represent the solutions or features that help alleviate AI developer frustrations. Lastly there are the gain creators. Gain creators describe how the product or service delivers the gains identified in the customer profile. These represent the features that provide value to AI developers and contribute to their desired outcomes.

The relationship between the customer profile and the value proposition fundamentally informs us about RDA by identifying how the needs, challenges and desired outcomes of AI developers are addressed. The value proposition addresses these needs and challenges by offering solutions and delivering value through data annotation. Pain relievers meanwhile directly address the pains identified in the customer profile, while gain creators deliver the desired gains. The purpose of using this canvas is to identify how the characteristics and features of RDA as described by the job characteristics align with the wants and needs of AI developers, in order to say something about their motivations. While literature gives insights to some of the potential gains and pains related to AI developers that RDA is hoping to address and mitigate, the relationship between the customer profile and value proposition and the severity of each pain and gain motivating AI developers to or not to transition towards RDA methods is still unclear.

3. METHODOLOGICAL APPROACH

3.1 Research Method

In order to answer the research question. This study will employ a qualitative research approach, seeking to identify factors influencing the decision-making process of AI developers in selecting different data annotation approaches. This will involve the conducting of in-depth interviews with developers of AI, users of annotation services and its providers. By utilising a qualitative approach this paper seeks to fulfil its instrumental role in providing understanding of data annotation alternatives, and shed a light on the reasons behind the lack in adoption (Baxter & Jack, 2008).

The goal of this research method is to allow interviewees to give insights into the Gains and Pains associated with data annotation practices related without influencing their answering. These problem-centred interviews offer a good approach for investigating the implicit dimensions of expert knowledge and integrating their experiences in search for answers (Döringer, 2021). Later in the interview, interviewees will be introduced to the RDA method of data annotation, after which their opinion will be collected. The study will focus on individuals who have expertise and experience in the field of machine learning for AI development, particularly those with experience in data annotation. This includes the focus on researchers who specialize in AI development techniques and applications, requiring annotated datasets, from the University of Twente. Other experts include industry practitioners of data annotation and the director of the AI AnnotatieLab as an example of RDA service provider. While the AI developers can provide insight into the elements of the value proposition Canvas, it is the AI AnnotatieLab that is able to provide insights on the extend of the value proposition. What makes these AI developers experts is their practical experience in the field. Interviewees may have firsthand experience with different data annotation services, including both traditional approaches and emerging alternatives like RDA. This offers real life perspectives on the Pains and Gains associated with data annotation for AI development. In-depth knowledge and understanding of the complications of data annotation. Their insights should provide valuable insights into the factors influencing the selection of data annotations service providers by giving information on the interaction of the customer profile and the value proposition of RDA. Additionally, alongside the data collected from the interviewees, data was collected through desk research to gain a deeper understanding of the context of the newly presented information from the interviews.

3.1.1 Interview Questions

The interview questions will form a baseline for the interviews to get the necessary information for answering the research question. These interview questions are aimed at getting the interviewee to describe their experiences with the pains, gains and customer jobs related to data annotation. The importance here lies in uncovering the role that data annotation plays in how they can perform their job. As such, all questions are separated into their respective categories, each question having an explanation of what they seek to answer in order to allow for easy follow up questions. Later in the interview, after the interviewee have described their own experiences, answering questions related to the Customer Profile, their opinion on the Value proposition of RDA will be gathered. This will be done by asking their opinions and feedback on how well the Products and Services described by RDA addresses the needs and challenges they have and how they believe RDA could or would influence their work.

3.1.2 Data Analysis

Every interview has been recorded and transcribed in order to successfully collect all data from the interviewees. The interview data management and analysis were performed using ATLAS.ti. To reduce and identify patterns in the transcribed text from the interviews, the data analysis strategy of a thematic analysis has been used. This is a method for identifying, analysing, organizing and reporting themes found within the data (Nowell et al., 2017). This method involved the six—phased method as documented by Braun and Clarke (2006) applied as a iterative process moving back and forward between steps with every new findings. All transcripts were initially read over to familiarize with the content, noting down initial impressions and findings. In combination with this, the value proposition canvas was used to identify possible themes. These findings were used to make initial codes, codes representing categories used for forming themes and patterns. ATLAS.ti allowed for preliminary codes to be organized into more significant themes and categories, supported by an iterative analysis of the interviews, giving an resulting analysis that showed which key topics appeared most commonly in the data and the patterns throughout the different interviews (ATLAS.ti, 2024). The data analysis included ranking the importance of the customer jobs found through the interviews as well as the pains and gains associated with them. This analysis allowed for the connection with and the examination of the value proposition of RDA to be made.

4. RESULTS

The following section presents the results from the interviews, supplemented by additional literature research. The results are organized into the two parts of the value proposition canvas, the customer profile, and the value proposition. Having looked at how RDA's value proposition relates to the customer profiles, a short section on how it does not is also provided. The interview participants will be referred to by the abbreviation IP, followed by the number associated with their interview. In Appendix B, a corresponding outline of each section can be found.

4.1 A Breakdown of the Customer Profile

The customer profile is used to categorise the findings according to their respective areas, to give a clear overview of the right side of the value proposition canvas. However, with a small sample size, caution must be taken in regard to the factors of the customer profile as these factors might not be transferable to all annotated data users. The customer jobs identified from the interviews for developers of machine learning applications relate to their tasks, problems and needs that they try to address. All participants identified annotated data as a need for their own roles or those of AI developers. This need requires the task of annotating, which they address by either annotating it themselves or having it done by others.

This differentiation is one of the crucial findings from the interviews, the distinction between the different customer segments based on how they solve their need for annotated data. All three customer segmentations share the customer job of needing annotated data to perform their job, this job ranging from AI development to research into human behaviour within the group of interviewees (IP1, IP2, IP3, IP6, IP8). There were two main customer segments existing of people who have data that they need annotating, the “do-it-themselves” and the “outsource to others” (IP1, IP2, IP3, IP4, IP5, IP6, IP8). As the names suggests, “do-it-themselves” are those who annotate the data they need to do their job themselves while the “outsource to others” choose to outsource the data annotation process to others. There is also a group of future potential customers who did succeed in obtaining data that has already been annotated and met their needs, thus not needing annotation services nor doing annotation themselves for now (IP6, IP8). These three groups were established by looking at the tasks they do and needs they have in relation to customer jobs, according to which a division of customer profiles was made each with some of their own unique pains, gains, and customer jobs. All groups hope to gain high-quality data with the annotated data they have, as higher-quality data means better training of the AI model.

4.1.1 The Outsource to Others

Firstly, turning to the “outsource to others” who choose to outsource their data annotation tasks to third-party services. This group recognizes the value of leveraging external expertise to meet their annotation needs, allowing them to focus on core activities such as model development and innovation. This segment often faces challenges related to maintaining in-house quality and scalability, leading them to seek reliable outsourcing partners. Despite the convenience and efficiency offered by outsourcing, these customers express concerns about data privacy, quality control, and effective communication with external vendors.

4.1.1.1 Customer jobs

One unique task of the client to outsourcing is the need to be able to communicate towards the annotators what their needs and expectations are, be this as an initial explanation, coding book or guidelines (IP1, IP2, IP8). The importance of providing and having an explanation in data annotation work, whether performed in-house or outsourcing, cannot be overstated (IP1, IP2, IP4, IP8). However, although it is not only important for the “outsource to others”, but also necessary to enable outsourcing. These explanations serve as a cornerstone for ensuring consistent, accurate and high-quality annotated data which are critical to the development of machine learning models (IP1, IP2, IP3, IP6, IP8). Interview participant two talks about taking theory as a baseline for an annotation guide, but as it often involves doing something new, theory will not be able to give the full picture.

Furthermore, when outsourcing, the importance of considering domestic and international services was noted, falling under the common theme of contextual and cultural understanding. Some data may require cultural & contextual background knowledge that may be difficult for annotators abroad to understand even with access to a guide, causing them to be unable to do their job properly if they have to start evaluating data as a part of the annotation process (IP2, IP7, IP8).

Lastly, when choosing to outsource, it is up to the client to check the annotated data for quality and perhaps correct it themselves or send it back if necessary (IP1, IP5, IP6). And with outsourcing clients, one must not only ensure data privacy inhouse, but also the outsourcing partner must ensure compliance with data security policies and relevant regulations (IP3, IP4, IP8).

4.1.1.2 Pains

When interviewees were asked about outsourcing options, a common emphasize was placed on their reluctance to outsource (IP1, IP3, IP4, IP5, IP6, IP8). This reluctance came paired mostly with their association of outsourcing with privacy concerns, lack of quality annotated data as well as ethical considerations (IP1, IP3, IP4, IP6). Interview Participant 1 even labels data annotation services as having a bad reputation due to the outsourcing to low-wage country workers with bad working conditions.

Although several factors influence the reasons to outsource annotation, it appears that awareness is also playing a major role. Most interviewees were not aware of any of the different outsourcing options, showcasing even a problem of awareness of the existence of this service market (IP3, IP4, IP5, IP7). Though also mostly unfamiliar with online labour platforms services, familiarity with gig work showed lack of trust in quality and concerns the ethicality of the working conditions (IP1, IP3, IP4). Regarding the ethical considerations of outsourcing showcase that the menial nature of data annotation work is a barrier towards doing the work themselves, but also creates reluctance towards putting other people to the task of only doing that kind of menial work (IP1, IP3, IP6).

What emerges from the interviews is that the biggest concern of outsourcing the data annotation process is related to the privacy and security of the data. Privacy concerns are the major factor when working with external organisations in regard to data handling (IP1, IP3, IP4, IP6, IP7). The sensitive data forms the backbone of machine learnings model, and any mishandling of this data can lead to severe privacy breaches (IP1, IP2, IP3). Data that is to be annotated may contain private information that an individual or organisation does not want to share with others (IP1, IP3). Organisations are so afraid of their data falling into the hands of competitors that they, as interview participant three says it, “watch over it like a dragon”. This fear has become a significant obstacle to outsourcing, as the value of their data is considered a critical asset. This concern aligns with the resource-based view theory, which posits that a firm's sustainable competitive advantage is derived from its valuable and inimitable resources (Barney, 1991). Relying on online gig platforms for data annotation aggravates these privacy concerns (IP1, IP3, IP4, IP6). Interviewees believed that it would be challenging to ensure consistent privacy standards through the usage of such platforms (IP1, IP3, IP6). Moreover, there is a difference between privacy-sensitive data and non-personal data, which can make it virtually impossible to outsource without proper advance planning (IP2).

Multiple interviewees stated that for them to consider the outsourcing of data annotation work, quality would have to be guaranteed as well as data privacy ensured (IP1, IP2, IP3, IP5, IP7). Interviewees noted themselves that if the annotated data did not meet their quality expectations, more work would have to be spent on correcting it (IP1, IP3, IP5). Ultimately, if data annotation is not done properly by a third party, the costs can become difficult to manage. As Interview Participant 1 describes outsourcing through online labour platforms, 'It's cheap, but often the quality is not good enough at all and so in the end not being cheap at all, because you still have to go and check and correct everything'.

Lastly, if the annotation process is outsourced to a foreign country, establishing expectations and needs through communication will just be a lot harder, especially if there is a need to return annotated data because of errors (IP1, IP6).

4.1.1.3 Gains

The main reason the annotation process is outsourced is that it saves a lot of time by doing so (IP1, IP6). Time that allows in-house focus on more important tasks (IP1, IP3, IP6). The cost of having an AI developer do annotation work himself is seen as a ridiculous expense (IP1, IP3, IP6). Interview participant six says that if the amount of data increased, the company would 100% have it outsourced, as it is just too expensive for your employees to continue doing it themselves. Some support for labour platforms was expressed in the way that they were cheap options for collecting data, and if the annotation were to be of equal quality it would also become a valid option for that (IP1, IP2).

4.1.2 The Do-It-Themselves

For the "do-it-themselves", the need for annotated data is solved by doing it yourself. This group often express a strong preference for maintaining direct oversight over their data, ensuring security and refinement throughout the annotation process. Despite the time-consuming and repetitive nature of this work, these developers often see it as an integral part of the learning and development process. While these people do the work themselves, it does not necessarily mean they would have chosen to do it themselves if they had the decision (IP5, IP6).

