universität wien

# MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

## Taming Crowds: How practitioners address uncertainty when crowdsourcing data sets


verfasst von / submitted by

Dipl.-Ing. Fabian Fischer, BSc


angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2021 / Vienna 2021

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | UA 066 906 |
| Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: | Masterstudium Science-Technology-Society |
| Betreut von / Supervisor: | Univ.-Prof. Dr. Ulrike Felt |

# Taming crowds

**How practitioners address uncertainty when crowdsourcing data sets**

Fabian Fischer

## Acknowledgements

This thesis investigates a topic that was my motivation to start my STS studies. As such, it has been accompanying me for several years. I lost sight of the thesis several times, which made completing it a challenge and, at times, an arduous endeavour. However, I am glad that I persisted as it was a process full of learning opportunities – about STS, the topic of research, writing and myself. This thesis would not be possible without support from many sides.

First, this thesis would not exist without the practitioners who were willing to participate in my research. I am grateful for the insights they provided and the openness shown during the interviews. Thank you.

I want to thank my supervisor Ulrike Felt for at crucial times giving me the right amount of push to make necessary progress and overcome my inner barriers and for providing support in the height of a COVID-19 summer as I wrapped up this thesis.

Some of my peer students were instrumental for me to finish my STS studies and this thesis. Anna, David, Ruth, and Tobi, thank you for a great start into STS. Together with Melanie and Stefan, you have been accompanying my studies and remained important contact points to talk about the highs and lows of studying and STS. Veronika and Sophie, you provided vital, regular support that was important for me to keep going towards the final phase of this thesis.

Finally, I want to thank Xandi, who had to endure me during my years as an STS student while also working for a living. This meant putting up with me being buried in texts, in front of my computer, in class. Often it was endless days and weeks where I was not always the most pleasant company and where my spare time was often limited. You repeatedly reminded me of life outside work and study and provided support throughout the years – thank you!

# Contents

# 1. Introduction

As I write this thesis, the global COVID-19 pandemic has been a catalyst to accelerate the adoption and use of digital technologies in all areas of life. In times when physical contact between humans has become problematic, they have become a panacea to keep our societies working while being physically distanced. Prior to this forced push of digital transformation, the advent and widespread adoption of the internet in the early years of the new millennium and of smartphones since 2007 brought massive shifts in the ubiquity of digital technology and communication. As Vertesi and Ribes put it, "[m]icrochips are no longer confined to hefty machines in the corner of the laboratory or even the workplace desktop: they are in our homes, our pockets, our clothing, sometimes under our skin" (2019, p. 1).

In the past decade, tremendous progress has been made in the area of so-called 'Artificial Intelligence' ('AI'). There has been considerable hype around 'AI', being advertised as powering television sets, personal assistants on smartphones and their speech recognition, interaction with lights and other parts of 'smart' homes, biometric applications (fingerprint, iris, and face recognition), language translation, and 'self-driving' cars.

With the successes of 'AI' have come concerns. Concerns over chatbots turning into misogynist Neo-nazis overnight (Metz, 2016), face recognition systems being plagued by racism and sexism (Buolamwini & Gebru, 2018), 'self-driving' cars killing pedestrians (Wakabayashi, 2018) and drivers (Lee, 2010), targeting of voters based on their alleged voting preference and other information inferred from their behaviour on the social networking site Facebook (Scott, 2018), and gender stereotyping machine translations (Zou & Schiebinger, 2018). All these concerns point towards the question of how it is possible that this new, advanced technology seems to adopt the dark side(s) of our societies. Data sets play a crucial role in this, and studying their production can contribute to understanding these phenomena. This has received little attention compared to the research on the consequences of data-driven technologies (Bechmann & Bowker, 2019).

But first, dear reader, let me clarify what I understand as 'AI' for the scope of this thesis. I characterise the core technologies of the mentioned types of 'Artificial Intelligence' as comprising knowledge bases, logic, and machine learning – knowing that the term 'AI' is contested and that not all things labelled 'AI' actually use any of these. Knowledge bases are structured repositories of 'facts', e.g. that a penguin is a bird. Logic allows reasoning over knowledge, e.g. that birds lay eggs and, consequently, so do penguins.[1] Machine learning aims to 'learn' how to solve a

[1] This is a common example for reasoning over knowledge bases, as penguins cannot fly, unlike most other birds. Which already hints at one of the core issues with this kind of knowledge base.

problem without being told explicitly how to do so, but by being 'trained' by data or feedback. It is especially machine learning that has seen tremendous progress in the last decade. This is due to the breakthrough of deep learning in 2012, a then-novel approach to machine learning. It significantly outperformed previous attempts at automated image recognition (Krizhevsky, Sutskever, & Hinton, 2012). Since then, it has been applied to a multitude of problems – most of the examples mentioned above are likely to be powered by some form of deep learning.

Machine learning itself can be differentiated further. Reinforcement learning is a form of machine learning where positive and negative feedback is provided to the learning system that then figures out on its own what is the 'right' way to do things. A popular example is automated gameplay, where the machine learning system never gets introduced to the game's rules but has to figure them out itself. Supervised machine learning can be described as learning by example: Based on a set of training data that is adequately described, the machine learning algorithm learns how to identify, e.g., what kind of pet is shown on a photograph. Unsupervised machine learning, on the other hand, has no clear example to learn from. Instead, it has to figure out on its own that a data set contains, e.g., images of two different types of pets (such as dogs and cats).

Both supervised and unsupervised machine learning rely on data to learn from, and many of the machine learning approaches benefit from large amounts of data. This was another reason why many forms of machine learning have become increasingly competent at solving problems: The amount of data available has been increasing rapidly as more and more aspects of our lives are being mediated, recorded and measured by digital technologies.

This brings me back to the aforementioned concerns. One explanation of 'misbehaving' 'AI' is the data it gets trained on: Of course, confronted with misogynist and Nazi content, a chatbot will learn from this bad influence (Metz, 2016). If German documents usually record doctors using the generic masculine but nurses frequently with the female-specific term 'Krankenschwester', then it's no surprise that a machine will adopt this stereotype when translating from a language with gender-neutral nouns (Kolly & Schmid, 2021). The lack of data in some areas also became a concern: Pedestrian crossings of highways are so rare it's difficult to collect data for this kind of event. And the sheer amount of data has increasingly turned problematic, too, as, e.g., Facebook got criticised for the amount of data it collects and sells about its users' behaviour (BBC News, 2018).

Against this background, how data sets get produced warrants attention. Recently, scholars have been calling for more thorough investigation (Bechmann & Bowker, 2019; Jaton, 2021b) and documentation (Gebru et al., 2020) of how data sets are produced. One important aspect of

data set production, especially for supervised machine learning, is the labelling and annotating of data. While these two terms frequently get used interchangeably, I prefer to differentiate them as follows: Labelling assigns labels (effectively categories) to data points (images, documents, texts, words, etc.), whereas annotation selects parts of a data point (a bounding box around a face, nouns in a sentence, pedestrians in a photograph, etc.). Sometimes labels then also get added to these annotations.

One common approach to get data sets labelled and annotated is crowdsourcing. You, the reader, probably encountered many times that websites ask you to pick those images out of several that show school buses, traffic lights or cats, similar to figure 1. This system is called reCAPTCHA[2]. On the one hand, this is used to check if you are a human and not some form of automated access to the website, e.g., by a so-called bot. But on the other hand, it is also a mechanism to generate huge data sets containing labelled images for machine learning. Each time you pick these images, you help guide some machine learning system to identify school buses, traffic lights or cats.


Figure 1: This is an example of reCAPTCHA, a method of collecting labelled images, in this instance of cat images. Source: https://en.wikipedia.org/wiki/ReCAPTCHA

Crowdsourcing is a well-established approach for the labelling and annotation of data (Geiger et al., 2020). It means, at its core, that many people participate in labelling parts of the data, the so-called 'crowd'. Often, these people are not trained in any particular way, such as visitors of arbitrary websites that use reCAPTCHA. Instead, each data point is usually presented to several people who each label it independently. The sheer number of people is expected to compensate for the lack of expertise for the application by cleverly integrating these multiple labels and annotations to generate what the machine learning system then learns from. As I will show, this process is fraught with uncertainties, among them concerns over the 'crowd' that often gets rendered problematic (Doan,

[2]CAPTCHA is an acronym for Completely Automated Public Turing test to tell Computers and Humans Apart. A Turing test, in turn, being a test of a machine if it can imitate human behaviour so well that humans cannot distinguish it from real human behaviour any more.

Ramakrishnan, & Halevy, 2011; Oleson et al., 2011), how to create suitable 'tasks' to distribute to the workers and how often each data point should be annotated by the 'crowd'. Consequently, this study asks the question *how do practitioners that crowd source data sets address uncertainties during this process.*

In addition to the ongoing debates about data and 'AI', there is also a very personal motivation why I decided to ask this question. Before I started my endeavour as a social scientist, I was trained as a computer scientist, and I worked several years on projects that could be put under the banner of 'AI'. My first encounter with crowdsourcing was early during my professional career as a computer scientist. I was working on so-called natural language processing problems, i.e. attempts to make a computer understand text written by humans (at least to some extent). One of the problems we've been tackling at the time has been 'sentiment analysis'. This research area aims to identify if a piece of text conveys positive, neutral, or negative sentiment. One widely-established source to train and evaluate sentiment analysis algorithms was customer reviews on websites with the ubiquitous star rating mapped to a positive–negative spectrum. In our case back then, this was limiting, as the texts we were analysing were news and social media texts. Hence we turned to crowdsourcing to get a usable data set that was relevant to our application scenario.

In one case, my colleagues and I attempted to take the context of an individual sentence into account. We were aware that sentences are contextual, and properly identifying the intended emotion without an understanding of this context would be difficult. To get an idea of how much context would be required, we varied the amount of provided context. After the crowdsourcing, we encountered difficulties using the resulting data: Too frequently, the crowd workers' answers varied a lot, making the data set rich in ambiguities and contradictions (some workers labelling a data point as negative, whereas others labelled the same data point as positive). We were confronted with how to make use of this kind of data. In the end, we failed to resolve the uncertainty that resulted from these ambiguities, and budgetary constraints prevented investing in further research.

During that time, I was also increasingly wary of the attempt to establish an 'objective' 'ground truth' about the conveyed emotion of a text, as I deemed it frequently subjective and context-dependent in more ways than we addressed: A text could mean different things to the author and the audience, the author's intention can be ironic or sarcastic, etc. Consequently, I lost interest in this particular kind of research. Instead, I increasingly saw this kind of work as problematic. I saw crowdsourcing attempts as frequently adhering to a universalistic worldview, where there's one correct answer, and it's only an issue to identify it.

This personal experience and uneasiness with crowdsourcing eventually lead to the master thesis you are reading right now. I wanted to investigate how other practitioners do this kind of work and with what kind of assumptions they approach crowdsourcing. And I wanted to move beyond my anecdotal, personal experience and do proper empirical work to better understand the phenomenon of crowdsourcing in practice.

As you may have noticed, I use the notions of 'practitioners' and 'in practice'. I could have studied academic debates in computer science, manuals and how-to guides of crowdsourcing platforms. However, drawing on an early contribution of Science & Technology Studies (STS), academic publications are polished, purified accounts of events that are, in practice, in the 'lab', considerably more messy and contingent (Latour, Woolgar, & Salk, 1986; Jaton, 2021b). Manuals are similarly problematic, as they describe how things *should be done*, not how they are done in practice. Hence, they are normative and idealised accounts (Suchman, 1995).

Coming back to the topic of 'AI', crowdsourcing is one important way to obtain data for machine learning, and I approached crowdsourcing from this direction. However, I encountered several obstacles during my recruitment efforts: Most of my respondents had no immediate plans to use the data for machine learning. Some of them showed interest in using the data for machine learning sometime in the future. Some respondents work in a field where machine learning is the main focus, and crowdsourcing is one way of obtaining the data. But crowdsourcing is also a research niche in and of itself. One respondent used crowdsourcing for machine learning twice, and all shared the goal of using crowdsourcing to create data sets with useful labels and annotations. Consequently, I deem it a non-issue that there were no clear plans for 'AI' applications: The issue of uncertainty arises regardless of future uses for machine learning. In order to make the data set in any way meaningful, these issues have to be addressed by the crowdsourcing practitioners.

As implied above, I chose to conduct interviews with crowdsourcing practitioners. These qualitative interviews, five overall, were done in a semi-structured way. I chose the type of questions to evoke ethnographic accounts of the respondents' actual practices (Spradley, 2002). After transcribing the interviews, I used Situational Analysis (Clarke, 2005) to analyse the material.

As a theoretical framework, I draw on Actor-Network Theory (ANT), with elements of both 'classic' ANT from the 1980s and more recent developments in ANT. These works provide sensitivities and concepts that are particularly well suited to study complex, socio-material networks, such as crowdsourcing, that involve both humans (my respondents, experts and crowd workers) as well as non-humans (data, algorithms, crowdsourcing platforms).