4.1.2.1 Customer jobs

The primary task that this group has is to of course annotate themselves. This involves labelling data so that it can be analysed and understood (IP1, IP2, IP3, IP4, IP5, IP6). While some choose to do it themselves because of data privacy reasons, for others there is too little data to consider it at all (IP2, IP3, IP4). As with outsourcing, a need for these developers is maintaining control over their data to ensure its security and privacy (IP3, IP4). In addition, annotation work for the "do-it-themselves" requires constant change to realize improvements in the AI training (IP3). Even though one can make a codebook with annotation rules, other interesting events can be observed which cannot be placed under any of the predefined categories (IP1, IP2, IP3). As said in one interview "You often see annotation as a combination of top down and bottom down (IP2)" meaning that annotators do not just try to interpret the unknown context of the data, but also apply known context to the data.

Lastly, there may be another need that developers try to address by annotating themselves. Namely by annotating themselves they also come to learn more about the data (IP2). For some, this is precisely part of the research as they do not yet know what they are looking for from the data (IP2, IP4, IP8).

4.1.2.2 Pains

Regarding the pains involved in the annotation process for the "do-it-themselves", the cost related to doing data annotation themselves was paramount (IP1, IP3, IP4, IP5, IP6). This is not just cost in terms of monetary costs, but also cost as in time consumed. Even with data sets that were so small that the AI model trained with them did not do a decent job, data annotation is repeatedly mentioned as a very repetitive and time-consuming process (IP1, IP3, IP4, IP5, IP6). In one case a team of developers was collectively working full-time for a fortnight annotating data (IP6), time that according to the interviewees would have been better spent doing what they are paid so much to do, and not repetitive annotation (IP1, IP3, IP6). Cost and time of doing data annotation inhouse as such were frequently mentioned as pains associated with the work (IP3, IP4, IP5, IP6). This is further confirmed by the interviewees who see every option for outsourcing as a method for cost savings if done correctly (IP1, IP2, IP3, IP4, IP6, IP7). Interview participant six mentioned that he would not be able to work for more than an hour at a time without taking a break, as it would drive him crazy.

Yet, even as some interviewees see the annotation process as a waste of valuable time, one interviewee mentioned a lack of budget to outsource annotating as a reason they have done and will do the data annotation themselves (IP5). It is often not the developers themselves who oversee the budgeting and the decision to outsource, but those above them (IP3). And according to interview participant three "You should not convince the developer, but the one who is in charge of the budget" as they are often not as technically oriented. Moreover, the pains grow when the amount of data grows to the point where it becomes unrealistic for an individual to do it themselves, in case the pain of a time-consuming task becomes too great and the option to outsource more attractive (IP2, IP3, IP5, IP6, IP7).

4.1.2.3 Gains

One of the positive outcomes that interviewees seek by annotating themselves is the assurance of data security (IP1, IP3, IP4). Because of strict laws regarding privacy and the competitive nature of data, developers may and can choose to keep a close eye on the process. Similarly to the pain associated with outsourcing related to the resource-based view theory, by annotating the data in-house, risks otherwise associated with outsourcing such as potential data breaches or misuse are mitigated.

Other benefits sought through doing the annotation work themselves have to do with the challenges of when it is either unclear how to actually annotate the data or when annotation work is part of the learning process required for their work (IP2, IP4, IP5, IP8). This iterative process, described by one interviewee as "quite the trial and error" (IP3), brings the advantage of enabling developers to experiment and make incremental changes, leading to progressively better model performance (IP1, IP2, IP3). Moreover, by doing the annotation work themselves, it allows them to make sense of their data to support their work, find patterns and improve annotations (IP2).

Furthermore, because the quality of the annotated data has a significant impact on the quality of the model that is trained with it, there is a considerable benefit in being able to control the quality by doing it yourself (IP5, IP6, IP8).

4.1.3 Pre-annotated data

Lastly, for the group of future potential customers who neither annotate data themselves nor have their data outsource, but rather choose to obtain pre-annotated data. Pre-annotated data refers to data that has been collected or bought from others that has already been annotated with the attributes relevant to the tasks. Though the annotated they obtain can meet their current needs and allows them to do their jobs, future problems have been noted by the interviewees. An explanation for why this type of customer exists in contrast to what has been said in

2.1.1 about the market being data starved is that these kinds of developers do not require as much data as others, often because they are working on a proof of concept (IP6, IP7), or because their data is actually widely publicly available.

4.1.3.1 Customer jobs

As for tasks, it is important for this group to be able even find pre-annotated data that meets their expectation and needs. Finding data online is often not a problem, but if this data is not already annotated, they would still have to get it annotated because of which they would no longer belong to this group (IP5, IP6, IP7, IP8). This group therefore includes task of evaluating the data they come across for quality and fit for their needs. If the pre-annotated data is not good enough, they must make corrections to it or work with what they got (IP5, IP7).

4.1.3.2 Pains

A major problem that is encountered when buying pre-annotated data is the lack of customization for their needs, either requiring the data annotation to be adjusted or forcing them to work with what they got (IP6, IP7). Not only can buying pre-annotated data be more expensive than doing it yourself (IP6), but it is also often not enough to make a satisfactory AI model (IP6, IP7). If these people were to continue with the development of their AI, not only more but also scenario specific data for the model would be required, making these developers opt to collect data themselves or collect this data through services to meet their needs (IP2, IP3, IP6, IP7). This new data would then need to be annotated putting them at the crossroad of choosing which type of customer to be.

4.1.3.3 Gains

The main gain to having the opportunity to purchase pre-annotated data directly is that it is immediately available. Especially for those who do not have their own data, it is very attractive to have annotated data immediately available with which to work (IP6, IP7).

4.2 Results of the RDA Value Proposition

4.2.1 How the RDA Value Proposition properties relate to the Customer Profiles

To understand whether the characteristics of RDA-based data annotation alone are enough to motivate organisations to move away from online labour platforms or to motivate organisations to outsource instead of doing it themselves, the identified factors of the value proposition are related to the customer profiles.

4.2.1.1 Products & Services

To reiterate, and once again clearly put down what features RDA offers, the features that seek to alleviate the pains and create the gains as identified in the customer profiles will be analysed.

The most notable feature of RDA services related to its value for AI developers is the annotation of their data according to their needs and expectations. RDA describes a model for services offering annotation of client data. The first feature of RDA related to its value is the open and direct interaction opportunities between annotators and clients. RDA also places a strong emphasis on data security and privacy. Open communication allowing transparency further seeks to support that. A core aspect of RDA is its responsible commitment to the workers. Annotators are offered long-term employment and comprehensive training, which ensures they are skilled and fully aware of the importance of privacy regulations. RDA is committed to providing good working conditions for its employees. Inspired by the job characteristics model (Hackman & Oldham, 1975), RDA ensures that annotators work in environments that promote job satisfaction and productivity. This focus on worker well-being seeks to translate into higher-quality annotated data for clients.

4.2.1.2 Pain Relievers

The interviews revealed that the feature of RDA alleviates the pains for the different customer profiles. The way RDA services are seen to alleviate the main pain is done by the existence of the service, which should take work away from AI developers allowing them to spend their valuable time on other parts of their work (IP1, IP2, IP3, IP4, IP5, IP6, IP7, IP8).

Through transparency and certificates about data safety and working conditions, RDA is believed to address ethical concerns about the working conditions as well as the security of the data and processes (IP1, IP3, IP7, IP8). Transparency is an important factor, being able to provide the information of what goes on with and who has access to their data exactly is seen as crucial for ensuring privacy for those who seek to outsource (IP1, IP3, IP6). To realize the pain reliever of data security and compliance, providers of RDA services must enforce strict data handling protocols to ensure that sensitive information is protected throughout the annotation process (IP3, IP4). During interviews, it became clear that following laws for handling data is highly significant and the only way to guarantee proper data handling (IP4, IP8). The method for showcasing that an organization follows the appropriate laws, regulations and policies related to data privacy and security, as mentioned by interview participant 8, is by obtaining certifications. Certifications are a guaranteed method of ensuring data privacy according to industry standards. Only after obtaining these certifications can an organisation become trustworthy (IP8). The European GDPR Institute provides a platform for data processors to ensure a structured and efficient means for GDPR compliance demonstrating data security (EU-GDPR-Institute, 2019). There are also internationally recognized certifications of data protection such as the ISO/IEC 27001 and ISO/IEC 27701 that show to others that a system is in place to manage risks related to the security of data handled, respecting all the best practices and principles enshrined in the standard (ISO, 2024).

Improved working conditions are considered an effective and essential factor for AI developers when considering an outsourcing option (IP1, IP3). RDA successfully emphasizes long-term employment and training, providing a stable and secure working environment. This stability not only is believed to ensure higher quality in data annotation, but it also aligns with the ethical values of many AI developers who seek to contribute positively to society (IP1, IP4, IP6). Interviewees believe that having people work on a wider range of tasks, and having some ability to choose which tasks to work on themselves gives annotators more autonomy increasing their working conditions (IP1, IP3, IP6). If the work provided through RDA can provide a better future for workers, this is only seen as a bonus and advantage over others in terms of responsibility of the clients (IP1, IP3).

Adding onto worker wellbeing, ethical considerations regarding working conditions are a strong motivator for choosing RDA services above others. Some interviewees show reluctance towards choosing online labour platforms due to ethical dilemmas related to unstable income, lack of benefits, and minimal worker protections (IP1, IP3, IP4). This not only improves the working conditions for annotators but also further helps to ensure that the annotated data is of high quality. To be considered RDA, the promotion of fair labour practices by offering stable employment, fair wages and benefits towards annotators is crucial. This is not only related to the working conditions for annotators but also enhances the quality of the annotated data.

Moreover, there is the ISO 45001, an international standard that specifies requirements for an occupational health and safety management system (ISO, 2018). The standard establishes criteria for health and safety policies, and adopting the standard shows that the organization is committed to worker health, safety and wellbeing (ISO, 2018).

Lastly, the initial explanation of how to annotate and the provision of a guideline are believed to be essential for establishing transparency, accountability and above all task significance, providing workers with deeper understanding of their tasks. A well-made guide helps prevents mistakes, if it can provide contextual understanding and establishes quality expectations (IP1, IP2, IP8). Lastly, several interviewees noted that the effectiveness of outsourcing relies on communication (IP1, IP6). Interview participant six mentioned that it would be problematic to communicate without a fixed communication line during the annotation process, emphasizing that establishing this line before starting the task is just as important. Interviewees believe that the quality of the annotation depends on the preparation of the AI developer as the responsibility of providing annotators with the know-how on how to annotate their specific data lies with them (IP1, IP2, IP4, IP8). The client-annotator interaction provided by the direct contact and open communication opportunities is also shown to create trust in the RDA process, combatting reputation issues (IP1, IP3, IP6).

4.2.1.3 Gain Creators

The main feature of RDA that is also seen as a plus from the interviewees is the responsibility towards workers in the form of good working conditions, this is seen as a gain creator as it has also been connected to the expectation of higher quality annotated data (IP1, IP3, IP6). Providing a service for annotating data creates the benefit of streamlining the annotation process, helping AI developers save time to focus on core activities (IP1, IP3, IP6).

One of the presumed benefits of RDA, the collaboration relationship between client and annotator has been cited as a benefit to both parties during the interviews (IP1, IP6). An open communication channel would allow for continuous feedback and understanding of the data for the annotator giving more meaning to the work while also providing the client with more control over the iterative process (IP1, IP2, IP3, IP6). This has been noted by the interviewees as adding control over the quality of annotated data allowing iterative improvements (IP1, IP2, IP3, IP6). As data annotation is an iterative process, constantly making small changes to the annotations to see improvements in the AI training (IP3). This can become quite the trial-and-error procedure when it is difficult to predict in advance whether one or any dataset is going to lead to a well-trained AI (IP3).

RDA service providers are also believed to save clients' money in the long run, even when having to ask for higher rates, because if RDA providers are able to deliver guaranteed quality annotations, the need to check and correct the data diminishes resulting in even less work (IP1, IP6). Considering this, it was mentioned that it is those who decide on the budgeting that have to be convinced of the added value from working with an external organisation as they often lack the technical knowledge to realize the gains it could create (IP3).