In the remainder of this thesis, I will first situate my research in relevant related literature, covering studies of data, data sets and 'AI', crowdsourcing, and (coping with) uncertainty. In section 3, I will then provide more details about Actor-Network Theory as my theoretical framework and the sensitising concepts relevant for my study. Section 4 will present the research questions that guided my study. Following this, in section 5 I will describe in detail my empirical material and the methods used to analyse it. Section 6 will then present the empirical results obtained from my interviews before discussing them more broadly in section 7. Finally, a conclusion will round off this thesis.

## 2. State of the art

This thesis touches on several strands of research. First, uncertainty and how people cope with it are closely linked to my thesis and literature on this topic informs my analysis. Second, as I am studying the production of data sets, studies of data and, more specifically, data sets are related. Because data sets are a crucial part that enables 'AI', studies of 'AI', especially those that focus on the underlying data, are of interest, too. Third, crowdsourcing is likewise a relevant field of research.

### 2.1. Dealing with uncertainty

Uncertainty and ways to cope with it have a rich history in STS as a research topic. One of the early accomplishments of STS was to document how scientific practice is much messier, much more conditional than it is often portrayed in (scientific) publications and presentations to a wider public (Latour et al., 1986). STS scholars have repeatedly argued how newly created knowledges do not lead to closure and certainty. On the contrary, the increasing scientification of society has contributed to an increase in complexity and an increase in uncertainty (Nowotny, Scott, & Gibbons, 2008). Consequently, "[t]he generation of uncertainties is as inherent to, and endemic in, research as it is to contemporary life" (Nowotny et al., 2008, p. 37). These uncertainties can be identified as positive or negative – a decision which is in itself uncertain and volatile. As Callon, Lascoumes, & Barthe argue, the positive attitude towards uncertainties is to recognize them as "a starting point for an exploration intended to transform and enrich the world in which we decide to live" (2009, p. 257), or as Nowotny et al. phrase it, a positive attitude encourages experimentation and risk-taking. However, the negative attitude sees uncertainties as threatening and aims to eliminate or at least reduce them.

How people and institutions cope with uncertainties is neither random nor entirely contingent (Nowotny et al., 2008). Some coping strategies attempt to spread the uncertainties spatially across personal, professional, and personal lives and temporally across life spans. On the one hand, spreading uncertainties can reduce their threatening effect, e.g., through assurance mechanisms and other forms of solidarity. However, these mechanisms are under stress in an era of increased individualisation, leading to internalised coping methods. In the context of professional careers, one way of doing this is to live 'portfolio' biographies, i.e. the accumulation of qualifications, experiences, occupations and training, to be prepared for uncertain and changing future income opportunities. A different form of spreading, however, has to be kept in check: By establishing "appropriately structured environments", local uncertainties are being prevented from becoming global (Nowotny et al., 2008, p. 37).

An additional aspect of uncertainty and techno-scientific research and development is relevant to my research. MacKenzie (1998) identifies a pattern that indicates that people estimate the uncertainty of certain findings and procedures differently based on their social distance to the actual practices, a phenomenon he called 'certainty trough'. He identifies three different groups. First, practitioners directly involved in knowledge production are well aware of the limits of their knowledge, the contingencies involved in the production of 'facts' and consequently, the number of uncertainties abound. The second group are 'outsider' critics who question broadly the institutions where the knowledge gets produced and consequently doubt the knowledge itself. Third, and this is the trough part, are users loyal to the institution but not themselves involved in the knowledge production. They dismiss the broad scepticism of the 'outsiders' but are not aware of the limitations aware to the actual producers of the knowledge. Thus, how close people are to the actual practice impacts how they estimate the uncertainties entangled with these practices and the knowledges they produce.

A particular way of coping with uncertainties is the quest to make uncertainties calculable (Hacking, 1990; Lupton, 1999; Nowotny et al., 2008). In modernity, uncertainty turned into chance, which is the foundation of risk: The indeterminacy of uncertainties was conquered by estimating the likelihood of events, making the uncertain known or at least knowable. Through methods of calculability, chaos should be tamed, and an indeterminate world should become manageable. With the invention of risk, the term 'uncertainty' got a different connotation, being left to those events that resisted calculation. Consequently, their probabilities could not be estimated or were unknown. The distinction between risk and uncertainty was muddied towards the end of the twentieth century, and risk has increasingly become synonymous with 'bad risk' and danger. Lupton points out that risk and probability calculations set out to tame uncertainty but can paradoxically lead to heightened anxiety because so much focus gets put on risk.

Hacking (1990) focuses on the development of parts of probability theory, an important building block for the calculability of risk. He traces how many statistics and probability theories were developed for or applied to the jury system in France during the French Revolution, involving many important figures of this field, most notably Laplace and Poisson. In court trials, many uncertainties have to be settled: Is the defendant guilty or not? One way to find an answer is juries, i.e. a group of (lay) people who have to find an answer to this question. At the time, France didn't have any experience with juries at court trials. This led to new questions that Laplace, Poisson and others tried to answer. They were, among others: How big should the jury be? How many choices does the jury have? By what majority should the jury decide? These

questions show that by establishing a jury, the uncertainty shifts from the defendant to the jury and its members.

Interestingly, before the French introduced juries, England already had a jury system in place (Hacking, 1990). There, however, the jury had to decide unanimously, i.e. in consensus through deliberation. Hacking also notes that a jury could only decide between guilty and not guilty in England, whereas in Scotland, there was a third option, 'not proven'.

From this, one can quickly draw analogies to the crowdsourcing of data sets: How many annotations per data point are necessary? Which annotations are accepted or allowed? And how to integrate the different annotations into the final annotation used for the training data? Even the question of which majority should decide is relevant to many crowdsourcing projects because majority voting is frequently used to integrate the annotations from multiple crowd workers. The strength of the majority also has implications on which questions to delegate to the jury. This issue was troubling Laplace: "How confident do we want to be that a jury has convicted rightly?" (Hacking, 1990, p. 90) For Laplace, even a strong majority of 10:2 jurors was too error-prone to be suitable for the death penalty. As a consequence, Laplace argued for its abolishment. While decisions addressed with crowdsourcing are not as drastic as a conviction for death, it does open up the question if there are debates among practitioners about the amount of agreement necessary in crowdsourced data and what kind of decisions should be crowd-sourced or delegated to a machine learning system trained on crowd-sourced data.

Referring to the work of anthropologist Mary Douglas (1966/1969), Lupton (1999) draws a connection between classification and risk. The argument is that all cultural classification systems fail to accommodate all things because some may fit no category (anomalies) while others may fit multiple categories (ambiguities). These things are sources of anxiety and turn out 'risky'. There are several ways how to cope with things that defy the existing classification system. First, ambiguous things can simply be put into one of several candidate categories. Second, anomalies can simply be removed physically so that they no longer challenge the classification. Third, the classification system can be adapted so that it can accommodate the anomalies successfully. Fourth, a new residual category that labels the anomalies as dangerous can be introduced. And fifth, cultures may turn to mythology or other cultural practices that give these anomalies a meaning on a different level than the classification system.

In their study of the International Classification of Diseases (ICD), Bowker and Star (1999) investigated the struggles of the World Health Organisation to find a suitable, working classification scheme. On the one hand, in their research, it became visible how social values and norms inform the compilation of the ICD. On the other hand, some strategies to cope with

the inevitable uncertainties came to light. In line with Douglas (1966/1969), the ICD contains several residual categories. The ICD adopts the aforementioned strategy of spreading: ""Not elsewhere classified" appears throughout the entire ICD, but nowhere as a top-level category. So since uncertainty is inevitable, and its scope and scale essentially unknowable, at least its impact will not hit a single disease or location disproportionately." (1999, p. 25)

I could not find many sources that specifically deal with uncertainty and crowdsourcing or 'AI'. In her 2019 essay *Doubt and the Algorithm: On the Partial Accounts of Machine Learning*, Amoore explores the relationship between machine learning algorithms and doubt. She argues that a machine learning algorithm's output always "pertains to the 'ground truth': a labelled set of training data from which the algorithm generates its model of the world" (pp. 4–5), with the labels often created by humans, e.g., by using crowdsourcing platforms such as Amazon Mechanical Turk. Drawing heavily on Donna Haraway's (2001) concept of 'partial perspective', Amoore argues that algorithms provide a partial perspective in so far as their output is generated from this 'ground truth' and that their output is 'truthful' in relation to this 'ground truth'.

What is particular to algorithms is their necessity to reduce the multiplicity that is present in the data (and the internal workings of machine learning algorithms, such as thresholds, weights and probabilities) to a single output and "that moment of decision is placed beyond doubt" (Amoore, 2019, p. 5). This is important, according to Amoore, because it is this condensing to a single output that allows algorithms to create actionable output. Against this observation, Amoore argues for calling into question the grounds on which a machine-learning algorithm comes to its output, i.e. the 'ground truth', to reinstate doubt in data, as the data was never "settled and certain" (p. 18). But it is not only the data that is doubtful. It is every step of applying machine learning algorithms where "doubt lives and thrives and multiplies" (p. 17).

There are two important theoretical motivations that I can discern from Amoore's essay. First, Amoore argues that algorithms are not alone in their capacity to give only *partial* accounts (related to their 'ground truth' in the case of machine learning), but that this is a well-known problem of humans, too. Second, she argues that by creating a single output, algorithms close other possibilities, and, by being the foundation of (political) decisions, they also foreclose possible futures: "The claim to a ground truth in data that pervades our contemporary political imagination precisely closes the door to the future, offering algorithmic solutions to close the gap and resolve the difficulties of decision." (2019, p. 19) By calling into question the data that is the foundation of the algorithm, and all the steps necessary to calculate the single output, the future can be opened again. In section 2.2.2 I will come back to the topic of 'ground truths', including the uncertainties and, as Jaton (2021a) calls it, hesitations involved in their production

that is close to Amoore's arguments.

## 2.2. Studies of data, data sets, and 'AI'

Data has become a ubiquitous resource and topic in recent years. Data has been publicly hailed as the new oil (The Economist, 2017), as a saviour of politics (MIT Technology Review, 2013) and threat to democracy (Larson, 2018), as a way to modernise businesses forming part of a "fourth industrial revolution" (MIT Technology Review Insights, 2020), and improve our health and fitness (Cha, 2015). A massive increase in data being produced, circulated, and consumed due to the ongoing digital transformation has been the basis for these hopeful claims (Kitchin, 2014).

Data has become ubiquitous and easily processed due to increasing processing speeds and storage capacities, with the rapid growth of data production being called a deluge (Kitchin, 2014) and explosion (Jasanoff, 2017). Historically, data collection and analysis has generally been the domain of governments, a situation that changed in the course of the twentieth century (Rieder & Simon, 2016; Kitchin, 2014; Desrosières, 1998). In the last decade, data, especially in the form of so-called 'Big Data', has become an increasingly hot topic in STS debates and interdisciplinary research frequently drawing on STS (boyd & Crawford, 2012; Iliadis & Russo, 2016).

These studies critically question assumed traits of ('big') data that have been hyped frequently: Data, particularly in large amounts, often come with claims to objectivity and accuracy. This claim is founded on the idea that data is a resource (like oil) 'out there' that simply needs to be discovered, that "exist in and for themselves" (boyd & Crawford, 2012, p. 667). But this is not the case. Instead, data is always produced, usually with a particular goal in mind. What kind of data gets recorded, circulated and consumed are important choices.

There's often a claim that the data speaks for itself, without the need for interpretation. Big Data claims to produce 'facts', but it needs an interpreting researcher even for data to be imagined as data (boyd & Crawford, 2012). As such, there's no such thing as 'raw data', as some form of 'cooking' is inevitably involved (Bowker, 2005). Consequently, data has to be understood as a "human-influenced entity" (Muller et al., 2019, p. 4).

Vast amounts of data – Big Data – are additionally often claimed to be exhaustive. They are supposed to provide a panoptic view of the world, an "all-seeing, infallible god's eye view" (Kitchin, 2014, p. 133). Contrary to this claim, data can only provide oligoptic "views from certain vantage points, using particular tools" (p. 133). As a consequence, data can be, e.g., "skewed by gender, race, income, location and other social and economic factors (not everybody

uses Twitter or Facebook, or shops in a particular store, or is on a particular phone network, etc.)" (p. 154).

It is particularly Big Data that also resulted in a rush of new forms of knowledge-making. It has significantly changed how many areas of research work and how research gets thought (boyd & Crawford, 2012). Berry (2019) argues that the "cult of data-ism" (p. 45) is a turn away from critical reason to data-deterministic thinking, which claims that there's "an abstract and metaphysical standard by which human action and society can be judged" (p. 45), leading to a new form of authority.

Another closely related insight is that data is always contextual and loses its (original) meaning when taken out of this context. As an example, data on 'friendships' on social network platforms is not the same as other forms of friendship and cannot be reasonably taken as the same (boyd & Crawford, 2012). Big Data approaches tend to create abstractions that can provide value but limit the possible inquiries. Context is particularly "hard to interpret at scale and even harder to maintain when data are reduced to fit into a model" (p. 671).

### 2.2.1. Data sets and their production

I differentiate data sets from data more broadly. While data can encompass diverse kinds of data points, maybe spread across many places and storage facilities, data sets are meaningful collections of data points for a specific purpose, resulting in a coherent set of data (Jasanoff, 2017). The design of data sets begins when the goal and the expected outcome of an application based on the data set gets defined (Miceli et al., 2021).

Jaton (2021b) conceptualises data sets as a form of inscription (Latour et al., 1986). Looking at data sets this way, it becomes clear that it's an active process to inscribe data, so it becomes a data set. As with all inscriptions, choices have to be made what to include and what to leave out. Bechmann and Bowker (2019) study 'AI' through the lens of classification theory. They argue that classification plays an important role in the 'AI' ecosystem, including the production of data sets. In addition to what to include and what to exclude, classification takes place when ignoring classes at the margins when compiling the data sets, and when defining potential categories (e.g. in the form of labels). In the latter case, Bechmann and Bowker also note that too few categories can result in useless information. In contrast, too many categories can lead to increased bias and randomness when conducting the categorisation. In general, the classes used for classification are historically, culturally, socially, materially and institutionally contextual, following the person(s) who conduct the categorisation (D'Ignazio & Klein, 2020; Bechmann & Bowker, 2019; Bowker & Star, 1999).

Generating a data set can require the combination of data from different sources. To achieve coherence, these data sources have to be integrated, involving many small decisions, e.g., how to address ambiguities that may arise during the process of integration (Diesner, 2015). Erroneous decisions can affect the "accuracy of the data and obtained findings" (p. 2). More broadly, the quality of the classifications taking place is crucial to the "perceived success or failure for a given social context" (Bechmann & Bowker, 2019, p. 3).

As Bowker and Star (1999) argue in their seminal work *Sorting Things Out*, it is human to classify, often taking on mundane and personal forms. But classification can have far-reaching consequences, as it can promote certain points of view while silencing others. Roughly along the lines of these arguments, D'Ignazio and Klein see data as a result of "unequal power relations" (2020, p. 38), illustrating it with the example that there are no government records of femicides in Mexico, which prompted researchers and activists to collect this data themselves as a tool to demonstrate the scope of femicide to Mexico's Congress.

Unequal power relations are also present in a different way, as there's often an imbalance between those that collect data and those that originally created the data (Miceli et al., 2021). Nowhere is this imbalance more apparent as with 'Big Tech' companies such as Google, Facebook, and Amazon, who collect tremendous amounts of data generated by their users and have (near) monopoly status in today's digital economy. But, as Miceli et al. argue, there are also imbalances between different types of workers who participate in the production of data sets, e.g., between those who define the problem to be solved, data scientists who formalise the problem and devise how and which data should get collected and how it should be processed, and those workers who then actually do the collecting, labelling and annotating. This topic is particularly controversial when it comes to crowdsourcing, as I'll discuss in section 2.3.1.

Several scholars have noted that the processes necessary to create data sets are often not documented (Bechmann & Bowker, 2019; Geiger et al., 2020; Miceli et al., 2021) and there are benefits to be had in documenting decisions and processes (Diesner, 2015; Jaton, 2021b). Several reasons can motivate a proper study and documentation of data sets and their production. First, for people involved in the production of data sets, proper documentation preserves knowledge and can potentially lead to improved work practices (Miceli et al., 2021). Second, it can disclose the data set's specifications, which can help future users choose appropriate data sets (Holland, Hosny, Newman, Joseph, & Chmielinski, 2018; Miceli et al., 2021). Third, it can be important for accountability reasons – towards clients (who may have commissioned the data set) and towards society at large. However, Miceli et al. (2021) note that striving for accountability can, if taken to the extreme, lead to forms of worker surveillance.

Geiger et al. (2020) study papers that used machine learning for tasks on Twitter and how well they documented the creation of the used data sets. They are interested, among other things, in the documentation of who the annotators were, if one data point got labelled by multiple people and if their agreement was calculated and disclosed, the compensation of the annotators and the instructions provided to the annotators. Their findings are that the amount of information varied widely, with clear room for improvement.

Documentation can, however, struggle to make explicit tacit and implicit knowledge. Taking up the concern over power relations at play, this includes the documentation of hierarchies, world views, and interests, as these are often taken for granted (Miceli et al., 2021). One approach is the development of standardised procedures to document the data sets, encompassing the motivation for the data set, its composition, a description of the collection process and recommended uses (Gebru et al., 2020). However, to make the mentioned implicit assumptions explicit, Miceli et al. argue for more reflexivity, including the workers' social position, the relations among various stakeholders in the process, and questioning the epistemologies at play. Additionally, people and institutions involved in the data set production have to be accounted for as important forms of influence, leading to a sharpened understanding that there is no such thing as 'raw' data.

One popular approach to producing data sets is crowdsourcing, which I will take a closer look at in section 2.3. But there are also more traditional employment arrangements, as companies specialise in data set production that provide their services to customers. Miceli et al. (2021) studied this kind of company and noted that frequently, workers came from marginalised communities, including refugees and slum residents. Kazimzade and Miceli observed that most workers at the companies they studied "have never received training on general knowledge regarding data-driven systems and machine learning, and many find it very difficult to reflect on the use and impact of their annotations" (2020, p. 5). The workers are also not aware of the goal of their work, i.e. what kind of product should be created with the resulting data set. This highlights the mentioned unequal distribution of power between the client and these workers. This also encompasses the way the workers have to, e.g., label data, as they usually have to follow rigid and standardised classification schemes provided by the client that allows no room for the workers' personal, subjective assessment of the data.

### 2.2.2. From data (sets) to 'Artificial Intelligence'

As I have written in the introduction, data sets are a crucial part of what powers 'Artificial Intelligence'. The other part is algorithms that process the data, learn from the data and reason over the data. When looking at data sets with regards to 'AI', a few additional strands of

research become relevant.

But first, let me tackle the term 'algorithm'. This term has been causing debate among social scientists but also within computer science (Vardi, 2012), as an algorithm can be many things to many different people and depending on context (Gillespie, 2016; Kitchin, 2017). As a consequence, a variety of definitions is used. This discontinuity and vagueness resulted in some criticism from computer scientists and mathematicians who claim to have precise definitions and clear boundaries of what an algorithm is and what it is not (Seaver, 2017).

Beyond the academic fields, 'algorithms' have become a hot topic in public discourse in recent years. There, 'algorithm' is often used as a very broad and inclusive rhetorical figure (Gillespie, 2016), again leading to a rather vague understanding of an algorithm. Notions such as 'Facebook's algorithm' or 'the Google algorithm' are frequently used in news media and discussions centred on these companies' services.

Criticism hinging on this vagueness of what an algorithm is was also raised from the social sciences. Many characterisations when discussing algorithms have described algorithms as elusive yet very powerful entities, as mythical creatures (Ziewitz, 2016). Ziewitz draws a parallel to the language of politics that "tends to privilege the figure of the lone decision maker at the expense of more complex realities" (p. 6). Ziewitz and Seaver (2017) argue for using the fact that the term algorithm is unstable in a productive way to understand algorithms as multiples. Berry (2019) warns that abstractly using the concept 'algorithm' "can obscure the specificity of computational instances" (p. 44); thus, critical engagement with algorithms has to attend to the specific (material) instances of algorithms.

Loosely along the lines of Seaver and Ziewitz, many studies of algorithms conceptualise algorithms as a complex, heterogeneous socio-technical assemblage (Ananny, 2016; Gillespie, 2016; Kitchin, 2017). Kitchin lists many elements of the assemblage: "Systems of thought, finance, politics, legal codes and regulations, materialities and infrastructures, institutions, inter-personal relations" (p. 17) as well as "all kinds of decisions, politics, ideology and the materialities of hardware and infrastructure" (p. 17) all shape the production of algorithms. Looking at algorithms as complex, heterogeneous socio-technical assemblages de-mystifies them and makes their messiness, situatedness as well as political and cultural embeddedness visible (Ziewitz, 2016).

Returning to data sets, it is clear that they are an important part of algorithms: Algorithms operate on (input) data and create (output) data. The connection between algorithms and data is particularly tight in the case of machine learning algorithms, as their behaviour is a direct consequence of the data they get trained on. Thus, the production of training data warrants particular attention. But machine learning algorithms will subsequently operate on

data, creating new labels and annotations. As such, algorithms are a dual entity: They are *doing* things, but they also *need* things (Jaton, 2021b).

Training data is frequently based on so-called 'ground truths'. In his monograph *The Constitution of Algorithms*, Jaton (2021b) describes how a 'ground truth' is used for supervised machine learning: First, a set of data points is assembled into a data set. These data points then get manually labelled or annotated. Data points and their labels form the 'ground truth'. This is subsequently split into two parts, a so-called training set and an evaluation set. The machine learning algorithm gets trained on the training set, using both the data point and its label. On the evaluation set, the trained algorithm then has to predict the correct label. Depending on the number of correct and incorrect predictions, performance measures can be calculated that signal the quality of the machine learning algorithm.

'Ground truths' are crucial to 'AI' research beyond training and evaluation of the algorithms themselves. Instead, 'ground truth' data sets frequently get published. They can then become a common point of reference for research teams competing in the development of algorithms that perform well on these data sets. Jaton's (2021b) study vividly shows three things. First, machine learning research cannot progress without an appropriate 'ground truth' available to researchers. (In Jaton's case study, it was the detection of the visually most important part of an image.) Second, these 'ground truths' are designed for a very clearly designed problem. However, this problem formulation is often constrained and insufficient for new application areas (in that case study, the assumption was that there's only one clearly significant element in the image). The research team studied by Jaton attempted to create an algorithm with more robust real-world usability, but this kind of robustness could not be shown on the established 'ground truth'. Consequently, the research team attempted to produce a new 'ground truth'. Which, thirdly, was difficult to establish in the machine learning community. Thus, establishing a recognised 'ground truth' in the community requires convincing arguments and is not a matter of simply making the data set freely available.

Several scholars have problematised the term 'ground truth'. As with the term 'raw data', Miceli et al. (2021) argue that this notion places the people involved in the creation of the 'ground truth' as external, as "outside the object of research" (p. 169). Bechmann and Bowker (2019) argue that 'ground truths' can cause problems, partly through assumptions that are implied. One assumption is that there *is* a 'ground truth', ignoring confirmation bias. Second, 'ground truths' are ahistorical, i.e. they imply that the correct labels remain temporally stable. Bechmann and Bowker make this case for behavioural predictions based on social media activity, an area where this issue is aggravated. A third assumption is that the 'ground truth' will remain

truthful and without performative effects itself on the future. Again, Bechmann and Bowker argue that this applies to the prediction of user behaviour, but I'd argue that this can affect many other applications, too.

An important, controversial, and highly researched topic is that of the 'bias' of machine learning. While computers have often been positioned as a means to overcome human biases and discrimination, it turns out that biases also plague machines (Barocas & Selbst, 2016). Biases are also a source and explanation for many unwanted phenomena that I referred to in the introduction (stereotyping machine translation, racist and sexist face recognition, 'self-driving' cars incapable of handling certain situations). Biases in data are seen as imperfections, and there is the suspicion that the data inherit the prejudices of the humans involved in their generation. This can, in turn, lead to (continued) discrimination of disadvantaged populations (Eubanks, 2018).

There has been an increasingly active interdisciplinary community of researchers that have engaged critically with this issue under the banner of 'fairness', meeting at conferences such as the ACM FAccT[3] and FAT ML[4] conferences. A lot of this research comes from technologists concerned over potential unwanted consequences of machine learning, often adopting a technology-focused approach to algorithmically-mathematically identifying and preventing biases. The research is too diverse and too voluminous to cover here in detail.

Still, two papers seem worthwhile to mention explicitly: First, a seminal paper by Friedman and Nissenbaum (1996) highlights that this debate has a decades-long history and is not restricted to machine learning. Second, sensitivities from STS have increasingly been brought into the discussion, such as the importance of the social context and the view of machine learning as *socio*technical system (Selbst, boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). In their paper, Selbst et al. emphasise that approaches focusing only on the technical, that abstract from the social context where the machine learning will be used, have inherent limitations. Of the five 'abstraction traps' they identify, three appear relevant to my study: First is the insight that the social context where a ('fair') machine learning system will be used has to be modelled, i.e. accounted for, and cannot simply be abstracted away. Second, a system that may work well in one particular social context may break down and do harm in a different social context. And, third, many concepts cannot simply be formalised mathematically because the "meaning of social concepts such as fairness, ... can be procedural, contextual, and contestable, and cannot

[3]FAccT stands for Fairness, Accountability, and Transparency. The annual ACM FAccT conference has taken place since 2018. More information at https://facctconference.org.