Furthermore, interviewees expressed a preference for having a regular point of contact for quick communication, making it easier to address issues and implement necessary changes when problems arise. (IP1, IP2, IP3, IP6). Through the collaboration from both parties involved, high quality of annotated data is expected as expectations are known and met (IP1, IP6, IP8). The collaboration with external annotators was also named as an opportunity to get feedback on their research in addition to a way to thread new insights that they themselves may have overlooked (IP1, IP5). As this is a two-way collaboration, opportunities for open communication can ensure that questions are asked promptly when annotation instructions or guidelines are unclear, allowing for issues that the client may not have anticipated to be addressed, enabling both parties to learn and improve (IP1).

Another benefit is that by complying to the privacy regulations as mentioned in the pains, RDA services can assure clients of robust data security practices aiming to equal the privacy gains from the "do-it-themselves" (IP8). Having long-term, known employees would also help with this, considering that their names can be included in the project plan of clients meeting data security laws (IP1, IP4).

4.2.2 Insurmountable hurdles for RDA

There are some pains and gains that AI developers seek that RDA is not able to meet.

What emerges from the interview with participant 1 is that costs for those who are already outsourcing to online labour platforms is a major issue when switching, as they are already accustomed to the low prices the online labour platform has and are fine with the quality they receive, the higher rates that RDA providers would ask for to provide better working conditions are deemed too expensive (IP1).

Furthermore, one other gain RDA is not able to provide is the instantly availability of annotated data, which will be the same for any source of outsourcing as annotating takes time. Therefore, there is no way for a service to offer the instant availability of annotated data that pre-annotated data buyers currently enjoy.

Lastly, when data can be split into objective occurrences, allowing for judgment and observation without cultural interference or background knowledge, the need for collaboration dissolves together with its added value (IP2). And as interview participant two further suggests, a domestic data annotation service provider could outsource objective data to low-wage countries, while complying with data regulations, and maintain the subjective annotation part domestically, providing a potentially cost-effective alternative to RDA.

5. DISCUSSION

The discussion will begin with an interpretation and explanation of the results, looking at how the research question was answered. Furthermore, the practical and theoretical implications of the research are outlined, followed by the limitations. A final conclusion is given after which future research needs are explained.

5.1 Interpretation of Results

The present study was designed to determine the factors motivating or demotivating AI developers to transition towards RDA methods for data annotation. Returning to the question posed at the beginning of this study:

"Why do AI developers (not) transition towards Responsible Data Annotation methods?"

By having conducted interviews with experts in the field and analysing the responses, a clearer understanding of the intentions and preferences of data annotation users has been gained.

A noteworthy observation from the customer profiles is that the pains and gains of the two groups, “outsource to others” and “do-it-themselves”, seem to mirror each other, what a frustration is for “do-it-themselves” appears to be a benefit of outsourcing and vice versa. By keeping the annotation process in-house, the “do-it-themselves” group seeks to avoid the pains associated with outsourcing and leverage the gains associated with maintaining direct control over their data annotation activities while the “outsource to others” seek the benefits of not having the pains of doing it themselves.

The value proposition of RDA aligns with the insights from the interview with the director of the AI AnnotatieLab. Customers that used to fall under the “do-it-themselves” category are willing to pay the premium in order to help society in the form of their learning-work programme. Combining other interviews and what interview participant one identifies as their biggest barrier to RDA adoption, awareness of these types of services, reasons to be “do-it-themselves” can be interpreted.

AI developers who choose to keep data annotation in-house primarily do so to avoid the pains associated with outsourcing. These include concerns over data privacy, quality control, and the perceived ethical considerations of external providers. By maintaining direct control over their annotation processes, developers can ensure data integrity and confidentiality, addressing their top priorities without the need for extensive external oversight. The in-house approach allows for immediate feedback and adjustment, ensuring that the data meets specific internal standards and requirements.

The primary motivator for outsourcing data annotation is the potential for cost and time savings. Outsourcing can streamline the annotation process, freeing up internal resources and allowing AI developers to focus on their core competencies. By leveraging external expertise, companies can achieve high-quality annotations more efficiently. However, this decision often hinges on the balance between the immediate cost savings and the long-term benefits of superior data quality. Outsourcing to online labour platforms or lower-cost providers can be appealing due to reduced expenses, but these options may compromise data quality and ethical standards.

5.1.1 Conclusion; Why Transition Towards RDA

RDA offers a compelling option for data annotation outsourcing by aligning with ethical standards and ensuring high-quality, privacy-focused data annotation. The first motivator for AI developers to transition to RDA is the emphasis on the ethical treatment of workers. Secondly, the RDA model is believed to deliver high data quality, translating into more reliable AI training. By ensuring annotators are well-trained and motivated RDA can enhance the accuracy and utility of annotated data. RDA services typically offer robust privacy protections, addressing a significant concern for many AI developers. By ensuring that data is handled securely, RDA reduces the risks associated with outsourcing sensitive information.

Yet as the value proposition of RDA, as seen in the results, is believed to address both the pains they face and gains they desire, there are several barriers that may deter AI developers from adopting RDA. The premium associated with RDA services can be a significant deterrent, especially for developers with limited budgets or those handling smaller volumes of data. The initial expenses may outweigh the perceived long-term benefits for some organizations. A lack of awareness about the benefits of RDA services is a critical barrier. Many potential users are not fully informed about how RDA can address their pains and enhance their gains, leading to reluctance in transitioning from existing methods. Developers accustomed to in-house annotation or existing outsourcing arrangements may find it challenging to switch to a new model. This resistance is often fuelled by satisfaction with current processes or a belief that the benefits of RDA do not justify the effort and cost of transition. Lastly, while RDA emphasizes data privacy, the perception that outsourcing inherently compromises security can still deter developers. Ensuring data privacy in a way that convinces sceptical developers remains a significant challenge for RDA providers.

In conclusion, the high importance of privacy and ethics in data annotation underscores the need for existence of responsible practices. Having analysed the findings from the interviews, RDA is believed to address these needs effectively, providing a competitive alternative to online labour platforms by ensuring high privacy standards and ethical integrity improving the annotated data quality.

5.2 Practical Implications

From a practical lens, the results have implications for providers of data annotation services and their potential users.

This paper has identified the need for organisations providing data annotation services in a responsible manner to effectively showcase their capabilities and benefits to address common misconceptions about the quality and privacy of their services. For those who currently handle annotation in-house, it is crucial to clearly communicate the advantages of RDA services and justify the associated costs to their funders. For organizations that already outsource, it is important to highlight how RDA can provide benefits beyond what online labour platforms offer.

Additionally, the RDA model must adapt to address not only responsibility towards workers, but also the integrity and security of the data itself. Changes are required, adapting from the results, in order for RDA methods to be not only a more ethical but also economically sustainable alternative for data annotation services, motivating all customers from its characteristics to choose RDA. From the results, it is clear that RDA by itself does not have the characteristics to persuade everyone to transition. Changes that could increase the possibilities of offering this kind of service at a low price while guaranteeing that responsibility to data and employee should be looked at given that this could convince current “do-it-themselves” to start outsourcing more and convince the users of online labour platforms to switch.

Another implication for service providers, be this RDA services or not, is regarding the handling of sensitive data. The study reveals that privacy concerns are a major barrier to outsourcing. Providers must demonstrate how they mitigate risks associated with handling sensitive information to alleviate client reluctance and build trust. Given the high stakes involved in outsourcing privacy-sensitive data, service providers need to prioritize and clearly communicate their strategies for ensuring data security.

However, there are further implications regarding privacy. With all the risk that outsourcing carries for the client should something go wrong, resulting in their reluctance to outsource, the need to offer services for annotating privacy-sensitive data has to be considered. To add to this, if privacy and quality can be guaranteed for low prices by services in low-wage countries, what place do domestic services such as the AI AnnotatieLab even have?

5.3 Theoretical Implications

The combination of the results provides support for the various theories used to write this report and the development of the RDA methodology.

This paper has served as a practical example of how the model of Hackman and Oldham can be used to study pain relievers and gain creators. The study was able to relate task significance, skill variety and feedback to the different pain relievers and gain creators of RDA. While the Hackman and Oldham model highlights these dimensions, it falls short in addressing certain critical aspects such as data privacy and security. To cover these gaps, additional theories are to be explored to understand how to create more pain relievers and gain creators if possible.

The resource-based view could help explain why AI developers may prefer to keep data annotation in-house due to strategic value of their data and the desire to maintain control over sensitive information. This theory provides insights into the competitive advantage gained from unique resources, including well-managed, high-quality data annotation processes (Barney, 1991).

Furthermore, the transaction cost theory can complement explaining the decision-making process involved in outsourcing. The transaction cost theory considers the transaction as the most basic unit of measure and focuses on how much effort, resources, or cost is necessary for two parties to complete an exchange (Williamson, 1981). Transaction costs are defined as the costs beyond the cost of the product or service that are required for the exchange between two entities (Sarkis et al., 2011). Transaction cost theory also presents a rational view for evaluating 'make versus buy' decisions related to data outsourcing, possibly explaining the mirroring of the customer profiles (de Camargo Fiorini et al., 2018).

Lastly, as responsible data annotation is framed within the context of corporate social responsibility, CSR frameworks emphasize ethical considerations, fair treatment of workers, and job security, which are crucial for understanding the motivations behind adopting RDA methods. CSR helps frame RDA within the broader context of responsible innovation, aligning business practices with societal expectations and ethical standards (Jarmai et al., 2020).

5.4 Limitations

During the research for this report several limitations were encountered that may have affected the results. Firstly, the scope of the study was constrained due to a bias from the limited pool of interview participants, primarily comprising researchers from the University of Twente. Only three interviews were conducted with individuals outside the institutions. This may have skewed the findings towards the perspectives common at the University of Twente and its researchers.

Because of this, the findings may not be applicable in other fields of AI development aside from that of pioneering research which takes place at the university.

Secondly, the research faced delays due to late responses and sudden cancellations by some interviewees. These unexpected circumstances disrupted the planned schedule, leading to tight time constraints. Firstly, these delays affected the thematic analysis process, leading to a more rushed evaluation. This directly reduced the depth of the data collection efforts. Secondly, the cancellations resulted in missing perspectives that would have contributed to a broader understanding of the topic. These delays thus limited the depth and breadth of the results.

Lastly, there was a lack of knowledge among participants regarding online labour platforms, its usage for annotation, and the controversies surrounding them, with none of the interviewees had personal experience with annotation work through these platforms. Because of this, it was necessary to explain the concept of online labour platforms to these participants, which may have introduced interviewer bias.

5.5 Future Research Recommendations

Based on the implications derived from the study and considering the limitations identified, some suggestions for future research can be made.

As the costs of data annotation, both in terms of time and money, done inhouse or outsourced is a critical factor to the decision to outsource, a cost-effectiveness analysis of the service should be done providing detailed analyses and case studies to help developers and managers understand the long-term savings and efficiency gains potentially gained from outsourcing. This future research should also investigate whether RDA reduces the costs and efforts associated with initializing data annotation outsourcing practices. Additional investigation ought to be done on the actual quality differences of objectively and subjectively annotated data between services covered by RDA, online labour platforms and others. To elaborate on this further, the value of services where the data annotation process is split into two as described by interview participant two will need to be examined. The opinions of actual users of such services and whether the benefits of the RDA model are also applied were not explored in this study.

The theoretical foundations on which RDA is based needs expanding. By integrating insights from various domains, such as the resource-based view and transaction cost theory, future research should better understand the value proposition of RDA. The value of annotation services in relation to sensitive data, such as that pertaining to individuals or competitive advantages, needs to be considered. The possibilities of collaborating in order to annotate even sensitive data for others and still meet those data security requirements should be looked at, after which it should also be considered whether it is worth it. Future research needs to clarify the advantages and drawbacks of outsourcing big data initiatives such as data annotation. Researchers are encouraged to investigate whether data annotation activities should be conducted outside or within an organization through the lens of transaction cost theory.

Finally, future studies should examine differences in adoption across a larger and more diverse pool of interview participants, extending beyond the group of researchers from the University of Twente. It is important to include individuals from companies developing AI for various purposes, such as commercial enterprises, government agencies, and non-profits. Additionally, exploring the perspectives of international participants can provide a more comprehensive understanding of how RDA services are perceived and utilized globally.