[4]Here, FAT is a superseded acronym for, again, Fairness, Accountability, and Transparency. This conference was held annually from 2014–2018. More information at https://www.fatml.org/.

be resolved through mathematical formalisms" (p. 61).

Relating to these debates, but turning our attention back to 'ground truths', Jaton (2021a) (re-)positions biases as inherent to 'ground truths' and not necessarily as something negative. On the contrary, Jaton asserts that biases are a necessity for machine learning, that it is biases that machine learning actually learns from. In line with researchers' concerns of biases, Jaton notes that several choices have to be made when 'ground truthing' machine learning algorithms: First, the problematisation, i.e. the decision of what problem to be addressed. Second, the data to collect, and, third, the labelling of this data (which is the subject of this thesis). Due to these decisions to be made, Jaton highlights that "[i]t could be otherwise" (p. 6). Following from these observations, he argues that it is the acknowledgement of these choices and possibilities, or the lack of acknowledgement, that is of significance. Jaton notes that many organisations opened up, inviting researchers (e.g., sociologists) to investigate the processes that shape machine learning systems. In contrast, others – especially the powerful actors in the 'AI' industry – are "reluctant to make hesitations and uncertainties visible" (p. 8).

Newlands (2021) highlights that there is considerable effort necessary to make 'AI' work. She conceptualises the 'AI' supply chain as riddled with human labour, as the "chain of collection, curation and custody of data from source to model, passing through the hands of potentially infinite numbers of data workers, data brokers and data scientists on the way" (p. 2). Crowdsourcing is often one or several links in this supply chain. At the same time, 'AI' vendors are usually going to great lengths to make this human work invisible, but strategically have to 'lift the curtain', when, e.g., explaining a client how "mundane human effort" (p. 2) is required to make their product 'work'.

## 2.3. Crowdsourcing

In the previous section, I have mentioned crowdsourcing already several times as a means to create data sets. The term 'crowdsourcing' can convey many meanings besides data set creation. It can describe services such as Uber and AirBnB, where a 'crowd' of drivers and landlords, respectively, operate through commercial platforms that provide a unified experience for customers (Ashton, Weber, & Zook, 2017). It can also mean citizen science in the sense that many citizens contribute little pieces to a larger data collection and problem-solving effort (Nowotny, 2014). Even cities turn to crowdsourcing as a participatory process to develop new policies (Ashton et al., 2017; Brabham, 2013). Crowdsourcing can also be a means to accomplish what has long been the sole domain of state actors (Jasanoff, 2017). What these examples share is the out*sourcing* to a *crowd*, i.e. a large number of (potentially anonymous) people. This is also

the source of the portmanteau of crowdsourcing, which most scholars trace back to a WIRED article by Jeff Howe (2006).

Brabham (2013) points out that crowdsourcing is no unified research field because it is spread across multiple disciplines with little exchange between these disciplinary debates. While the term crowdsourcing has been applied very broadly in the years since the article by Howe, Brabham defines crowdsourcing as

> "an online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals. Online communities, also called *crowds*, are given the opportunity to respond to crowdsourcing activities promoted by the organization, and they are motivated for a variety of reasons" (p. xix, emphasis in the original)

This definition does not quite cover cases such as Uber and AirBnB, but it covers well the form of crowdsourcing that I investigate in this thesis.

### 2.3.1. The work of crowd workers

Why would one turn to the crowd? What kind of work is it that the crowds do? One of the motivations to resort to crowdsourcing is that humans are seen as better qualified than computers to solve some tasks, such as "language translations, survey responses, information gathering" (Brabham, 2013, p. xx). Brabham provides a problem-focused typology to differentiate crowdsourcing between knowledge discovery and management, broadcast search, peer-vetted creative production, and distributed human-intelligence tasking. The latter problem concerns the analysis of large amounts of information, frequently by splitting up larger tasks into small 'micro-tasks', and fits the case of this study because this "approach to crowdsourcing is appropriate when a corpus of data is known, and the problem is not to produce designs, find information, or develop solutions but to process data" (p. 50).

A hope associated with crowdsourcing is that, under favourable circumstances, experts can be outperformed by 'the crowd' (Brabham, 2013). At the same time, creative crowdsourcing (e.g. designing t-shirts) is largely done by professionals who have a solid career in design. In contrast, for scientific crowdsourcing, crowd workers may even have PhDs in the respective field. This puts some doubt on the portrayal of the crowd as amateurs, as Brabham notes. The motivation why people participate in crowdsourcing varies widely, too. For example, creative professionals may be motivated by the possibility to build a portfolio for future employment, a motivation that is not, I argue, likely to be relevant for people involved in 'human intelligence tasking'.

When it comes to the creation and processing of data, other terms than 'human intelligence tasking' and 'crowdsourcing' are used for the same phenomenon, too. As Irani and Silberman (2013) argue, each term has slightly different connotations. Calling crowdsourcing platforms 'micro-labour marketplaces' emphasises the market dynamics (pricing, transaction management and choice of task by the workers). Calling it 'human computation' and 'Humans-as-a-Service' frames the crowd workers as a resource that can simply be plugged into a larger computational system.

Berry (2019) studied Amazon Mechanical Turk as a case study for the development of what he calls Critical Theory of Algorithms. Berry takes a normative stance when he argues that this crowdsourcing platform transforms "labour ... into a commodity through an interface." (p. 48) Marketplaces for 'micro-labour' – AMT and other forms of the 'gig economy', e.g., Uber – show how social conflict gets embedded in computational technologies. What these platforms make possible is to create an "unending stream of labour-power on demand in a similar fashion to an electricity or water supply" (p. 49).

Gray and Suri (2019) see crowd work as a continuation of contingent work, a kind of work that (seemingly) needs no professional training nor particular skills and which doesn't require nor need full-time jobs. Historically, contingent work was, e.g., "[f]arm wives sewing [and] young black women tallying numbers by longhand" (p. 58). Similar to these 'others', crowd workers get "devalued because the tasks they do are typically dismissed as mundane or rote" (p. 58), often at a physically remote location. Partly due to this perception, crowd workers possess no cultural influence, highlighting the aforementioned power imbalances when creating data sets (Miceli et al., 2021).

Additional concerns are that crowd workers get paid less than people doing the same in traditional employment arrangements (Brabham, 2013). Brabham notes that, at the time of writing his book, regular users of Amazon Mechanical Turk earned only $2 per hour. Closely related is the concern that crowdsourcing can undercut professionals, leading to an erosion of ethical standards built up by professional associations. Brabham argues that much of the criticism comes from professionals that were struggling and look to crowdsourcing as a scapegoat.

In his book chapter, Berry (2019) engages critically with a particular crowdsourcing application developed at the MIT that integrates Amazon Mechanical Turk with Microsoft Word in various ways (Bernstein et al., 2015). In the developers' words, what this project does is to place "workers in productive tension with one another" (p. 90). It achieves this by splitting the task into several small sub-tasks and letting one worker choose the best solution created by other workers, effectively supervising their work. Berry criticises that

> "labour is inscribed into the system, but also de-humanized and reified into pure labour power, which is abstracted from its human form. But more than this, it is the purity of the algorithmic "pattern" that, stripped of normative content and de-contextualized, appears to justify and encourage potentially exploitative and unjust labour practices." (p. 54)

As Berry notes, these practices are already widespread in the 'gig economy' in the form of ride-sharing (Uber), food delivery (Mjam, UberEATS), and domestic work (Taskrabbit).

Crowd workers are usually working "insulated from other workers" within their tailored user interfaces (Berry, 2019, p. 49). Despite this isolation, some crowd workers are organising. One means of organising is Turkopticon, a platform that allows crowd workers to share their experiences with employers (Irani & Silberman, 2013), a similar project is the platform Dynamo (Salehi et al., 2015). Other attempts at worker organisation are the collectives TurkerNation, MTurkGrind, and the Reddit /r/HITsWorthTurkingFor (Berry, 2019).

However, this does not mean that crowd workers, platform operators and customers of these platforms are on equal footings. There are considerable power imbalances at play. Most critically, people who request crowd workers to solve a task via Amazon Mechanical Turk have far-reaching rights: They may decide not to pay the crowd worker after the fact, e.g., if they are not content with the work, and they can filter crowd workers based on 'approval ratings' and depending on the workers' solutions to test tasks (Irani & Silberman, 2013). Some customers even pay only the agreeing majority if asking several crowd workers to complete a particular task. Crowd workers can contact the 'requester' in such cases, but there's no legal recourse possible, and frequently these requesters do not respond. This can lead to wage theft.

One key element to make workers invisible is a wrapper, the application programming interface (API), that many crowdsourcing platforms provide (Irani & Silberman, 2013; Berry, 2019). By invoking functions of this API, programmers can access crowd workers the same way they would invoke software programs. This is an important form of abstraction in software engineering where the person using an API does not care how exactly the mechanisms behind the API work as long as the results are correct. This way, the fact that people actually complete these requests becomes largely hidden. As Irani and Silberman argue, "by hiding workers behind web forms and APIs, AMT helps employers see themselves as builders of innovative technologies, rather than employers unconcerned with working conditions." (2013, p. 613)

### 2.3.2. Representativeness of the 'crowd' and its consequences

Another important issue is the representativeness of the crowd. An important question is who contributes to the production of data and what exclusion mechanisms are in play. Adams and Brückner (2015) have shown that these forms of inclusion and exclusion have clearly visible effects on Wikipedia. How is Wikipedia relevant to my research? In two ways. First, Wikipedia can also be understood as a form of crowdsourcing, even though it is structurally quite different from the processes that I study. This crowdsourced nature of Wikipedia promises democratization of knowledge, both by making access free as well as by enabling participation. But the authors argue that several aspects of Wikipedia limit the democratic potential and the level of quality, particularly the non-representativeness of contributors and the necessity to conform to a certain jargon in order to succeed as a contributor.

Second, Wikipedia is often used as a frame of reference for a broad public. Building on this function as a frame of reference, it gets used as a data source for a myriad of 'AI' applications. This is often through DBpedia, a (machine-readable) knowledge base that is based on extracted, structured information from Wikipedia (DBpedia Association, 2021). Studying the processes of Wikipedia can, thus, provide insights into a cascading set of modern 'AI' applications that build on top of it.

Newlands and Lutz (2020) study the exclusion due to physical barriers. Even though crowdsourcing on, e.g., Amazon Mechanical Turk gets portrayed as an additional source of income for underprivileged people – because people from all over the world can participate – Newlands and Lutz point out that this requires access to the internet with a suitable device. Their study focuses on people whose internet access has only ever been with mobile devices (skipping stationary internet devices such as PCs and laptops). Their findings were that mobile devices are disadvantageous to work via AMT for various reasons, including slow processing speed and the difficulty of multi-tasking. Users who rely solely on mobile devices cannot work as efficiently as crowd workers with access to a PC or a laptop, which directly impacts the earnings possible on the platform. On top of this, some people who use AMT to crowdsource tasks explicitly forbid the use of mobile devices and thus explicitly exclude workers who only have access to this type of device.

What these studies show is that there's not 'the' crowd, but crowd*s* in the plural – similar to the Deweyan notion of public*s* (Dewey, 1946; Ashton et al., 2017). There are many forms of inclusion and exclusion at play, some more subtle than others, all impacting the representativeness of 'the crowd', which in turn impacts the data sets that get produced through crowdsourcing.

# 3. Theoretical framing and sensitising concepts

For my study, I draw on Actor-Network Theory (ANT), one of the core analytical frameworks of STS. ANT's core contribution is the insight that not only humans can be actors but that non-human elements can be actors, too. Thus, non-humans can actively construct things. This is in contrast to social constructivism that privileges humans (Michael, 2016; Pinch & Bijker, 1984). Both crowdsourcing and 'AI' systems that use the resulting data sets can be understood as actor-networks. The inclusion of non-human elements makes ANT a good fit to analyse the crowdsourcing of data sets.

In this section, I will first describe the classical ANT formed in the 1980s across several influential works and the core sensitivities they established. I will then go into more detail about concepts and debates around ANT relevant to my study, namely that of inscriptions and different kinds of actors. Finally, I will address some more recent, relevant developments in the tradition of ANT, such as the focus on multiplicity, matters of concern and ontological politics.

## 3.1. Classical Actor-Network Theory

Actor-Network Theory is a theoretical framework that studies how relations among human and non-human actors are formed, resulting in the eponymous actor-network. ANT took form in the 1980s with three studies about the success of the French microbiologist Louis Pasteur (Latour, 1983), the domestication of scallops in the French St. Brieux Bay (Callon, 1986) and the Portuguese's success as seafaring nation (Law, 1986). Common to these studies is that they studied *how* – and not why – relations are formed. These relations, the proponents of ANT argue, are 'the social', something that is elusive and given in some other strands of sociology, leading to argumentations that 'the social' is the cause of certain phenomena, without ever explaining what 'the social' is supposed to be.

The early ANT studies introduced three methodological principles. First, *agnosticism* argues that the researcher has to be "impartial towards the scientific and technological arguments used by the protagonists", i.e., no point of view should be privileged (Callon, 1986, p. 200). Additionally, researchers should make no presupposed assumptions about the actors and their identity if the identity is still under negotiation.

Second, the principle of *generalised symmetry* extends Bloor's principle of symmetry from the Strong Programme of the Sociology of Scientific Knowledge (1991 [1976]) so that human and non-human actors have to be described and analysed using the same repertoire of language. Put differently, all arguments must be acknowledged and explained the same way – whether it affects human, natural or technical aspects (Callon, 1986). Michael (2016) notes that this principle is

also one key reason why ANT scholars attempt to use neutral language to describe actors and processes in actor-networks (even though the terms used by Callon and others have their own historical baggage).

The third principle *free association* means that any kind of association between all kinds of actors are considered possible, i.e. none are a priori denied or assumed. The "hypothesis of a definite boundary" between natural and social events has to be rejected by the researcher (Callon, 1986, p. 200). Instead, the researcher "follows the actors in order to identify the manner in which they build and explain their world" (p. 201).

In addition to these three principles, Callon (1986) defines the establishment of an actor-network by the network builder consisting of four "moments of translation" (p. 203) which are not strictly consecutive but can overlap. First, the network builder is involved in the act of *problematisation*, i.e. posing questions in a way that makes the scientists indispensable. Callon exemplifies these moments with a case study of three biologists that propose a new form of scallop farming in St. Brieux Bay. They try to convince the fishermen that their proposed form of farming is the only way for sustainable scallop harvests in the future. Still, they also have to convince the scallops to attach as expected to rods placed in the sea, something that eventually fails.

For other actors to be interested in the problem, the network builder also has to (re)define their identities and interests. This is the second moment of translation, the process of *interessement*. The identities and interests are (re)defined in a way that frames the other actors as interested in a solution to the problem defined by the network builder. Put differently, this problem is constructed in a way that cannot be solved satisfactorily without involving the network builder. This turns the network builder into an *obligatory passage point* (OPP). This works through the usage of devices of interessement that are placed between the actor of interest and competing actors that provide alternate identities. If successful, the actor of interest associates with the network builder, disassociates from other competing actors and the actor's identity is redefined. In case of success, this contributes to the validation of the problematisation.

*Enrolment* is the third moment of translation. At the stage of interessement, the established relationship is still fragile. By transforming questions from the problematisation into more certain statements, true alliances should be formed. Taking Callon's case of the scallops and the fishermen, during the moment of interessement, the devices to persuade the fishermen were, e.g., reports of the domestication of scallops by Japanese fishermen and curves that show the declining scallop population in St Brieux Bay. During the moment of enrolment, however, roles have to be defined and distributed (hence the name en*rol(e)*ment) involving trials of strength:

The scientists have to show that the scallop larvae not only anchor in Japan but in St Brieux Bay, too – which involves fighting currents and disturbances of the scallop larvae. Only if the scientists can successfully show that the larvae anchor are the roles sufficiently defined and distributed. At this point, the fishermen are not only interested in the vague goal of sustained scallop harvest but start backing the scientists.

Finally, the *mobilisation of allies* "renders entities mobile which were not before" (Callon, 1986, p. 216), leading to the role of a spokesperson that speaks for other actors in the network. However, by speaking for some, this role also silences others. In the case of the scallops, only a few larvae were observed and counted, but many did not attach. The counts were converted into curves and published. These curves are the mobilisation: They are mobile and can be circulated – different to the larvae attached in St. Brieux Bay. Assuming the paper is accepted as significant in the scientific community, the researchers can "speak legitimately for the scallops of St. Brieux Bay" (p. 216), and the scallop larvae provide active support to the scientists.

These four moments of translation motivated the term *sociology of translation* that has also been used for ANT (Callon, 1986; Latour, 2005). These translations are the displacements of actors during the building of the actor-network. Figuratively goals, interests and identities, but also literally humans and non-human animals, devices and curves get moved through "negotiations, intrigues, calculations, acts of persuasion and violence" (Callon & Latour, 1981, p. 279). If a network-building actor is successful, it can grow by accumulating successful translations. If these translations are stable, they become black-boxed: *Black boxes* contain "that which no longer needs to be reconsidered, those things whose contents have become a matter of indifference." (Callon & Latour, 1981, pp. 285) This also paves a way to move the analysis from the micro to the macro level: Actor-networks with stable, black-boxed relations can act as actors themselves. This way, one can see, e.g., a university as one actor, even though it itself consists of many actors and relations.

## 3.2. Hybrids and agency

As humans and non-humans form associations and create networks, they become increasingly intertwined. Because of this, ANT tends to speak of *hybrids* of humans and non-humans. Frequently, there is, quite simply, no reasonable possibility to try to separate humans and non-humans, as the interdependence is so strong. Law (1994) brings the formidable example of the director of the Daresbury laboratory: The director is so closely intertwined with non-humans and other humans that make his performance (or enactment) possible: His phone, his secretary, papers that circulate in the office, visitors he receives, etc. It is that "all these materials and

endless others together perform [the] Director of Daresbury." (p. 143) However, he points out that 'the Director of Daresbury' is also not reducible to the humans and non-humans that surround the person. It is him who embodies these relations. What this observation amounts to is that there exists *distributed agency* that is only possible through this hybrid of humans and non-humans (Michael, 2016).

Pickering (1995) observes that *agency shifts temporally* between agents in an actor-network. He illustrates this with scientists constructing a machine (which could well be some computer or algorithm). They first perform an active role when constructing this machine but become passive as they try out if it works as predicted. At this point, it is the machine that is active. If it turns out that the machine does not yet perform as intended, the agency shifts back to the scientist. This reversal of roles can take many iterations. What is important about this 'dance of agency', as Pickering calls it, is that it is not only the human who reconfigures the machine but that the machine also reconfigures the human's intentions. If the machine does not work as intended, it can readjust the scientist's goals.

### 3.3. Inscriptions and immutable mobiles

Actor-Network Theory has been paying particular attention to different forms of representations. A key concept of ANT are *inscriptions* as a particular form of representation. As Michael (2016) notes, ANT "is interested in the internal workings of these representations mainly insofar as these impact on what those representations can 'do'" (p. 23). These representations form a crucial part of the enrolment of actors.

Inscriptions are produced by inscription devices that can be complex, such as bioessays (Latour et al., 1986), but also through simple inscription processes such as counting the scallop larvae that anchored to the first capture device in St. Brieux Bay. The resulting inscriptions can be diagrams, graphs, or written text, such as a statement. What is important to note is that inscriptions "are regarded as having a direct relationship to "the original substance.""(p. 51), and subsequent debates will focus on the inscription instead of the 'original substance'. Inscription devices are prime examples of black boxes: The inner workings of inscription devices such as, e.g., the mentioned bioessay, are irrelevant to the scientist-user. What is important are the input and output.

The function of inscriptions is the *persuasion* of the reader. The less visible it is how the inscription was achieved, the more persuasive it is. This is also how 'facts' are constructed: "[A] text is seen to contain a fact once readers no longer feel the need to interrogate how that text was put together" (Michael, 2016, p. 30). The result is often statements that are stripped of

modalities and formulated with certainty, i.e. "[m]oving a modality from 'it is probable that A is B', to 'X has shown that A is B'" (Latour, 1983, p. 162). At this point, it is only with considerable effort that other actors can raise convincing arguments against these inscriptions. Based on these 'facts', the network-builder can make (new) claims.

An important property of inscriptions is their function as *immutable mobiles* (Latour, 1987). This means that they are highly mobile and consequently can be circulated widely and easily. At the same time, they retain their meaning when travelling between actors and contexts. But their meaning can still be challenged, or to put it differently, immutability is not guaranteed. This is most apparent when they encounter another immutable mobile, at which point a trial of strength occurs (such as during the phase of enrolment). One key aspect to the success of inscriptions as immutable mobiles is the 'correct' reading by the receiving audience (Michael, 2016). Only if this is the case do they stand a chance of succeeding in trials of strength. If they fail, "inscriptions can collapse catastrophically" (p. 40).

As I mentioned above, the creation of 'facts' allows an actor to make new claims. Latour and Woolgar (1986) call this 'cycles of credit': Actors get 'credits' for the creation of facts which can, in turn, be invested to create further facts. Similarly, inscriptions can be cascaded: The results of inscriptions form the base for new inscriptions, or, turning to the concept of input and output of inscription devices, the output of one or several inscription devices become the input for another inscription device (Latour, 1990). As Michael (2016) notes, this cascading effect makes it increasingly harder for other actors to problematise the resulting inscription.

### 3.4. Of actors, intermediaries, and mediators

As stated, the core of an Actor-Network are actors and their associations. And, as also discussed, devices of interessement, inscriptions, and immutable mobiles play an important role in the establishment of associations. Are these things different from actors, given the fact that non-humans can be actors, too? And if so, how?

Callon (1991) differentiates actors from *intermediaries*. Intermediaries are defined as "anything passing between actors which defines the relationship between them" (p. 134). He illustrates this concept with the example of a product that defines the relationship between the producer and the consumer. By defining the relationship, intermediaries also define the roles of the actors that form this association. Thus, intermediaries play an essential role in the formation of an actor-network. What sets actors apart from intermediaries is their capacity for authorship, something that becomes visible when a company trademarks a product or scientists write articles under their name. Consequently, inscriptions are one kind of intermediaries, but

computer software, contracts, money and even disciplined bodies are other types of intermediary. Put differently, an actor is a special intermediary in that it "puts other intermediaries into circulation" (p. 141). What is an actor and what is an intermediary is thus a question of empirical investigation.

In an endnote Callon (1991) already refers to *mediators* as a third concept in between actors and intermediaries. Latour (2005) is more detailed on the distinction between mediators and intermediaries. Whereas an intermediary transports the meaning faithfully without modification, a mediator can change the meaning as it moves about. Latour takes the example of a computer: As long as it functions properly, it can be taken as a case of an (albeit complicated) intermediary. But if it breaks down, it is a "horrendously complex mediator" (p. 39). Deciding if an entity acts as an intermediary or as a mediator becomes, again, an empirical question: "[T]here exist endless number[s] of mediators, and when those are transformed into faithful intermediaries it is not the rule, but a rare exception that has to be accounted for by some extra work" (p. 40). Thus it is up to the researcher to study if and how a mediator is turned into an intermediary by the network builder.

### 3.5. Multiplicity and matters of concern

Many early ANT case studies were interested in the construction of 'facts' and how actor-networks become stable. More recent works in the tradition of ANT have shifted their focus. Instead of showing that matters of 'fact' are constructed (the important contribution of early ANT), it aims to problematise realities: *Matters of concern* raise the question of ontology, of "what the *real* world is *really* like" (Latour, 2005, p. 117).

It is in the wake of controversies that what have been accepted as matters of fact become "highly uncertain and loudly disputed" matters of concern (Latour, 2005, p. 114). One example is that of spermatozoids that "used to be obstinate little machos swimming forcefully toward the powerless ovule" (p. 115), but now there's a dispute among scientists if it is not that "they are now attracted, enrolled, and seduced by an egg" (p. 115). It is, according to Latour, up to the "abilities of the collective to unify" (p. 116) these multiple realities. Mol (1999) argues that "reality is historically, culturally and materially located" (p. 75). As Michael (2016) notes, this implies that realities are open to critique and can be contested – which is precisely what happens during controversies.

The emphasis on reality is also in sharp contrast to postmodern thought that takes the stance of multiple perspectives on one singular reality or ontology (Latour, 2005; Mol, 1999). Instead, here, reality is assumed as, by default, multiple. Instead of observing a singular reality through

multiple perspectives, multiple realities are *enacted*. This also implies that there is no single, essential thing that can be encountered, but it is always a particular entity (Mol, 1999). As Michael (2016) puts it, "we never see 'the car' that happens to be red or rusted or broken but always the particular car ('this red car' ...)." (p. 121)

As Mol (1999) illustrates with the example of anaemia, the object anaemia varies from one stage in the hospital to the next. In the consulting room, anaemia is a set of visible symptoms of the patient. In the laboratory, however, anaemia is the haemoglobin level measured and compared to a standard level. In both cases, it is anaemia, but it's enacted quite differently, and these different realities "co-exist in the present" (p. 79). These different realities can lead to tensions, but they can also successfully co-exist, "collaborate and even depend on one another" (p. 83).

### 3.6. Ontological politics

Assuming that there are multiple realities of an object, the question becomes if there is a choice which reality to enact (Mol, 1999). Mol reframes this into the question of where this choice is located. Staying with the topic of anaemia, Mol presents two ways how the detection of anaemia could be enacted: In a clinical manner or through screening programmes using haemoglobin tests. She highlights that it is out of historical circumstances that countries adopted either of these two approaches, "but there was never a moment or a place where it was decided" (p. 79).