6. ACKNOWLEDGEMENTS

My utmost thanks to all the people who were able to make this research possible. Firstly, I would like to thank Dr Jeroen Meijerink for his active help in making this research better. With his support, insights and feedback, the whole research process was more trouble-free than it would have been without him. Secondly, I would like to thank all those interviewed for their participation and willingness to be interviewed.

7. REFERENCES

Annotatielab, A. (2024). *Onze Missie: .* https://www.annotatielab.nl/over-ons

ATLAS.ti. (2024). *Interview Analysis Tools*. https://atlasti.com/interview-analysis-tools

Barney, J. (1991). Firm Resources and Sustained Competitive Advantage. *Journal of Management*, *17*(1), 99-120. https://doi.org/10.1177/014920639101700108

Baxter, P., & Jack, S. (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2008.1573

Blanz, M. (2017). Employees' Job Satisfaction: A Test of the Job Characteristics Model Among Social Work Practitioners. *Journal of Evidence-Informed Social Work*, *14*, 1-16. https://doi.org/10.1080/23761407.2017.1288187

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Buhrmester, M., Talaifar, S., & Gosling, S. (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science*, *13*, 149-154. https://doi.org/10.1177/1745691617706516

Casilli, A. A., Tubaro, P., Le Ludec, C., Coville, M., Besenval, M., Mouhtare, T., & Wahal, E. (2019). *Le Micro-Travail en France. Derrière l'automatisation*.

Cushing, E. (2013). Amazon mechanical turk: The digital sweatshop. *Utne Reader*, 1868-1884.

de Camargo Fiorini, P., Roman Pais Seles, B. M., Chiappetta Jabbour, C. J., Barberio Mariano, E., & de Sousa Jabbour, A. B. L. (2018). Management theory and big data literature: From a review to a research agenda. *International Journal of Information Management*, *43*, 112-129. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2018.07.005

De Stefano, V. (2016). The Rise of the "Just-in-Time Workforce": On-Demand Work, Crowdwork, and Labor Protection in the "Gig Economy". *Comparative labor law journal: a publication of the U.S. National Branch of the International Society for Labor Law and Social Security [and] the Wharton School, and the Law School of the University of Pennsylvania*, *37*, 471-504.

Döringer, S. (2021). 'The problem-centred expert interview'. Combining qualitative interviewing approaches for investigating implicit expert knowledge. *International Journal of Social Research Methodology*, *24*(3), 265-278. https://doi.org/10.1080/13645579.2020.1766777

Ekbia, H. R., & Nardi, B. A. (2017). *Heteromation, and other stories of computing and capitalism*. MIT Press.

EU-GDPR-Institute. (2019). *GDPR Certification*. EU GDPR Institute. https://www.eugdpr.institute/gdpr-certification/

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254-280. https://doi.org/https://doi.org/10.1016/j.techfore.2016.08.019

Goods, C., Veen, A., & Barratt, T. (2019). "Is your gig any good?" Analysing job quality in the Australian platform-based food-delivery sector. *J. Ind. Relat.*, *61*(4), 502-527. https://doi.org/10.1177/0022185618817069

Graham, M., & Ferrari, F. (2022). *Digital work in the planetary market*. Mit Press.

Gray, M. L., & Suri, S. (2017). The humans working behind the AI curtain. *Harvard business review*, *9*(1), 2-5.

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., Fullaway, C. C., McIntosh, B. J., Leow, K. X., Schwartz, M. S., Pavelchek, C., Cui, S., Camplisson, I., Bar-Tal, O., Singh, J., Fong, M., Chaudhry, G., Abraham, Z., Moseley, J., . . . Van Valen, D. (2022). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.*, *40*(4), 555-565. https://doi.org/10.1038/s41587-021-01094-0

Hackman, J., & Oldham, G. (1975). Development of Job Diagnostic Survey. *Journal of Applied Psychology*, *60*, 159-170. https://doi.org/10.1037/h0076546

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: test of a theory. *Organizational Behavior and Human Performance*, *16*(2), 250-279. https://doi.org/https://doi.org/10.1016/0030-5073(76)90016-7

ISO. (2018). ISO 45001:2018. https://www.iso.org/standard/63787.html

ISO. (2024). *ISO/IEC 27001:2022/AMD 1:2024*. https://www.iso.org/standard/27001

Jarmai, K., Tharani, A., & Nwafor, C. (2020). Responsible Innovation in Business. In K. Jarmai (Ed.), *Responsible Innovation : Business Opportunities and Strategies for Implementation* (pp. 7-17). Springer Netherlands. https://doi.org/10.1007/978-94-024-1720-3_2

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/doi:10.1126/science.aaa8415

Kuhn, K., & Galloway, T. (2019). Expanding perspectives on gig work and gig workers. *Journal of managerial psychology*, *34*, 186-191. https://doi.org/10.1108/JMP-05-2019-507

Lamers, L., Dorenbosch, L., & Fermin, B. (2023). *Inclusief Platformwerk: Nieuwe kansen op waardevol werk via online arbeidsplatforms*. Nederlandse Stichting voor Psychotechniek.

Le Ludec, C., Cornet, M., & Casilli, A. A. (2023). The problem with annotation. Human labour and outsourcing between France and Madagascar. *Big Data & Society*, *10*(2), 20539517231188723. https://doi.org/10.1177/20539517231188723

Meijerink, J., & Keegan, A. (2019). Conceptualizing human resource management in the gig economy: Toward a platform ecosystem perspective. *Journal of managerial psychology*, *34*(4), 214-232.

Morgan, R. A., van Zoonen, W., & ter Hoeven, C. (2023). Lost in the crowd? An investigation into where microwork is conducted and classifying worker types. *European Journal of Industrial Relations*, *29*(3), 301-322. https://doi.org/10.1177/09596801231171997

Newlands, G. (2021). Lifting the curtain: Strategic visibility of human labour in AI-as-a-Service. *Big Data Soc.*, *8*(1), 205395172110160. https://doi.org/10.1177/20539517211016026

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis:Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*(1), 1609406917733847. https://doi.org/10.1177/1609406917733847

Osterwalder, A., Pigneur, Y., Bernarda, G., & Smith, A. (2014). *Value proposition design : how to create products and services customers want*. John Wiley & Sons.

Schmidt, F. A. (2019). *Crowdproduktion von Trainingsdaten: Zur Rolle von Online-Arbeit beim Trainieren autonomer Fahrzeuge* (417). (Hans-Böckler-Stiftung, Issue. https://hdl.handle.net/10419/201850

Schmidt, F. A. (2022). The Planetary Stacking Order of Multilayered Crowd-Al Systems. *Digital work in the planetary market*, 137-156.

Suleyman, M. (2023). *The coming wave: technology, power, and the twenty-first century's greatest dilemma*. Crown.

Tubaro, P., & Casilli, A. A. (2019). Micro-work, artificial intelligence and the automotive industry. *Journal of Industrial and Business Economics*, *46*(3), 333-345. https://doi.org/10.1007/s40812-019-00121-1

Tubaro, P., Casilli, A. A., & Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data Soc.*, *7*(1), 205395172091977. https://doi.org/10.1177/2053951720919776

Tubaro, P., Le Ludec, C., & Casilli, A. A. (2020). Counting 'micro-workers': societal and methodological challenges around new forms of labour. *Work Organ. Labour Glob.*, *14*(1), 67. https://doi.org/10.13169/workorgalaboglob.14.1.0067

Vallas, S., & Schor, J. B. (2020). What do platforms do? Understanding the gig economy. *Annu. Rev. Sociol.*, *46*(1), 273-294. https://doi.org/10.1146/annurev-soc-121919-054857

Williamson, O. E. (1981). The economics of organization: The transaction cost approach. *American journal of sociology*, *87*(3), 548-577.

8. APPENDIX A

Describing the Value Proposition Canvas regarding Responsible Data Annotation for AI development. Let the protocol guide the interview but ask follow-up questions based on the answers of the interviewee.

Introduction Question regarding Organization and Employee (Customer Jobs)

1. Could you provide a brief description of your work in relation to AI development?
    a. *Make sure they provide information about the following:*
        i. *Tasks & Responsibilities*
2. What kind of data do you use, pictures videos etc, and where or how do you obtain this data?
3. Could you explain to me the process of labelling or tagging of data?
    a. *What are the things you or others do with the data before you can train those models?*
    b. *Who is involved in this, inside or outside the organisation?*
        i. *Explain that this is your understanding of data annotation.*
    c. *For outsourcing this task, do you use online platforms? Which platforms do you use for this purpose? Such as AmazonTurk*

Pains Related to Data Annotation

4. What challenges or frustrations do you encounter regarding data preparation and annotation for AI development?
        i. *Refer back to definition of data annotation.*
        ii. *Possible explanation of the differences between preparation and annotation*
    b. *Make sure they provide information about the following:*
        i. *Quality and Accuracy*
        ii. *Efficiency and Speed*
5. What are the main drawbacks or concerns you have about the current methods of data annotation, particularly through online labour platforms?
    a. *Make sure they provide information about the following:*
        i. *Worker impact*
        ii. *Cost Structure*

Gains Expected from Data Annotation

6. What are the desired outcomes or benefits you hope to achieve with data preparation & annotation work?
7. How do you envision an ideal scenario for data annotation that would enhance your work as an AI developer?
    a. *In what ways could data annotation services further help you do your job?*
8. What positive impacts do you anticipate for your work if the data annotation process were improved or optimized?
    a. *In what form would these benefits need to come?*

Products & Services of Responsible Data Annotation (RDA):

First, explain the products and services of RDA's value proposition and say something about the AI AnnotatieLab in Hilversum as an example.

9. Based on your experiences, how do you think RDA methods address the challenges or pains you have encountered with traditional data annotation practices?
    a. *Relates to Pain Relievers as well as making them describe RDA in their own words.*
    b. *Make sure they provide information about the following:*
        i. *Skill Requirements and Training*
        ii. *Cost Structure*
10. How do the features of RDA align with your needs as an AI developer?
    a. *Make sure they provide information about the following:*
        i. *Characteristics of RDA*
        ii. *Relation to their needs*

Pain Relievers of RDA:

11. Which specific aspects of RDA do you believe effectively alleviate the frustrations or concerns you have faced regarding data preparation and annotation?
    a. *Make sure they provide information about the following:*
        i. *Expected Pains*
12. How do you think RDA methods could improve data annotation compared to the methods you currently use?

a. *Make sure they provide information about the following:*
    i. *Efficiency & Speed*
    ii. *Cost Structure*
    iii. *Data Security & Privacy*
    iv. *Scalability*
    v. *Quality and Accuracy*
    vi. *Worker Impact*

Gain Creators of RDA:

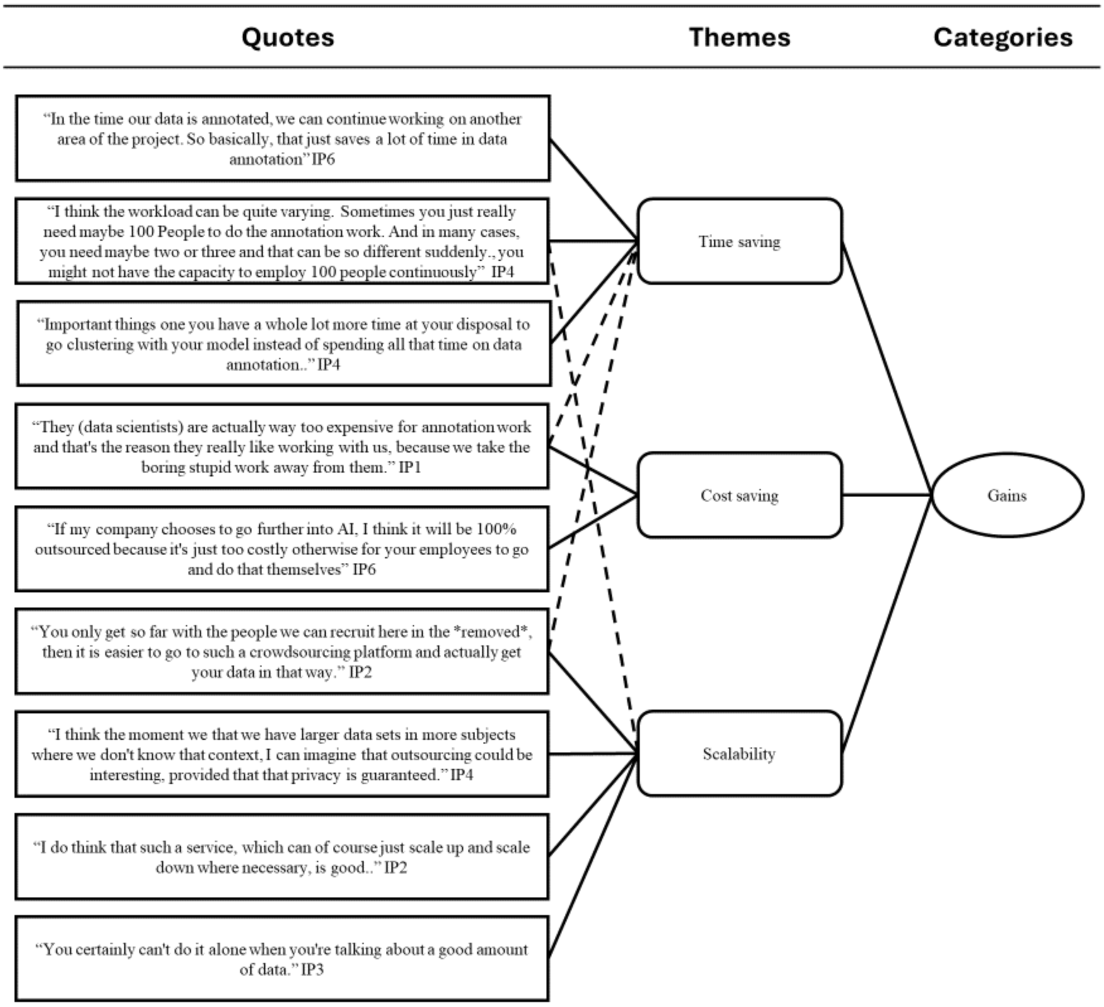
13. In what ways do you anticipate that RDA methods can contribute to achieving your desired outcomes or benefits in data annotation?
    a. *Make sure they provide information about the following:*
        i. *Expected Gains*
        ii. *Positive impacts of Data Annotation Services*
14. How do you think RDA can enhance your work as an AI developer?
    a. *Make sure they provide information about the following:*
        i. *Collaboration Opportunities*
        ii. *Quality*

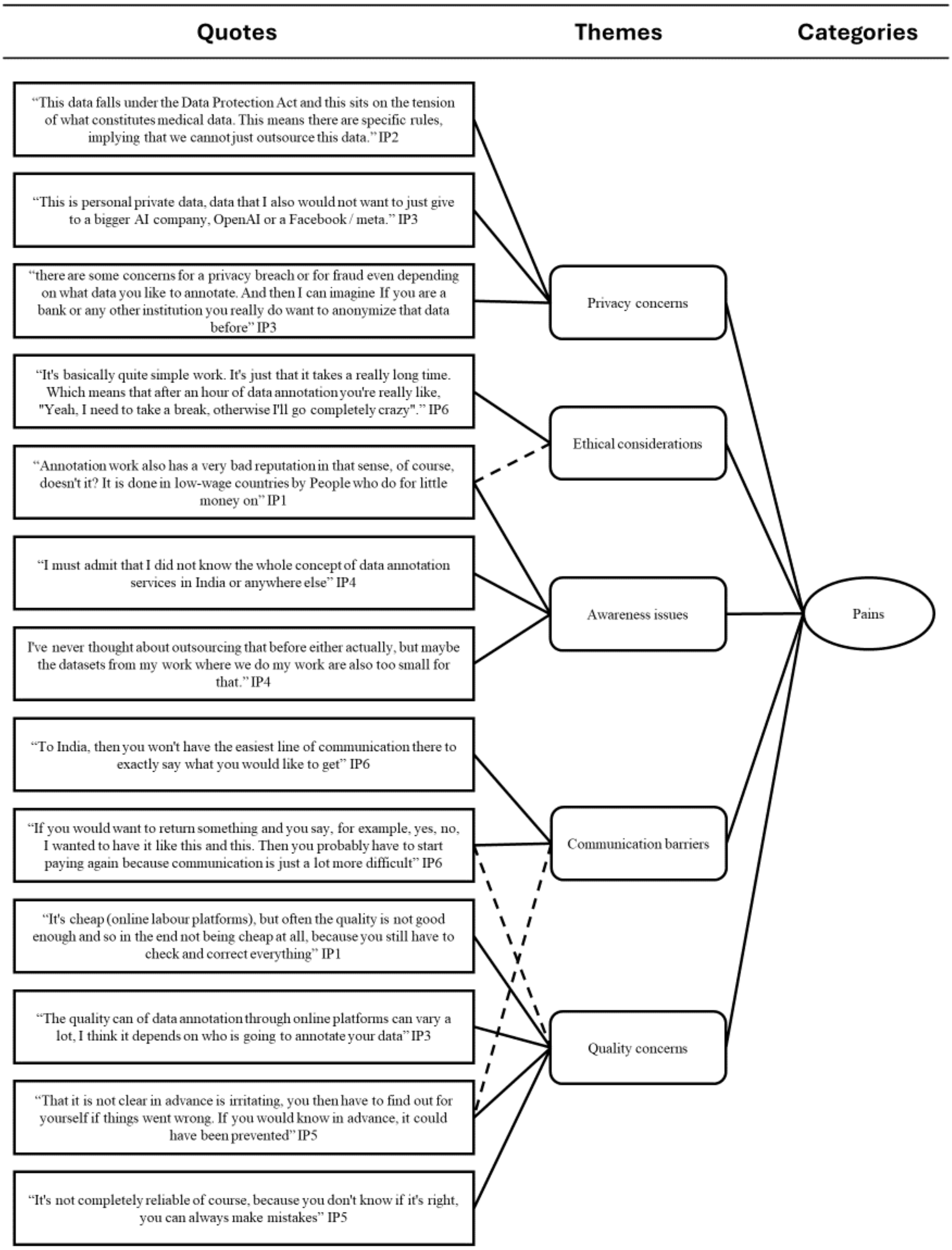
Final Questions:

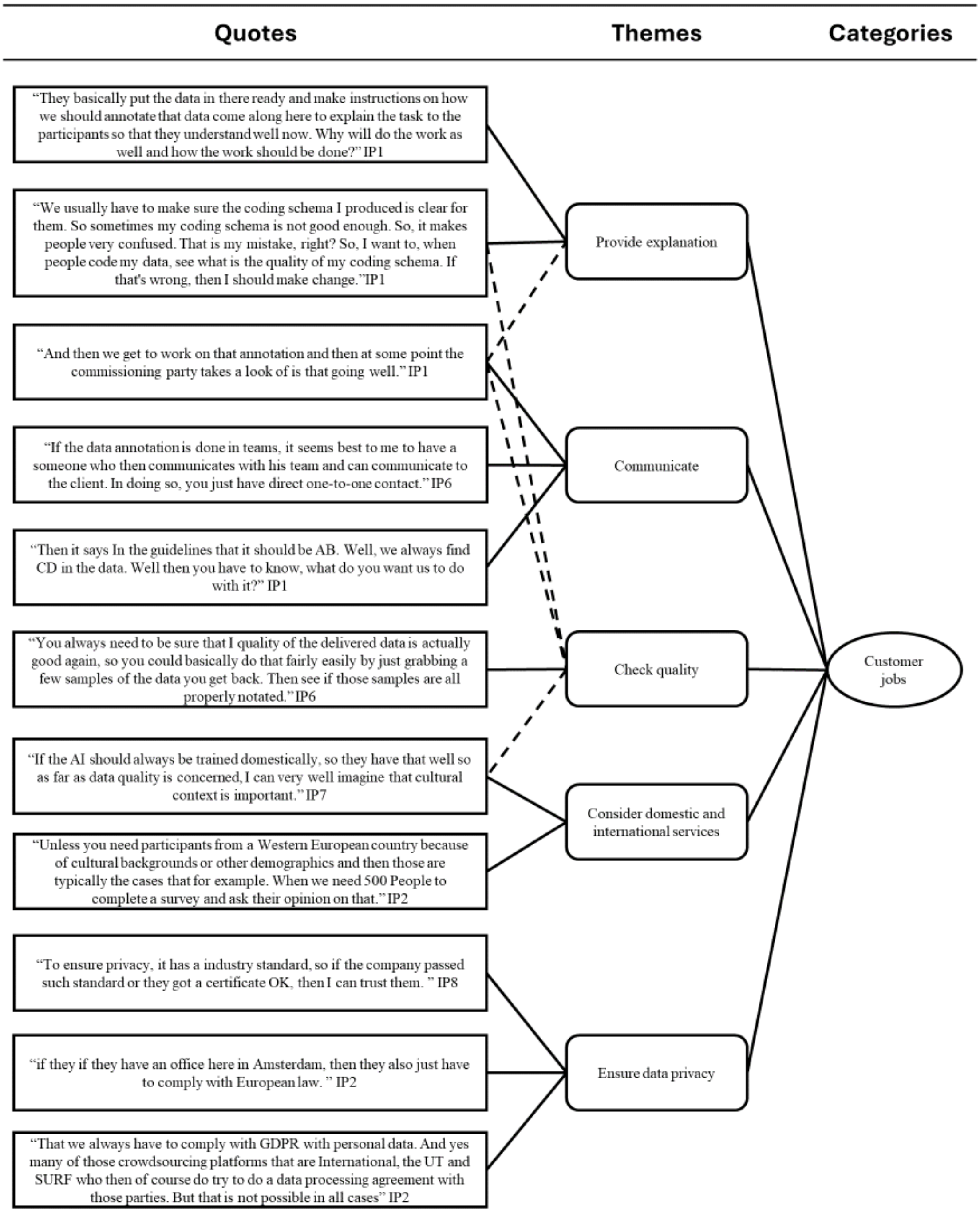
15. *Would you make use of RDA annotation services or the AI AnnotatieLab in Hilversum*
    a. *Could you explain you reasoning? Why yes, why not?*