*If* there was a decision to be made, frequently the arguments brought forward for or against a certain enactment actually *shift the site of decision* somewhere else, "to places where, seen from here, it seems no decision, but a fact" (Mol, 1999, p. 80). This can be the public opinion portrayed with confidence; it can be economic 'facts', etc. If one continues this thought, the options to choose among always shift further and further from the current situation.

To highlight that it is political – that is, that the conditions we live in are shaped with practices – which reality to enact, Mol (1999) introduces the concept of *interference*. Put simply, different enactments of a thing are linked to enactments of other things. In the case of anaemia, establishing the 'normal' haemoglobin level statistically raises the question of the population for which to calculate this 'normal'. As Mol argues, this usually involves differentiating into age groups (especially children), men and women, and even pregnant women. Thus, enacting the statistical 'standard' haemoglobin level implies enacting or interfering with enactments of 'children', 'men' and 'women', and 'pregnancy'. This further implies 'women' to be enacted biologically. A different form of setting the 'standard' haemoglobin level, the pathophysiological approach, compares the haemoglobin levels of the same patient to the levels of the patient in a

healthy situation; thus, no formation of groups is necessary.

Mol's concept of ontological politics sensitises us to ask where choices are being made and which models of relationships are mobilised. In line of ANT, Mol (1999) is highly critical of models that let 'customers' or 'experts' decide 'after the fact'. Instead, she argues that choices are often incorporated in the effects we seek, against which we identify the effectiveness of measures, but also in the techniques we use. As such, she prefers not to ask the question who chooses, but who *does* ontological politics and how to handle these incorporated choices.

## 4. Research question

This thesis contributes to the study of data (sets) and 'Artificial Intelligence' by investigating how practitioners address uncertainties during the process of crowdsourcing. How various forms of uncertainty are resolved are seemingly small procedures but can have profound impact on the resulting data sets. If these data sets are subsequently used to build 'AI' systems, it shapes these systems, too. My focus is not on manuals that describe best practices or (scientific) publications that re-present crowdsourcing processes to publics but on the messy actual practices 'on the ground'. Thus, my research is guided by the following main research question:

> **How do practitioners that crowdsource data sets address uncertainties during this process? (MQ)**

To obtain a rich picture of these practices, four sub-questions with different foci additionally guided the research.

First, it is interesting to know where practitioners locate the sources of uncertainty and how these sources create tensions and anxiety during the crowdsourcing process. The following sub-question addresses this issue:

> How do practitioners think that uncertainties get introduced during the crowdsourcing process? (SQ1)

Closely related to SQ1 is the composition of the crowd, i.e. who is tasked with the annotation process. More importantly for this research, however, are imaginations about the 'crowd'. As mentioned in the introduction, sometimes the crowd is being portrayed as deficient and framed as problematic (Doan et al., 2011; Oleson et al., 2011). This is particularly important for distinguishing between a 'lay' crowd and an 'expert' crowd and can impact the coping strategies for uncertainties. This question addresses how imaginations are affecting the coping strategies:

> How are imaginations of the 'crowd' impacting coping with uncertainties? (SQ2)

There are many ways how practitioners can address uncertainties. There are also many reasons they might choose a particular method, e.g. because it's an industry standard, internal standard in an organisation, personal preference, feasibility, and availability. The following question asks which kind of mechanisms practitioners use and why they chose the particular ways of dealing with uncertainty:

> How do practitioners choose methods to address uncertainties during crowdsourcing? (SQ3)

Some crowdsourcing platforms are very widely known, Amazon Mechanical Turk being the prime example. Some practitioners use tailored crowdsourcing platforms. These platforms and other technologies are a crucial infrastructure to the crowdsourcing process. Consequently, it is relevant to investigate their role in addressing uncertainties, whether they afford certain ways to deal with uncertainties and if they impose restrictions on how to address them, leading to my final sub-question:

> How is the choice of crowdsourcing platform related to the introduction and addressing of uncertainties? (SQ4)

# 5. Material and methods

As Bruno Latour and Steve Woolgar (1986) have shown in their influential early work *Laboratory Life. The Construction of Scientific Facts*, the day to day of scientific practice is considerably more messy and contingent than it usually gets depicted in (scientific) publications.

While the laboratory studied by Latour and Woolgar is different from my research project, one particular issue holds for the context of crowdsourcing data: Frequently, accounts and representations of successful procedures are portrayed as a result of a logical process when they are actually a lucky combination of analogical reasoning, experimentation and particular local circumstances. Similarly, manuals and guidebooks provide accounts of how things should be done, but people in practice often deviate considerably from these idealised and normative accounts (Suchman, 1995).

Along these lines, I noticed during my career as a computer scientist that many scientific publications depicted their use of crowdsourcing as a logical result of the problem formulation and established best practices. Other publications provided no or minimal details about the crowdsourcing process, effectively black-boxing the crowdsourcing, an observation validated by recent research on the topic (Geiger et al., 2020). This motivated me to investigate the actual, messy practice of crowdsourcing as practitioners do it and not to analyse textbooks and scientific publications.

## 5.1. Empirical Material: Interviews

Given the aim of this research, I decided to conduct qualitative interviews with crowdsourcing practitioners. I.e. I interviewed people who have themselves used crowdsourcing to generate data sets (or plan to do so soon). As I see this practice frequently tightly entangled with 'AI' use cases, I hoped to find respondents with the goal or potentiality that the data sets will be used as the basis for 'AI' systems. This could be, e.g., as 'knowledge base' that allows logical reasoning or as 'ground truth' for machine learning. These interviews help reveal facets, dead-ends, experimentation and local circumstances that may never be reported in publications and other representations of the practitioners' endeavours.

Qualitative methods are regarded as particularly suitable for gaining the in-depth understanding that I seek to establish (Silverman, 2000). Qualitative interviews with practitioners working on crowdsourcing fit the scope of this research well. I chose to conduct one interview each with a small number of practitioners working on different projects. This approach allows me to contrast and compare their approaches to address uncertainties and how they ended up using a particular approach. In combination, these interviews allow me to paint a differentiated and

multifaceted picture of crowdsourcing.

Semi-structured interviews are a fitting method to develop a detailed understanding of my interview participants' thinking, reasoning, and memories (Jensen & Laurie, 2016). Semi-structured interviews allowed me to focus on my research questions but be sufficiently flexible to react to the differences in the respondents' projects and practices. Not all respondents were at the same stage of the crowdsourcing process, and the use cases for crowdsourcing varied considerably. The choice of semi-structured interviews allowed me to pose follow-up questions relevant to the particular crowdsourcing project and skip questions that were not applicable. Semi-structured interviews allow the right amount of balance between structure and probing questions and allowing the interviewee's responses to guide the interview to some degree.

I conceptualise the data I obtained during the interviews as a collaborative result of my interaction with my interview partner (Finlay, 2012): My questions, utterings, and body language are all impacting the answers my respondents provide. Consequently, my role as interviewer has to be seen not as that of a detached, passive observer but as an active participant in the interview (Rapley, 2007). Similarly, the accounts of the interview respondents are not a "'reality report', never a merely a transparent window on life outside the interview" (p. 20) but a specific reality co-constructed by me, the interviewer, and the interviewee.

To put it into terms of Actor-Network Theory, I act as a mediator enacting particular actor-networks, not as an intermediary faithfully representing an actor-network 'out there' (Michael, 2016). This is particularly important in my case because I have been originally trained as a computer scientist with a focus on artificial intelligence, have spent several years as a researcher and developer in this field and was recently employed at the faculty of informatics at the TU Wien for two years. This didn't escape my respondents (and I didn't attempt to hide this information on ethical grounds), and I did my best to consider this circumstance when analysing the resulting data.

Ethnographic interviews inspired the interview questions (Spradley, 2002). This style suits my research questions and my theoretical approach very well, as Actor-Network Theory attempts to create "thick descriptions" (Geertz, 1973, as cited in Michael, 2016). Ethnographic interviews help to focus on actual practices that may be everyday practices and routines to my interviewees. Spradley characterizes several types of descriptive questions that are useful for answering my research questions. These are notably Grand Tour Questions, loosely modelled after a guided tour of a place, e.g., a university campus. Put differently, they can provide a big picture. Spradley notes that a Grand Tour Question can ask for a tour "through a sequence of events" (2002, p. 50). In my case, a Grand Tour Question provides an overview of my respondent's

crowdsourcing project.

In contrast, Mini-Tour Questions "offer almost unlimited opportunities for investigating smaller aspects of experience" (Spradley, 2002, p. 51), e.g. how specific episodes of the crowdsourcing project unfolded. A combination of these types of questions formed the core of my interview guide. A simplified version of my interview guide can be found in appendix B.

In total, I conducted 5 interviews, ranging in duration from 26 to 53 minutes. These interviews are my core empirical material. During the interviews, respondents also suggested reading articles and looking up additional information, e.g. on specific concepts. These documents, as well as my initial exploratory research on the topic of crowdsourcing of data sets, are distinct from this core material. They are background information to provide context to my analysis (Jensen & Laurie, 2016).

### 5.1.1. Recruiting

I aimed to recruit about 6–8 respondents who have used crowdsourcing to create data sets or are in the planning stage. Initially, I was focused on finding respondents who do this intending to use the data for 'AI' applications. It turned out that many respondents who I could find were focusing on crowdsourcing. 'AI' was largely a potential future use case, but most respondents had no immediate plans to use the data for this purpose. Thus it was not a strict selection criterion. The issues that I set out to address were present regardless of future uses for 'AI' applications.

My initial plan was to recruit respondents from the area of Vienna or easily reachable by public transport from Vienna. This way, face-to-face interviews would be possible while travel time and expenses would be kept minimal. I wanted to find interview partners at universities that have research units working on machine learning or other artificial intelligence techniques: They could be using crowdsourcing to create suitable data sets. The following universities fit this profile: TU Wien, University of Vienna, Vienna University of Economics, Johannes Kepler University (Linz) and the University of Applied Sciences Upper Austria. Each of these universities provides courses on machine learning. Recruiting was done primarily by contacting people at research units that appeared promising, i.e. where I could reasonably expect that someone was involved in such an undertaking.

In addition, I attempted to recruit at relevant meet-ups, gatherings of professionals, academics, and interested audiences with a focus on knowledge sharing and networking. In Vienna, meet-

ups on the topics of data science[5], deep learning[6], and software for data-heavy applications[7] meet more or less regularly. I attended some of these meet-ups, put up flyers with short information about my research and contact information, made short announcements if possible, and took the opportunity of the networking part (usually after the talks) to find respondents.

Finally, I attempted to find respondents from industry via AI Austria, the Association for the promotion of Artificial Intelligence in Austria[8]. This is an Austrian think tank and networking association of organisations active in 'AI'.

Fortunately, my recruitment and interviews took place before the COVID-19 pandemic hit Europe and made these endeavours even more complicated. Unfortunately, my plan didn't work out quite as successfully as hoped. First, it was hard to find practitioners that conducted crowdsourcing. My initial recruitment strategy was e-mailing machine learning departments and researchers at the mentioned universities. However, there were few replies, and those that I got were negative: No one seemed to do crowdsourcing. Until today, I am not sure if some of the research groups did do some form of crowdsourcing but didn't see it as the focus of their research and thus didn't see themselves addressed. After several attempts, I abandoned this unsuccessful approach.

In parallel, I attended some of the mentioned meet-ups. Unfortunately, while meeting regularly, the frequency of these meet-ups is not high, reducing the number of opportunities. Additionally, many facets of the topics covered by the mentioned meet-ups do not involve crowdsourcing. One meet-up was fruitful, and I could recruit one respondent. My recruitment attempt via AI Austria was unsuccessful, even though they promised to send out my call for respondents via their newsletter.

In winter 2019/2020, I made a more focused attempt to directly address individuals whose research was in machine learning, had publications that mentioned crowdsourcing, or were in some other way potentially using crowdsourcing (e.g., by being involved in projects or lecturing courses where crowdsourcing could matter). This way, by personally addressing the potential interviewees instead of more or less randomly addressing department heads and members, I managed to find more respondents, sometimes by referral from colleagues. One potential respondent wanted to first talk off the record to figure out what it was about – a talk that was insightful and took almost 90 minutes. Unfortunately, a proper interview didn't work out for various reasons (scheduling, being out of the country, etc.).

Eventually, I could conduct five interviews with practitioners. However, none of them was

[5]Vienna Data Science Group Meetup, https://www.meetup.com/Vienna-Data-Science-Group-Meetup/
[6]Vienna Deep Learning Meetup, https://www.meetup.com/Vienna-Deep-Learning-Meetup/
[7]Future of Data: Vienna, https://www.meetup.com/futureofdata-vienna/
[8]https://www.aiaustria.com/

active in industry projects. Most were academics at various stages of their career (PhD to senior post-doc) and one a freelancer, but working for an NGO and not for profit. Fortunately, the cases turned out to be very diverse, and I am content with the resulting material for the scope of this thesis. In section 5.1.3 I describe these cases in more detail.

### 5.1.2. Ethical considerations

All interviews were conducted face-to-face and audio recorded. All interviewees signed an informed consent form before the interview started, included in appendix C. Besides the usual content, I also disclosed that I might discuss the material with my supervisor and student colleagues.