9. APPENDIX B

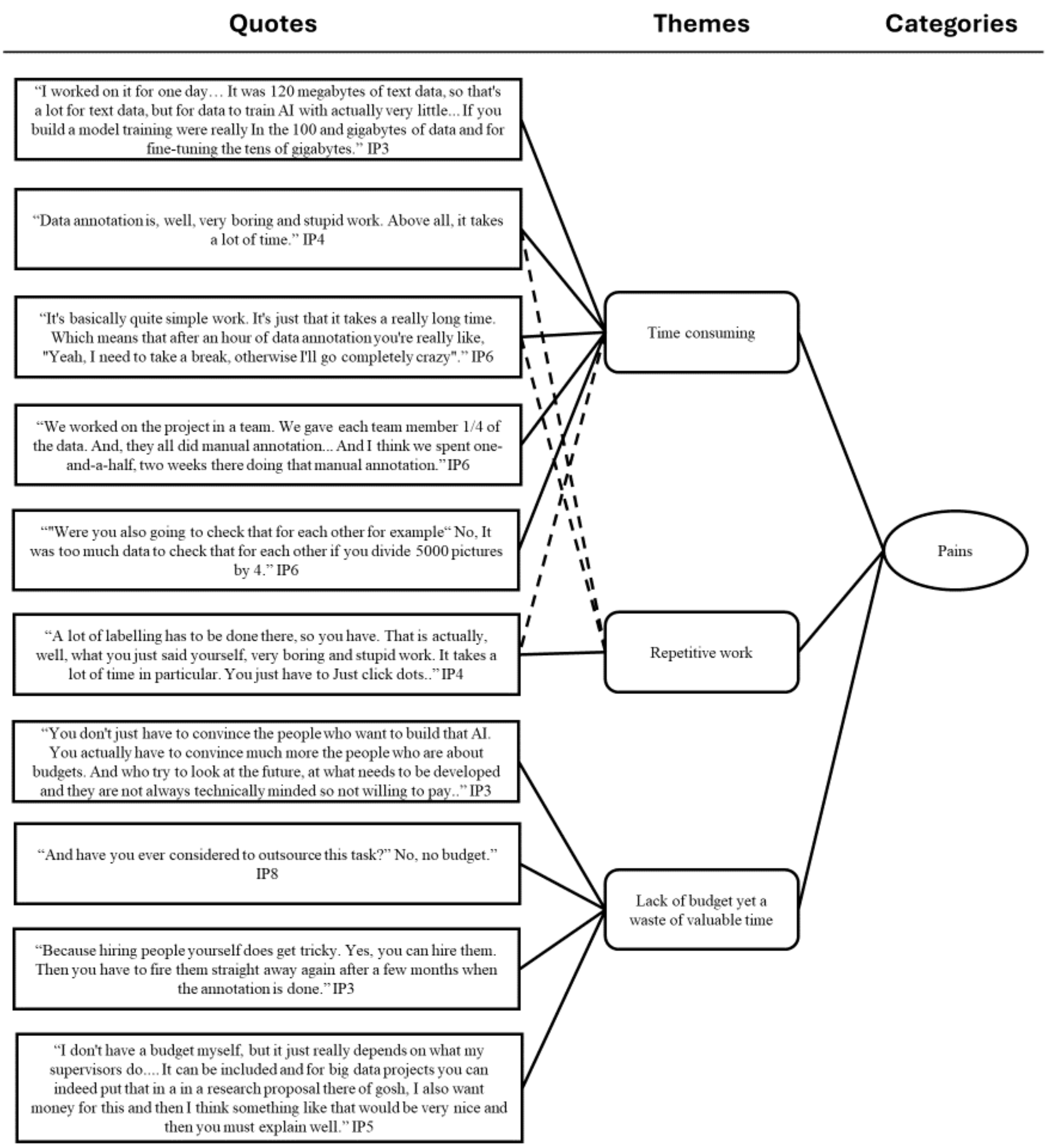
9.1 Outsource to others

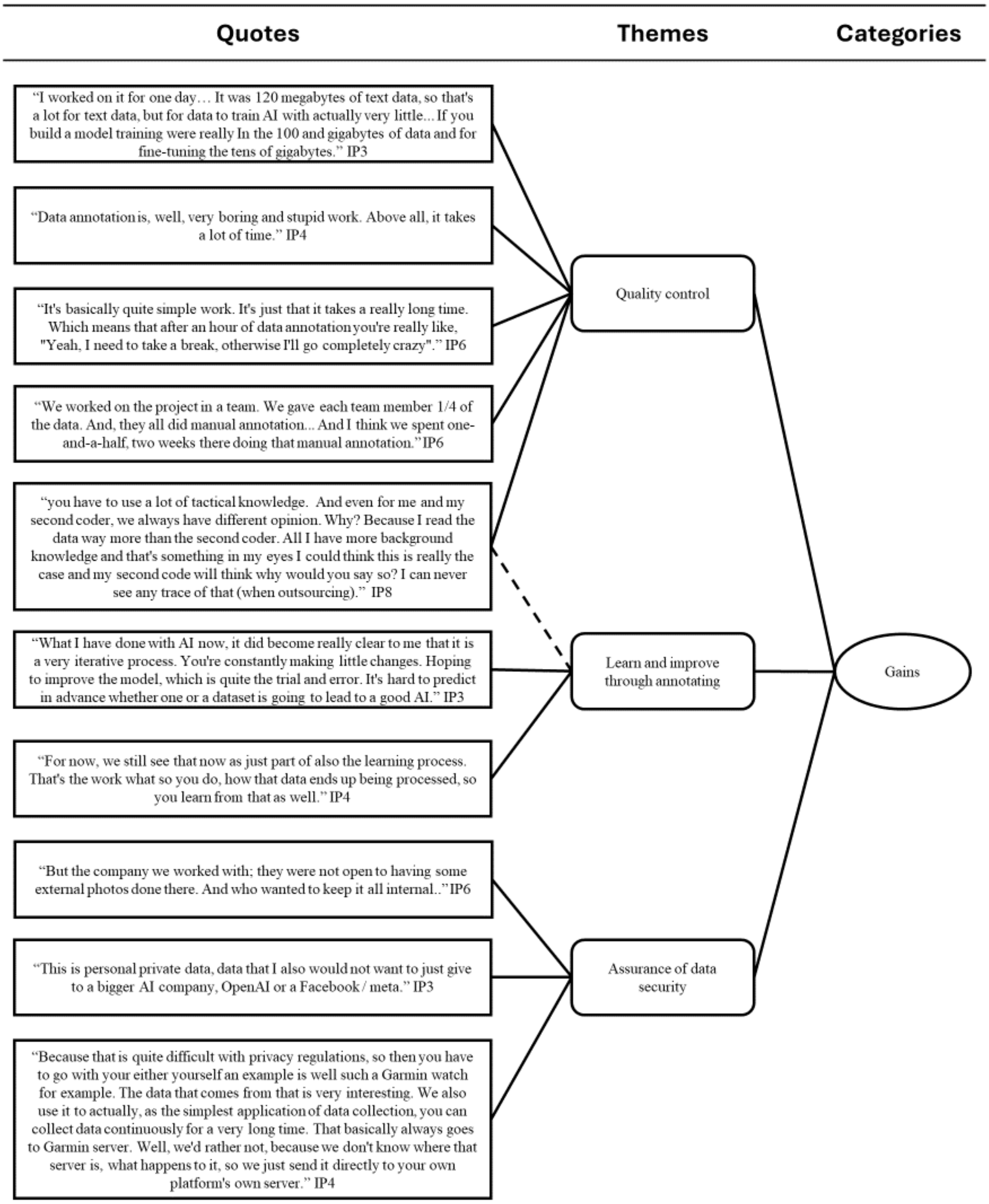


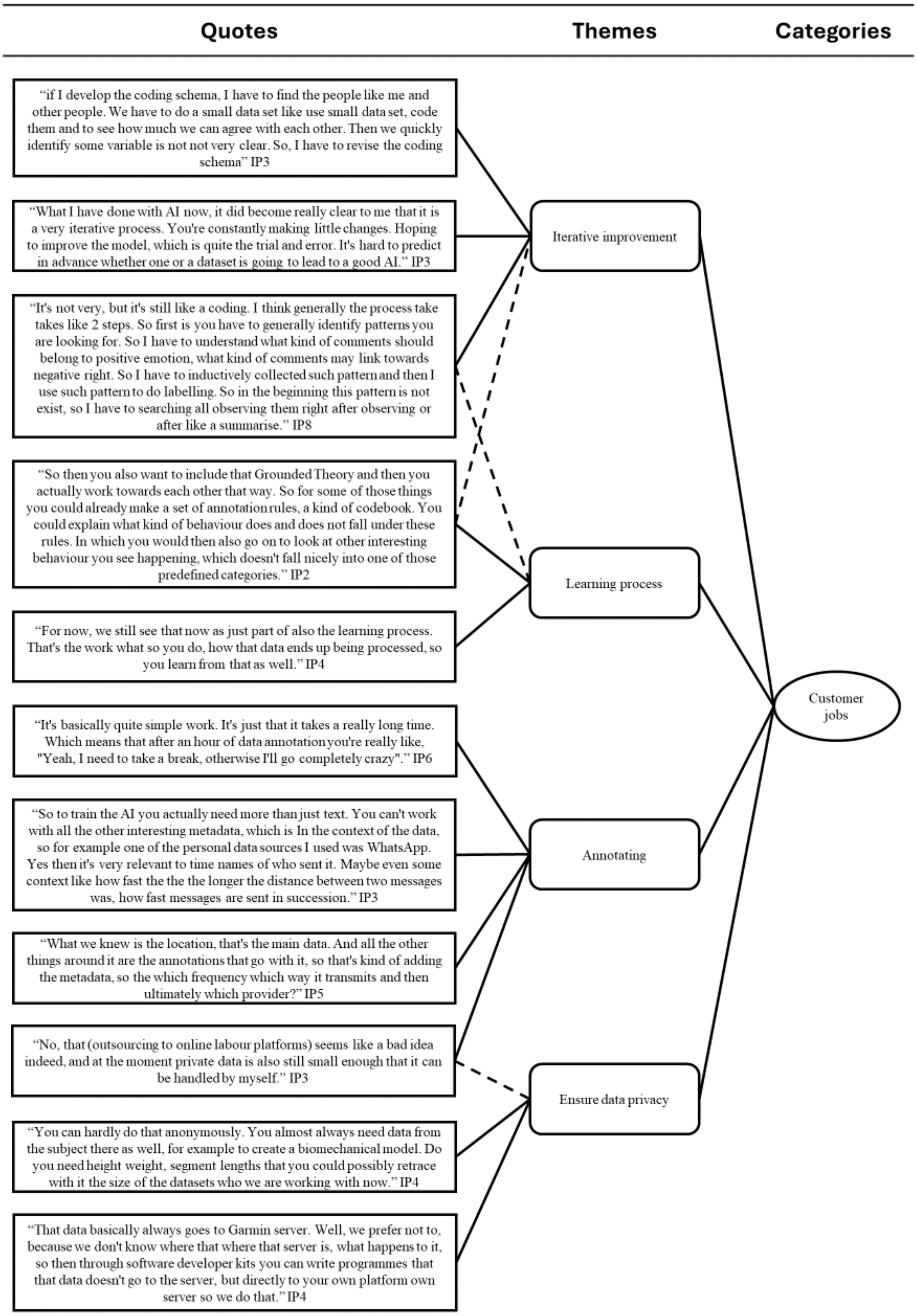

Quotes
Themes
Categories
“This data falls under the Data Protection Act and this sits on the tension of what constitutes medical data. This means there are specific rules, implying that we cannot just outsource this data.” IP2
“This is personal private data, data that I also would not want to just give to a bigger AI company, OpenAI or a Facebook / meta.” IP3
“there are some concerns for a privacy breach or for fraud even depending on what data you like to annotate. And then I can imagine If you are a bank or any other institution you really do want to anonymize that data before” IP3
“It's basically quite simple work. It's just that it takes a really long time. Which means that after an hour of data annotation you're really like, "Yeah, I need to take a break, otherwise I'll go completely crazy".” IP6
“Annotation work also has a very bad reputation in that sense, of course, doesn't it? It is done in low-wage countries by People who do for little money on” IP1
“I must admit that I did not know the whole concept of data annotation services in India or anywhere else” IP4
I've never thought about outsourcing that before either actually, but maybe the datasets from my work where we do my work are also too small for that.” IP4
“To India, then you won't have the easiest line of communication there to exactly say what you would like to get” IP6
“If you would want to return something and you say, for example, yes, no, I wanted to have it like this and this. Then you probably have to start paying again because communication is just a lot more difficult” IP6
“It's cheap (online labour platforms), but often the quality is not good enough and so in the end not being cheap at all, because you still have to check and correct everything” IP1
“The quality can of data annotation through online platforms can vary a lot, I think it depends on who is going to annotate your data” IP3
“That it is not clear in advance is irritating, you then have to find out for yourself if things went wrong. If you would know in advance, it could have been prevented” IP5
“It's not completely reliable of course, because you don't know if it's right, you can always make mistakes” IP5
Privacy concerns
Ethical considerations
Awareness issues
Communication barriers
Quality concerns
Pains


Quotes
Themes
Categories
"They basically put the data in there ready and make instructions on how we should annotate that data come along here to explain the task to the participants so that they understand well now. Why will do the work as well and how the work should be done?" IP1
"We usually have to make sure the coding schema I produced is clear for them. So sometimes my coding schema is not good enough. So, it makes people very confused. That is my mistake, right? So, I want to, when people code my data, see what is the quality of my coding schema. If that's wrong, then I should make change." IP1
"And then we get to work on that annotation and then at some point the commissioning party takes a look of is that going well." IP1
"If the data annotation is done in teams, it seems best to me to have a someone who then communicates with his team and can communicate to the client. In doing so, you just have direct one-to-one contact." IP6
"Then it says In the guidelines that it should be AB. Well, we always find CD in the data. Well then you have to know, what do you want us to do with it?" IP1
"You always need to be sure that I quality of the delivered data is actually good again, so you could basically do that fairly easily by just grabbing a few samples of the data you get back. Then see if those samples are all properly notated." IP6
"If the AI should always be trained domestically, so they have that well so as far as data quality is concerned, I can very well imagine that cultural context is important." IP7
"Unless you need participants from a Western European country because of cultural backgrounds or other demographics and then those are typically the cases that for example. When we need 500 People to complete a survey and ask their opinion on that." IP2
"To ensure privacy, it has a industry standard, so if the company passed such standard or they got a certificate OK, then I can trust them. " IP8
"if they if they have an office here in Amsterdam, then they also just have to comply with European law. " IP2
"That we always have to comply with GDPR with personal data. And yes many of those crowdsourcing platforms that are International, the UT and SURF who then of course do try to do a data processing agreement with those parties. But that is not possible in all cases" IP2
Provide explanation
Communicate
Check quality
Consider domestic and international services
Ensure data privacy
Customer jobs

9.2 Do-it-themselves

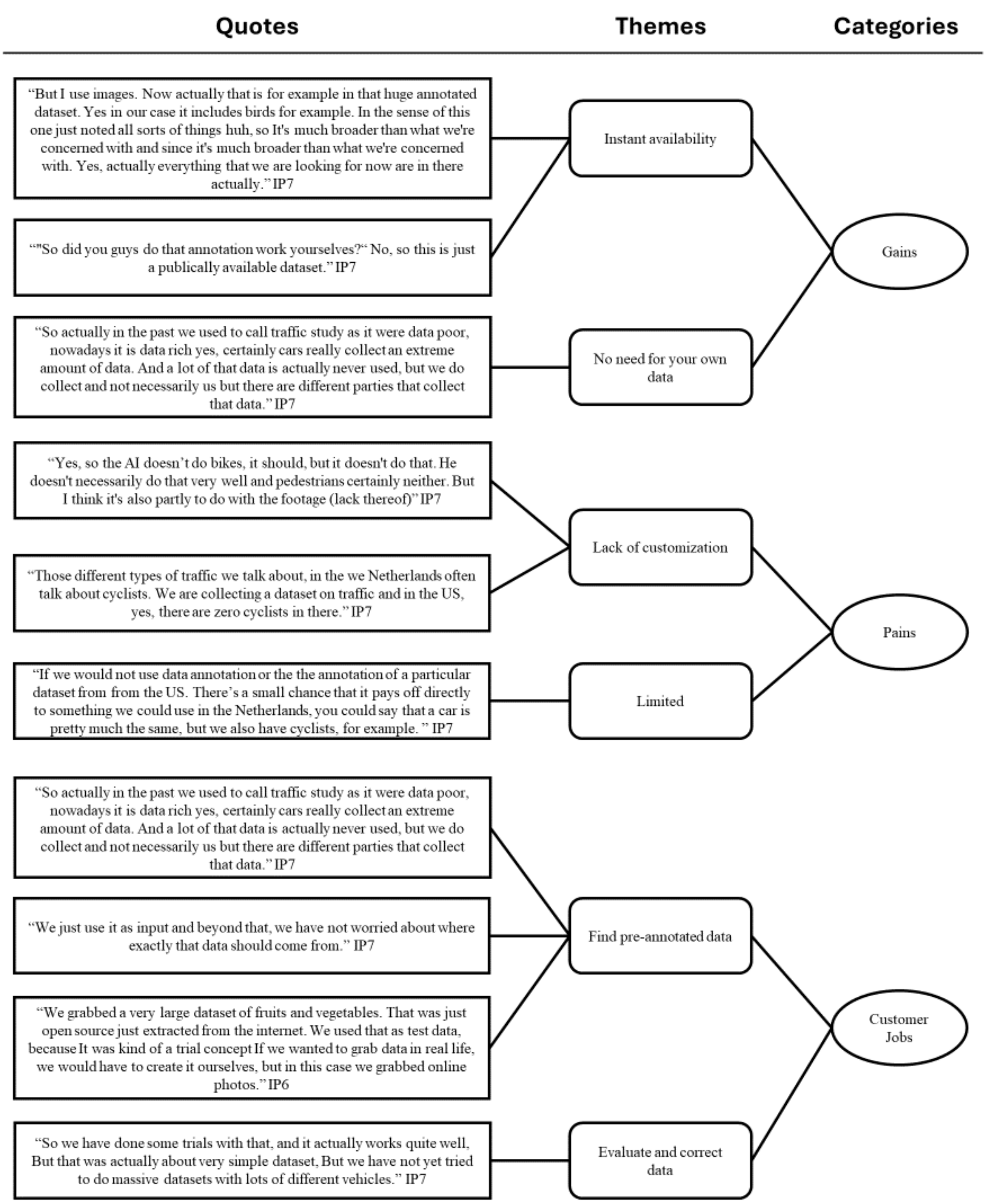



Quotes
Themes
Categories
“I worked on it for one day… It was 120 megabytes of text data, so that's a lot for text data, but for data to train AI with actually very little... If you build a model training were really In the 100 and gigabytes of data and for fine-tuning the tens of gigabytes.” IP3
“Data annotation is, well, very boring and stupid work. Above all, it takes a lot of time.” IP4
“It's basically quite simple work. It's just that it takes a really long time. Which means that after an hour of data annotation you're really like, "Yeah, I need to take a break, otherwise I'll go completely crazy".” IP6
“We worked on the project in a team. We gave each team member 1/4 of the data. And, they all did manual annotation... And I think we spent one-and-a-half, two weeks there doing that manual annotation.” IP6
“you have to use a lot of tactical knowledge. And even for me and my second coder, we always have different opinion. Why? Because I read the data way more than the second coder. All I have more background knowledge and that's something in my eyes I could think this is really the case and my second code will think why would you say so? I can never see any trace of that (when outsourcing).” IP8
“What I have done with AI now, it did become really clear to me that it is a very iterative process. You're constantly making little changes. Hoping to improve the model, which is quite the trial and error. It's hard to predict in advance whether one or a dataset is going to lead to a good AI.” IP3
“For now, we still see that now as just part of also the learning process. That's the work what so you do, how that data ends up being processed, so you learn from that as well.” IP4
“But the company we worked with; they were not open to having some external photos done there. And who wanted to keep it all internal..” IP6
“This is personal private data, data that I also would not want to just give to a bigger AI company, OpenAI or a Facebook / meta.” IP3
“Because that is quite difficult with privacy regulations, so then you have to go with your either yourself an example is well such a Garmin watch for example. The data that comes from that is very interesting. We also use it to actually, as the simplest application of data collection, you can collect data continuously for a very long time. That basically always goes to Garmin server. Well, we'd rather not, because we don't know where that server is, what happens to it, so we just send it directly to your own platform's own server.” IP4
Quality control
Learn and improve through annotating
Assurance of data security
Gains