While the content of my interviews is not overly sensitive, I still anonymised the interview material for this thesis as well as possible. It may still be possible that people familiar with my respondents' work can identify them successfully. This caveat was also made transparent to my interview partners.

### 5.1.3. Description of interview material

The following provides a short description of the cases covered by my interviews.

- **Interview 1: Verification of computer science models.** This interview was about a crowdsourcing use case that should allow an expert task related to formal modelling in the domain of computer science to be done in a parallel and distributed way. The crowd workers had to assess the correctness, or if deemed incorrect, the defect type of a model based on a textual description of what should be modelled. The setting is an Austrian university, and the crowd workers are computer science students. A particularity of this case is that students act as workers. Until the time of my interview, only data was used for crowdsourcing where experts had already set the correct results. I.e. there was little reason for uncertainty. But, since the goal is to apply this process to new data, how to deal with uncertainty in the future, how to combine different workers' responses in the absence of the correct answer is of interest to my interview respondent. This case will henceforth be denoted as C1.

- **Interview 2: Identifying and classifying offensive social media posts.** This interview was about a crowdsourcing project conducted by an NGO in an EU country. The aim was to identify offensive posts on social media and subsequently classify them. Crowdsourcing was a more systematic approach than previous projects by the NGO. The

workers were volunteers of the NGO. Their judgements were combined, based on certain rules, with a board of experts judging tricky cases. This case will be denoted as C2.

- **Interview 3: Detection of changes in audio recording.** This interviewee described two separate crowdsourcing projects they conducted. The first project, C3a, was about identifying when singing occurred in, e.g., opera recordings. The second project, C3b, was about identifying talk and music in radio recordings. In both cases, the goal was to identify if talk or singing happened and where it happened in a recording. For both projects, my interviewee paid students to do the work.

- **Interview 4: Improving support for crowd workers.** This respondent is researching crowdsourcing, in particular how the support for crowd workers to do the tasks can be improved. The task to solve by the crowd is the annotation of medical studies, particularly the patients' conditions, the medical interventions, and their outcomes. The platform used by this respondent is Amazon Mechanical Turk, and since the crowd workers are not expected to be medical experts, they need support to do the work well. This case will henceforth be called C4.

- **Interview 5: Qualitative Annotation in the Digital Humanities.** This respondent is a developer of a crowdsourcing tool for qualitative annotations of historical documents. The tool was developed for work that cannot be split into small, atomic tasks but requires context. As such, this interviewee's work sets itself apart from that of the other respondents. This interview was less about one particular project but about several projects that the tool developer has been involved in and that used their tool. This case will be called C5.

## 5.2. Situational Analysis

In order to analyse the gathered material, I used *Situational Analysis*, as developed by Adele Clarke (2005). Like Grounded Theory (Charmaz, 2016), an inductive analysis method pioneered by Glaser and Strauss, Situational Analysis is a method that builds on the coding of (textual) data. It is different from Grounded Theory (in the strict sense) with respect to the role of theories during the analysis: Situational Analysis does not pretend that theories emerge solely from the data but that theories always guide the analysis.

Situational Analysis provides a set of techniques that help to structure the codes, called maps (Clarke, 2005). These analysis aids and the pragmatic stance with regards to the emergence of theories from the data make Situational Analysis well suited to small research projects such as

this thesis.

Like Grounded Theory, Situational Analysis ideally builds on coded data but allows to create the maps without proper coding (Clarke, 2005) or to code only particularly interesting parts of data, making the analysis more time-efficient than classical grounded theory. I did a line by line coding for my research but didn't apply a code to each line. To make this task easier, to really focus on individual lines, I wrote a small piece of software that displayed only one line at a time and allowed me to enter a list of codes.

Coding and the creation of situational maps were done in two iterations. A first version was created after the first two interviews. This was done to get a first grip on the material gathered at this point in time. This initial, preliminary analysis informed my later interviews, leading to slight adaptations of my interview guide.

The first round of initial coding was very close to the data. I used this round also as a way to (kind of) translate German data into English codes. During and after this initial coding, I started to write memos about observations during the coding process.

A second round of coding was done after I did the initial coding for a subset of the interviews, partially informed by the memos. In this round, I focused on parts of the data that are relevant to the research question and attempted to abstract more than in the initial round. This was done, in part, to make comparisons between the individual interviews easier. This was, again, done with my small application.

Eventually, I imported all my codes from the second round into RQDA[9], a qualitative analysis package for the open-source statistical analysis software R. This allowed me to attach memos to codes and transcripts. Most importantly, it allowed me to organise the codes and material, e.g., by merging codes and adding codes to categories, features that I took advantage of to develop the final set of codes.

From the set of maps that Clarke (2005) developed, I focused on situational maps. These maps are a collection of "the most important human and non-human elements in the situation of concern" (pp. 86–87) based on codes that were formed from the analysed material. The importance given to non-human elements fits well my theoretical approach using Actor-Network Theory. In my case, the situation is the production of data sets by means of crowdsourcing.

I first created a messy situational map. It contains all the elements in the situation. I populated this map based on the codes from the second round of coding. As can be seen in figure 2, this map was pretty crowded. At the same time, it does not contain all codes. Instead, it forced me to reflect on which codes represented elements in the situation.

[9] https://rqda.r-forge.r-project.org/


Figure 2: This is the messy situational map after coding all five interviews. It contains the most important human and non-human elements of the situation.

Based on the messy map, I created several relational situational maps. For each of these maps, one important element is selected, and relations with all other elements and the type of relation is analysed. Again, this form of analysis is a good fit for studying actor-networks, with the focus being put on relations. I was more selective for these maps, only creating relational maps for the most important elements in the situation, e.g., 'atomic task', 'aggregation', and 'motivation'. Based on this map-based analysis, I created additional memos.

Based on the memos written throughout this analytical process and the analytical lens that Actor-Network Theory provides, I developed the analysis of the empirical results that I present in the next two sections.

# 6. Empirical results

In the tradition of Actor-Network Theory, I follow network builders: All my interviewees conducted some sort of crowdsourcing, and to do so, they built an actor-network. The networks that I am tracing based on my material are thus heavily influenced by my interviewing, the questions I chose to ask my interviewees, and the selection of interviewees. In this sense, my analysis is itself performative, or, as Michael (2016) puts it, my analysis can itself be understood as a mediator – I do not simply reproduce what's real, but by choice of methods, material and my analysis I draw a particular picture of crowdsourcing.

All interview participants identified their work with crowdsourcing, though most interviewees added some caveats and made distinctions (more on that in section 6.2.1). The crowdsourcing of data sets can be summarised as augmenting data with additional (meta)data. This augmentation, e.g., through labelling and annotations, is done to make sense of it to address a problem in a particular way. This also means that some kind of data set has already been produced before the crowdsourcing process starts, but for some reason, this data is not yet useful to address the problem. Obtaining additional metadata about the data by, e.g., describing objects visible in an image or classifying if social media postings are hate speech, adds utility to the data. This, in turn, either enables the analysis of the data itself or allows the data to be used for 'AI' applications, as described in the introduction.

From my interviews, a few core elements of crowdsourcing became clear, shared among most of the cases covered by my interviews:

- The **problem** to be addressed with crowdsourcing. This is, e.g., the effects of medication on patients as described in medical studies (C4), or the question of how widespread offensive content and hate speech is on social media and how frequently different kinds of offensive content occur (C2).

- **Original task** is a formalisation of the problem so it can be addressed with the available data. This can be, e.g., the classification of the presence and type of hate speech in social media (C2), the verification of a computer science model (C1), the annotation of core information in medical studies (C4), or detecting the presence of talk in radio recordings (C3b). As noted before, this can be generalised as augmenting the data set to increase its utility.

- A **data set** on which the task should be solved. What data set is to be augmented through crowdsourcing is informed by the problem. The data is frequently text (C2, C4, C5) and images (C5), but also audio recordings (C3a, C3b). In one case, it was figures

that represented a computer science model (C1). Generally, it is large(ish) data sets consisting of many data points (or documents) of homogeneous format that should be handled uniformly during crowdsourcing.

- **'Atomic tasks'** are the result of breaking down a larger task into small, quickly solvable tasks. Each 'atomic task' generally applies to one individual data point. This ranges from annotating three pieces of information such as patients, medical intervention and outcome into three separate tasks, each involving annotating only one part (C4). In other cases, a larger piece of information gets decomposed into the smallest parts, and single- or multiple-choice questions are asked about this part (C1, C2). As a tendency, the goal is to make the task as small and granular as possible – sometimes this is constrained by the context necessary to solve the task, as I will detail in section 6.4.3.

- **Distribution of tasks** to crowd workers is another core element. The 'atomic tasks' have to be distributed to crowd workers, who usually work on a set of tasks. This generally also involves access management – to regulate to whom the tasks get distributed, respectively, who gets access and can work on the tasks. To make distribution possible, infrastructure is necessary.

- **A platform** provides this infrastructure. It allows to create user interfaces that display the 'atomic tasks' to the crowd workers. In my interviews, the platforms ranged from the well-established Amazon Mechanical Turk (C4), makeshift appropriation of existing technologies not intended for crowdsourcing (C2), to creating dedicated platforms to varying degrees of sophistication (C1, C3a, C3b, and C5). Section 6.3 will take a closer look at the role of platforms.

- **Crowd workers** that solve the distributed tasks. Interestingly, only interviewee 4 positively characterised the people who solve the tasks as a 'crowd', something that is, so it seems, identified with laypeople. The remaining interviewees preferred to see their workers as experts to some degree. Section 6.2 will provide more information about the 'crowd'.

- An **annotation**, or augmented data point, is the result of a crowd worker solving an 'atomic task'. If an 'atomic task' is sent to several crowd workers, multiple augmented data points get produced.

- Often, a **combination mechanism** is used to create a final data set from multiple annotations (C1, C2, C4). In other cases, they are combined implicitly (C3). In C5, a negotiation among workers takes place, resulting in only one final annotation per data point.

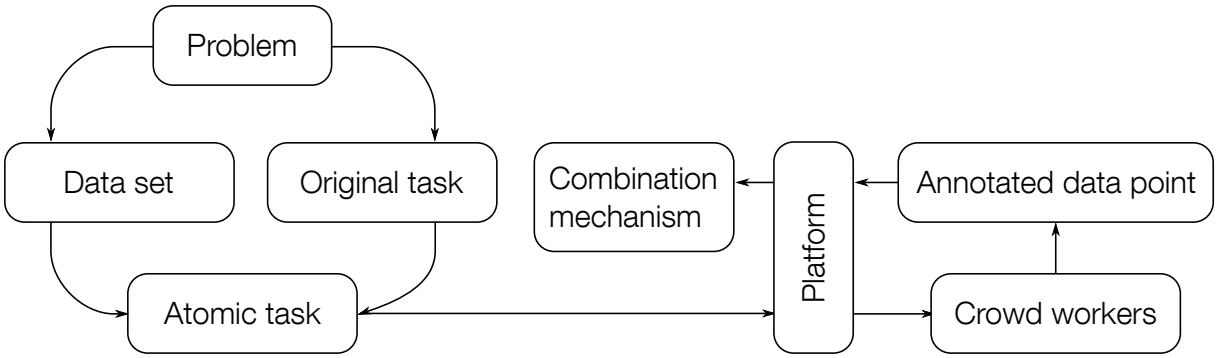


Figure 3: The core elements of a typical crowdsourcing situation. The problem informs what data set to produce as a basis for crowdsourcing. Based on the problem, an original task has to be formalised. For crowdsourcing, this task must be translated further and combined with data from the data set to form an 'atomic task'. These 'atomic tasks' can be sent via the platform to crowd workers who create annotations. These annotations are collected via the platform and eventually get combined.

Figure 3 shows these elements schematically. As I will show throughout this chapter, this actor-network is fragile, and a stable crowdsourcing process requires careful adjustments, experimentation and negotiations to make the crowdsourcing process successful. Practitioners have to solve several problems along the way: How can the problem be formulated into a task? How can this task be combined in a meaningful way with the data to form individual 'atomic tasks' that can be displayed to crowd workers? How can I get the crowd workers to complete the tasks as I intended it? How can I then create the final data set from the results of the crowd workers?

## 6.1. In quest of 'ground truth' – what constitutes uncertainty in crowdsourcing

As I am exploring how uncertainties can occur during crowdsourcing processes, I want to investigate what exactly can become a source of uncertainty in the context of crowdsourcing of data sets.

But first, let's take one step back. People turn to data to answer questions, to solve problems: Let us turn to social media to identify how frequent hate speech is. Let us annotate geographical locations in historical maps so they can be related to contemporary places. Let us check computer science models to ensure that they match the textual descriptions. In these cases, data should address the problem. But sometimes, it is 'AI' that should be used to answer these questions, which in turn needs data to learn from. In both cases, even if the data set is well-defined and the data itself is of good quality (whatever this means for a particular practitioner or application), it is often not yet in a state that could answer the question: The data needs to be processed

further to make sense of it.