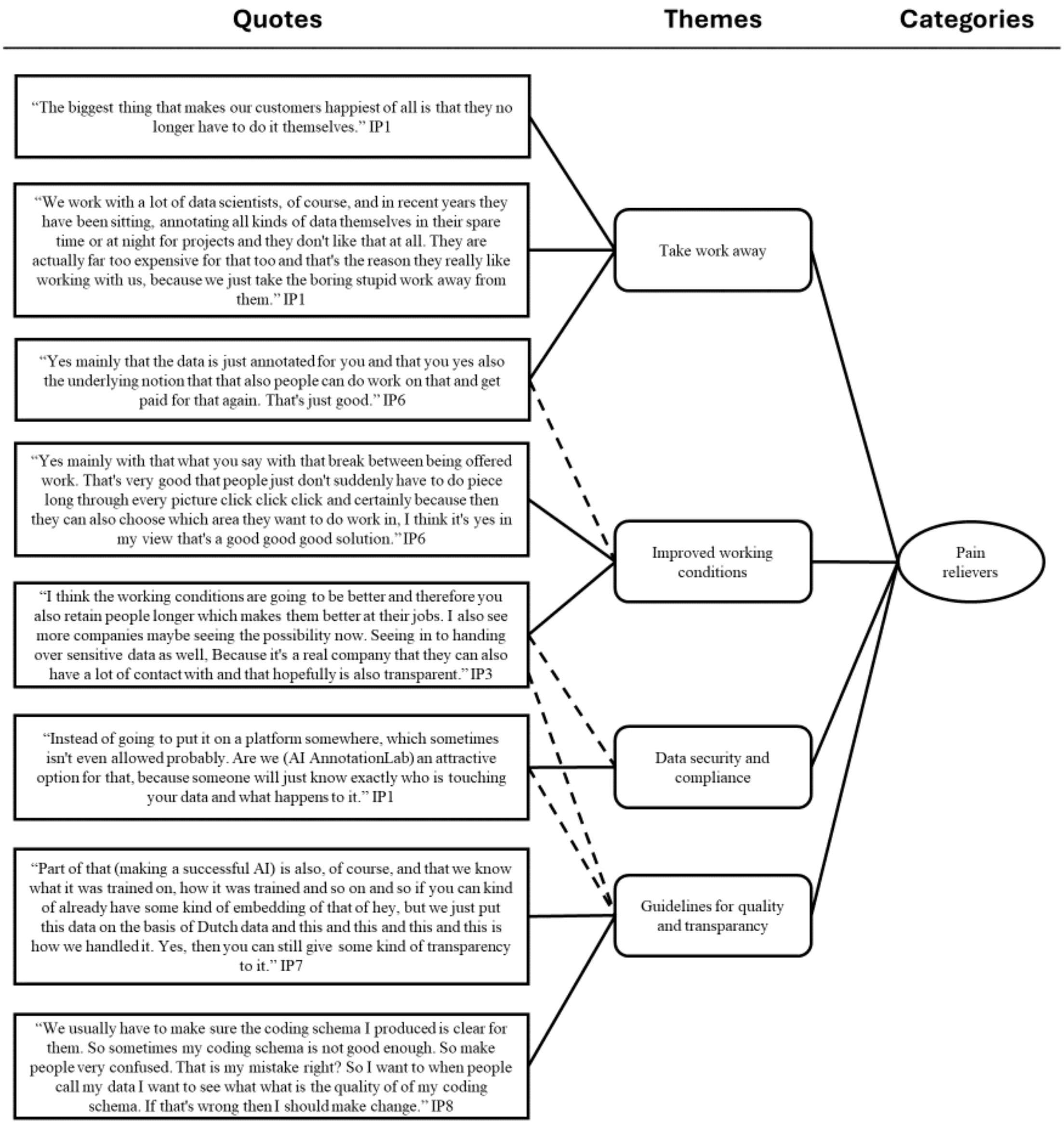


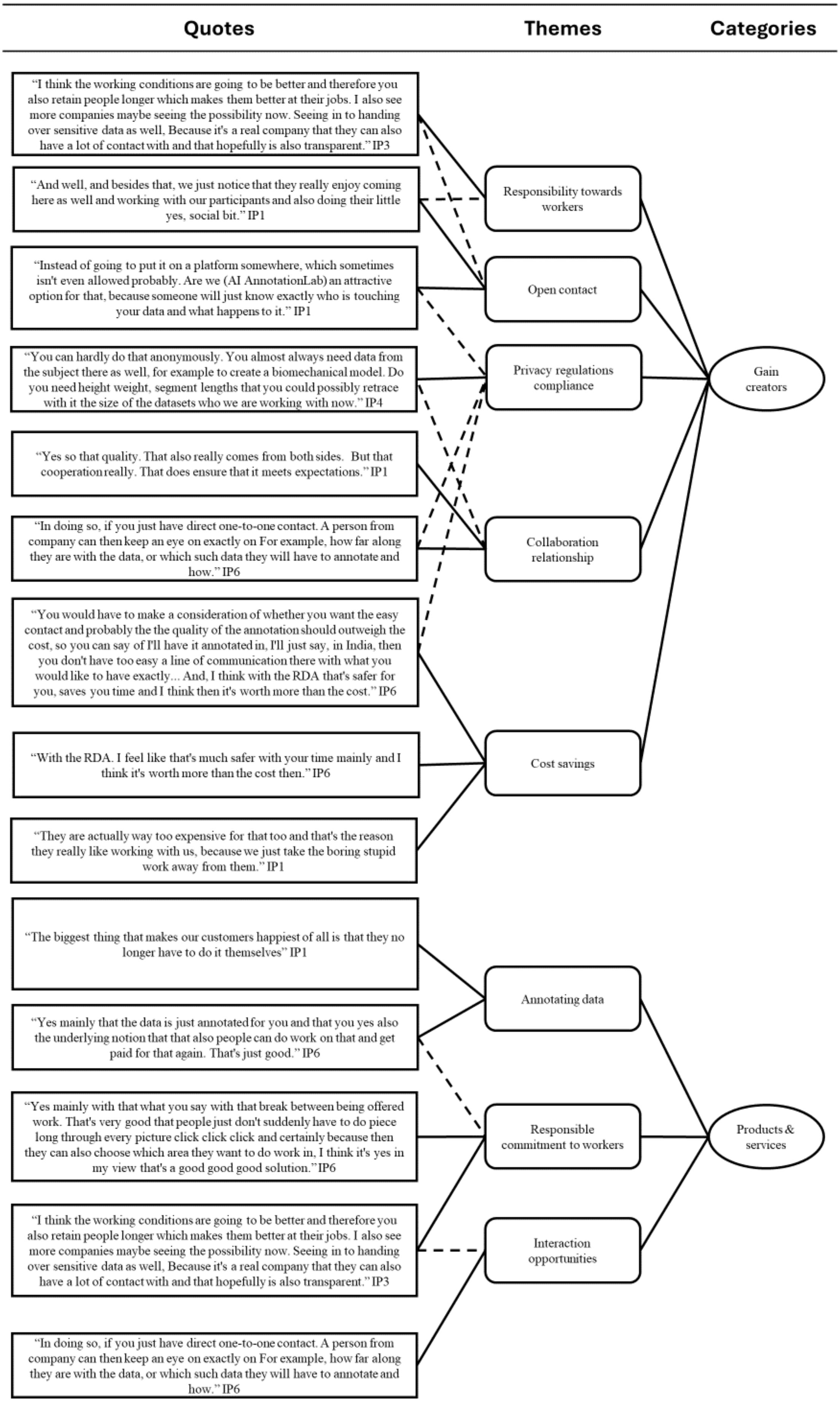
Quotes
Themes
Categories
"if I develop the coding schema, I have to find the people like me and other people. We have to do a small data set like use small data set, code them and to see how much we can agree with each other. Then we quickly identify some variable is not not very clear. So, I have to revise the coding schema" IP3
"What I have done with AI now, it did become really clear to me that it is a very iterative process. You're constantly making little changes. Hoping to improve the model, which is quite the trial and error. It's hard to predict in advance whether one or a dataset is going to lead to a good AI." IP3
"It's not very, but it's still like a coding. I think generally the process take takes like 2 steps. So first is you have to generally identify patterns you are looking for. So I have to understand what kind of comments should belong to positive emotion, what kind of comments may link towards negative right. So I have to inductively collected such pattern and then I use such pattern to do labelling. So in the beginning this pattern is not exist, so I have to searching all observing them right after observing or after like a summarise." IP8
"So then you also want to include that Grounded Theory and then you actually work towards each other that way. So for some of those things you could already make a set of annotation rules, a kind of codebook. You could explain what kind of behaviour does and does not fall under these rules. In which you would then also go on to look at other interesting behaviour you see happening, which doesn't fall nicely into one of those predefined categories." IP2
"For now, we still see that now as just part of also the learning process. That's the work what so you do, how that data ends up being processed, so you learn from that as well." IP4
"It's basically quite simple work. It's just that it takes a really long time. Which means that after an hour of data annotation you're really like, "Yeah, I need to take a break, otherwise I'll go completely crazy"." IP6
"So to train the AI you actually need more than just text. You can't work with all the other interesting metadata, which is In the context of the data, so for example one of the personal data sources I used was WhatsApp. Yes then it's very relevant to time names of who sent it. Maybe even some context like how fast the the the longer the distance between two messages was, how fast messages are sent in succession." IP3
"What we knew is the location, that's the main data. And all the other things around it are the annotations that go with it, so that's kind of adding the metadata, so the which frequency which way it transmits and then ultimately which provider?" IP5
"No, that (outsourcing to online labour platforms) seems like a bad idea indeed, and at the moment private data is also still small enough that it can be handled by myself." IP3
"You can hardly do that anonymously. You almost always need data from the subject there as well, for example to create a biomechanical model. Do you need height weight, segment lengths that you could possibly retrace with it the size of the datasets who we are working with now." IP4
"That data basically always goes to Garmin server. Well, we prefer not to, because we don't know where that where that server is, what happens to it, so then through software developer kits you can write programmes that that data doesn't go to the server, but directly to your own platform own server so we do that." IP4
Iterative improvement
Learning process
Annotating
Ensure data privacy
Customer jobs

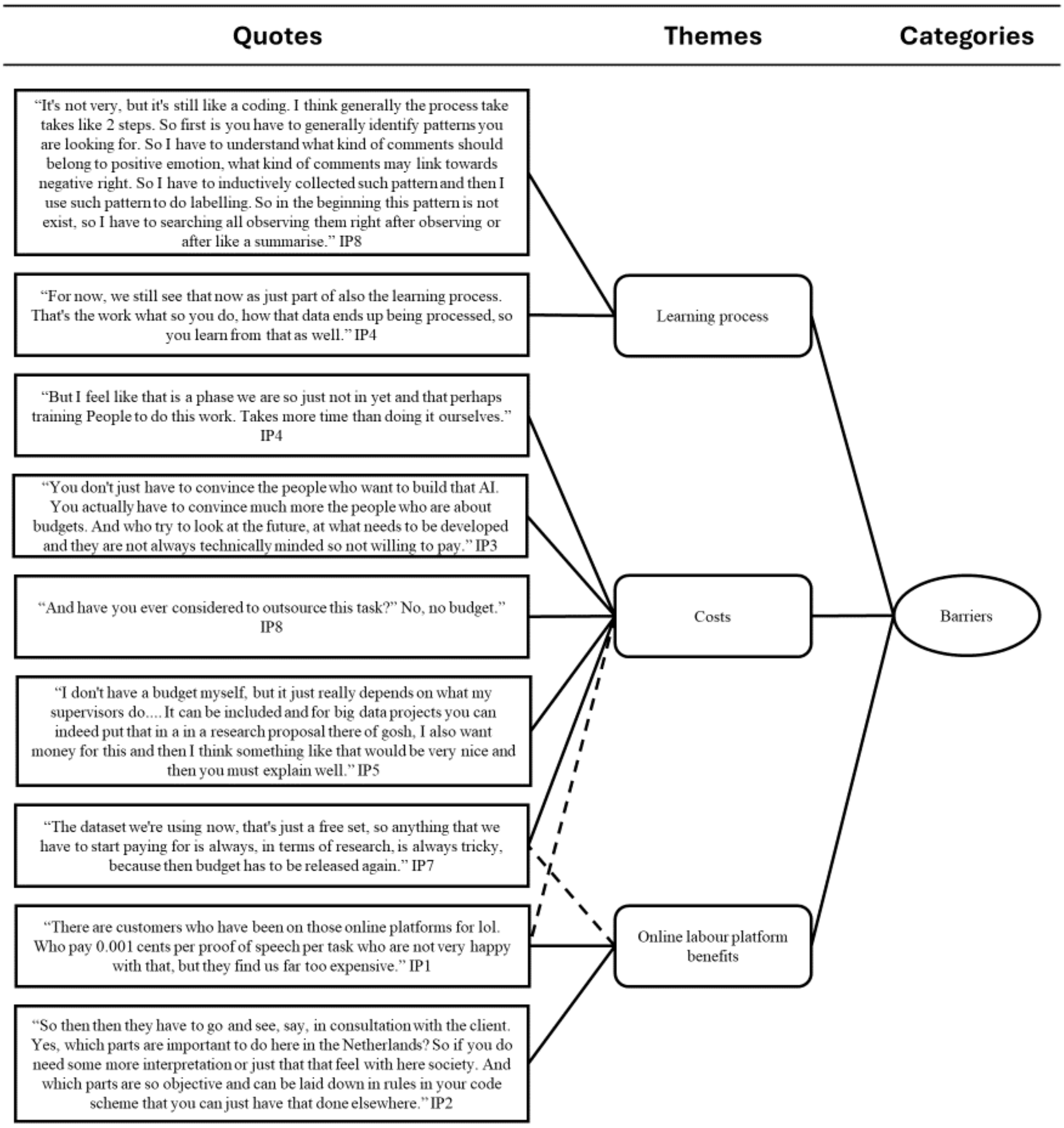
9.3 Pre-annotated data



9.4 Value proposition




Quotes
Themes
Categories
"I think the working conditions are going to be better and therefore you also retain people longer which makes them better at their jobs. I also see more companies maybe seeing the possibility now. Seeing in to handing over sensitive data as well, Because it's a real company that they can also have a lot of contact with and that hopefully is also transparent." IP3
"And well, and besides that, we just notice that they really enjoy coming here as well and working with our participants and also doing their little yes, social bit." IP1
"Instead of going to put it on a platform somewhere, which sometimes isn't even allowed probably. Are we (AI AnnotationLab) an attractive option for that, because someone will just know exactly who is touching your data and what happens to it." IP1
"You can hardly do that anonymously. You almost always need data from the subject there as well, for example to create a biomechanical model. Do you need height weight, segment lengths that you could possibly retrace with it the size of the datasets who we are working with now." IP4
"Yes so that quality. That also really comes from both sides. But that cooperation really. That does ensure that it meets expectations." IP1
"In doing so, if you just have direct one-to-one contact. A person from company can then keep an eye on exactly on For example, how far along they are with the data, or which such data they will have to annotate and how." IP6
"You would have to make a consideration of whether you want the easy contact and probably the the quality of the annotation should outweigh the cost, so you can say of I'll have it annotated in, I'll just say, in India, then you don't have too easy a line of communication there with what you would like to have exactly... And, I think with the RDA that's safer for you, saves you time and I think then it's worth more than the cost." IP6
"With the RDA. I feel like that's much safer with your time mainly and I think it's worth more than the cost then." IP6
"They are actually way too expensive for that too and that's the reason they really like working with us, because we just take the boring stupid work away from them." IP1
Responsibility towards workers
Open contact
Privacy regulations compliance
Collaboration relationship
Cost savings
Gain creators
"The biggest thing that makes our customers happiest of all is that they no longer have to do it themselves" IP1
"Yes mainly that the data is just annotated for you and that you yes also the underlying notion that that also people can do work on that and get paid for that again. That's just good." IP6
"Yes mainly with that what you say with that break between being offered work. That's very good that people just don't suddenly have to do piece long through every picture click click click and certainly because then they can also choose which area they want to do work in, I think it's yes in my view that's a good good good solution." IP6
"I think the working conditions are going to be better and therefore you also retain people longer which makes them better at their jobs. I also see more companies maybe seeing the possibility now. Seeing in to handing over sensitive data as well, Because it's a real company that they can also have a lot of contact with and that hopefully is also transparent." IP3
"In doing so, if you just have direct one-to-one contact. A person from company can then keep an eye on exactly on For example, how far along they are with the data, or which such data they will have to annotate and how." IP6
Annotating data
Responsible commitment to workers
Interaction opportunities
Products & services


Quotes
Themes
Categories
"It's not very, but it's still like a coding. I think generally the process take takes like 2 steps. So first is you have to generally identify patterns you are looking for. So I have to understand what kind of comments should belong to positive emotion, what kind of comments may link towards negative right. So I have to inductively collected such pattern and then I use such pattern to do labelling. So in the beginning this pattern is not exist, so I have to searching all observing them right after observing or after like a summarise." IP8
"For now, we still see that now as just part of also the learning process. That's the work what so you do, how that data ends up being processed, so you learn from that as well." IP4
"But I feel like that is a phase we are so just not in yet and that perhaps training People to do this work. Takes more time than doing it ourselves." IP4
"You don't just have to convince the people who want to build that AI. You actually have to convince much more the people who are about budgets. And who try to look at the future, at what needs to be developed and they are not always technically minded so not willing to pay." IP3
"And have you ever considered to outsource this task?" No, no budget." IP8
"I don't have a budget myself, but it just really depends on what my supervisors do.... It can be included and for big data projects you can indeed put that in a in a research proposal there of gosh, I also want money for this and then I think something like that would be very nice and then you must explain well." IP5
"The dataset we're using now, that's just a free set, so anything that we have to start paying for is always, in terms of research, is always tricky, because then budget has to be released again." IP7
"There are customers who have been on those online platforms for lol. Who pay 0.001 cents per proof of speech per task who are not very happy with that, but they find us far too expensive." IP1
"So then then they have to go and see, say, in consultation with the client. Yes, which parts are important to do here in the Netherlands? So if you do need some more interpretation or just that that feel with here society. And which parts are so objective and can be laid down in rules in your code scheme that you can just have that done elsewhere." IP2
Learning process
Costs
Online labour platform benefits
Barriers