Practitioners can turn to crowdsourcing as a means to process the data. Crowdsourcing augments the data with additional information in the form of labels or annotations, which help solve the problem. For example, interviewee 2 tried to identify offensive content related to human rights topics in social media postings. The questions asked of the data are thus: Do the social media postings contain offensive content? And if so, what kind of offensive content do they contain? Collecting social media postings in and of itself cannot answer these questions.

Similarly, interviewee 4 wants to identify where in a medical study's abstract patients and their conditions are mentioned. In this case, the information is present, but its location is unknown, or at least not formalised in a way that would allow a computer program to identify the location. Crowdsourcing, then, is itself a means to address uncertainty about the data by producing metadata (i.e. data about the data).

The production of metadata that helps addressing the problem is, however, not straightforward to accomplish. Each of the elements and their connections depicted in Figure 3 is fraught with uncertainties: How can the problem be translated into a suitable data set? How can a task be formalised that works with this data set? How can this original task, in turn, be translated into 'atomic tasks' for 'crowd workers' to solve? How to set up the platform in the correct way? Who is, or should be, the 'crowd', and how can they do good work? How should one proceed with the augmented data points? In the remainder of this subsection, I will look more closely at these elements and how they can introduce uncertainty.

### 6.1.1. Data as a source of uncertainty

As stated, practitioners turn to data to solve a problem. However, turning to data to answer the question makes the data problematic: What was once an abstract of a medical study becomes a problem: Where are the patients mentioned? Where in the text is their condition located? Similarly, a radio recording is no longer simply the recording of some radio show but poses a problem: Where do people speak? Where in the recording is music?

But let us look at this topic differently: What happens is that the data points have to be *enacted* in a way that fits the problem formulation. What might be enacted at a different stage as a radio show, as a form of entertainment, has to be enacted as a sequence with sub-sequences featuring human talk. What is somewhere else enacted a social media posting meant to share an opinion, to communicate with friends, families, and other publics, is now enacted as a data point transporting a certain kind of offensive content – or content that is not offensive at all. The enactments of these data points are *located* in a particular context – that of the task to

solve, the problem to address, and here it becomes a source of uncertainty.

To exemplify this further, assume that the task was to count the characters in the social media posting? Suddenly, the posting would be enacted as a string of characters with a clear length, a task easy to solve. The task could even be solved without involving a 'crowd' (but a piece of software, an algorithm, instead). Hence, one has to take the whole *task-data package* into consideration to deem the data a source of uncertainty.

But there is another layer to consider. In some cases, the practitioner who turns to crowdsourcing could solve the task themselves. It is merely that the amount of data is so large that doing it on their own would be too costly. To put it differently, it's a matter of scale why one would resort to crowdsourcing.

In the context of model verification, as conducted by interviewee 1, there appears to be a 'correct' way of solving a task. Based on established standards, it should be clear if a model is 'correct' or has defects and what kind of defect. In two interviews (1 and 4), this expected solution (i.e. 'ground truth') to the task was available. It was computer engineering and medical experts, respectively, that defined what the expected result is:

> "So in general, most of the time we know what is correct and not, also because this was a kind of . . . teaching exercise, so there was a well-known gold standard. So we had a correct model, and then the model with defects, so we knew exactly [where] there were defects . . . so most of the time we knew exactly what we expect for each task" (interview 1)

> "The data that I let the crowd workers annotate are already annotated by medical experts. That means I know already exactly, for each example, what the correct answer would be"[1] (interview 4)

As I will explain later, a (partial) data set where the 'correct' solution is known plays a crucial role at various points of the crowdsourcing process. More important at this point is that I understand the tasks in these examples to be presented as *matters of facts*. In both cases, the (scientific) community managed to unify what a 'correct' model is like and how one 'correctly' identifies the relevant information in study abstracts.

In some cases, however, the data can be a source for uncertainty even for the practitioner, e.g., because a data point is ambiguous, i.e. it could fit several categories, or because it is an anomaly, i.e. it fits no category well (Douglas, 1966/1969; Lupton, 1999). Thus, it can be, quite simply, difficult to decide how to solve the original task on the data, and hence the data itself becomes even more problematic. Additionally, ambiguities and anomalies can pose a problem

for the subsequent analysis or the 'AI' methods that should subsequently learn from the data because they often need clear instructions on what to learn. Thus, applying multiple labels in the case of ambiguities is often not an option.

Interviewee 2 remarked that they "[have] observed that many politicians are skilled in making statements that are not unequivocally offensive, but that are really on the line"[2] Here, the data points are clearly *matters of concern*: Is this social media post to be enacted as hate speech? Is that posting to be enacted as benign, maybe as sarcasm? Where should the line be drawn between insult and hate speech? These questions are not settled; they remain *disputed*. As I will show, in these cases, crowdsourcing can provide a benefit beyond scalability.

As mentioned above, in some cases, experts (that may not be directly involved in the crowdsourcing process) define a 'correct' label or annotation. In the context of crowdsourcing, these experts annotate a small data set with what is deemed the 'correct' answer – a so-called 'ground truth' (interview 3) or 'gold standard' (interviews 1 and 2) for the crowdsourcing process (see also Jaton (2021b)). At first, this may sound confusing because often crowdsourcing itself is used to produce a 'ground truth' (especially for 'AI' applications). The difference is that, usually, the 'ground truth' for crowdsourcing is only a small data set, whereas the analysis of social media over months concerns a (very) large data set. Likewise, machine learning algorithms often need tremendous amounts of data to be trained successfully. Hence it is again a difference of scale: crowdsourcing can require a small 'ground truth' data set (sometimes only a dozen data points) but can subsequently produce a (much) larger 'ground truth'.

Interviewee 2 also touched on a different way how data can become uncertain. When asked if they intend to use the crowdsourced data set for machine learning, they were cautious because the data gets produced in a particular historical context: "there has to, at least from my point of view, there has to be continued rating, eh, because what is up-to-date this year and what can be determined with the algorithm this year may be incorrect in two years, one has to continuously readjust."[3] This sensitivity for the historical context where data sets get produced didn't come up in other interviews. This points towards potential issues with many data sets that embed some form of temporal sensitivity, something that I'd argue is present in pretty much any kind of data: After all, the modelling of computer science problems will certainly change over time, just like the style of abstracts of medical studies may change, too. This puts limits to the re-usability of data sets in a temporally different context.

### 6.1.2. Task formalisation as a source of uncertainty

What should be done with the data can become another source of uncertainty in the crowdsourcing process. At first, the broad goal may appear clear: Identify defects in a conceptual model as well as the kind of defect (interview 1), identify offensive social media postings (interview 2), identify speech and vocals in audio recordings (interview 3), identify the location in medical studies that identify the condition, treatment and outcome (interview 4), and transcribing historical maps (interview 5). At a second, closer look, these tasks become fuzzy and increasingly unclear when they have to be formalised, and instructions for crowd workers have to be developed. How this can be achieved the best way is often unclear:

> "[T]he end goal of this would be to try to come up with a distributed process that big companies can use, you know, when they have to check two hundred ... pages of specifications against models, but also in several engineering disciplines you often have some kind of model and textual description or even no textual description, just some expert knowledge, and to we want to see whether you can do this in a more organised way" (interview 1)

Interviewee 1 is confronted with a model verification task that can include hundreds of pages that have to be compared to formal models. However, the broader application may even depend on implicit expert knowledge that has to be made explicit so it can be appropriated for crowdsourcing. Here, crowdsourcing is not only a way to solve a task by a crowd, but also a way to make a task more 'organised'. A translation has to take place, from the original task to something different. Interviewee 1 called these tasks 'atomic tasks'. Achieving this translation involved discussions with model verification experts: "this is actually a complex task, this ... takes up most of the preparation, and, first we discussed this with our domain experts, so this were the software engineering guys, ... what makes sense" (interview 1).

Whether this translation poses an issue is heavily dependent on the particular project. For interviewee 3, their projects had clear tasks that needed little work to be usable for crowdsourcing: "Those are tasks that are relatively easy also for non-experts. It was about, for example, to detect in an opera recording when someone is singing and when not, or to detect in radio recording when music is being played, when there's talk and when there's no talk."[4]

Pinning down the precise task to be solved by the crowd in the eyes of the person setting up the crowdsourcing is one issue. A related, but different issue, appears to be to make sure that the crowd workers conduct the task correctly, that they *read* the task correctly. This issue has come up in several interviews (interviews 1, 2, and 5). To illustrate this point, interviewee 2,

who worked with NGO volunteers as the crowd, did extensive piloting to make sure that the task is clear to the workers:

> "We had one training, then a pilot phase where we collected data for four weeks and rated it, and used this data to see, okay, which questions make sense, how did it run for our volunteers, what was easy for them, what was difficult, what went somehow wrong, to then redesign all the apps, algorithms, the data collection and the volunteer training with this information"[5]

Similarly, ensuring that the crowd workers correctly read the task provided the core of interviewee 4's research. They specifically wanted to improve the task design of prior research:

> In "this study that was conducted two years ago, they had plunked down a ten-page document to teach the [Amazon] Turk workers how it works. There it was written how exactly the task should work, ... this means three pages for the participants [of the medical study], three pages for the interventions, how one annotates this."[6]

At first glance, providing extensive guidelines on how to solve the task may seem appropriate to make it clearly defined. However, interviewee 4 seemed to disagree and saw this as a research opportunity. They continued:

> "And of course, nobody reads this, particularly not the Turk workers. And these ten pages contain lots of examples – it shows, for example, a sentence and one particular part is highlighted, and that should show how it works. And the big issue that I spotted is that you have this huge document of ten pages and you have to scroll through it and find which example fits well"[7]

Interviewee 4 sums this issue up quite nicely at a different point of the interview: "One doesn't want to overburden [the crowd workers], not to provide too much information, because the descriptions should be very short, precise, accurate and one doesn't want to overload them with information."[8]

Addressing tasks as a source of uncertainty is, in summary, an undertaking that needs a clear formalisation of the task and a delicate balance of providing the correct amount and type of information to make it clear to the crowd workers what they should do. At the same time, one has to prevent bloated instructions containing too much information, not all of which is relevant to the particular task and data point. Failing to achieve this can result in uncertainty creeping into the crowdsourcing process: It might be that there's no shared understanding between the crowd workers and the initiator, that the 'atomic task' fails to keep its meaning as it circulates,

or that due to information overload the crowd worker has difficulties 'reading' the task correctly. If this happens, it introduces uncertainty as it's unclear what task the crowd workers were actually addressing, if it is the same as the initiator expected them to.

What is at stake is to make sure that the 'atomic task' in combination with a data point can successfully work as an *immutable mobile*. Only if this is ensured can crowdsourcing become a success. As I will show in section 6.4.1, sometimes the 'crowd' gets trained to make sure that they 'read' the atomic task correctly.

### 6.1.3. Crowd as concern

One of the key elements of crowdsourcing is, as the name suggests, a 'crowd'. In all but one interview, the 'crowd' was not an anonymous 'crowd' of strangers, but either students (interviews 1, 3, and 5) or NGO volunteers (interview 2). Only interviewee 4 used Amazon Mechanical Turk and made the tasks available to users of the platform. Given the variance in crowd composition, the interviewees raised different concerns about the crowd. For interviewee 3, using commonly available crowdsourcing platforms, such as Amazon Mechanical Turk, was not the preferred option, and one of the reasons was the 'crowd':

> "Interviewee: First, it was not clear how good the quality of the data would be one would get
> Interviewer: Mhm, what do you mean by that?
> Interviewee: That there are people who quickly click through somehow to earn some money, and one has to devise some system, like an entry check, or intersperse examples throughout where one knows the answer, to check whether they're still paying attention"[9]

What this interview segment shows is the uneasiness to let strangers annotate the data. Trusting the workers to do good work appears to be important. In the case of an anonymous crowd, checks and a certain level of control have to be introduced that compensate for the lack of trust. I assume this lack of trust was less of an issue with the university students that the interviewee preferred to work with, as no checks to this extent have been found necessary in that case:

> "Interviewer: How did you know if [the data] was actually well annotated or not? Did you do something .. ?
> Interviewee: No, I didn't know. It's been a while, I believe I looked a bit at what they clicked for these hundred examples, and I also annotated these and looked if there were serious discrepancies. It could also be that I mailed one annotator and

> asked, hey, what about this here, I would have expected something different, but otherwise, I believe I may have randomly looked at the data ..., but everything looked reasonable"[10] (interview 3)

Interviewee 4 who worked with an anonymous crowd shares the concern towards so-called 'scammers' that don't solve the tasks in honesty: "And that's how I want to filter out that some scammers work on it that simply click through quickly."[11]

But even if the workers are trusted, there are other reasons how the workers can become a concern: "And then [the understanding of the task] also depends on [the] understanding of English as well, because this was in English" (interview 1). Interview 1, in particular, required some prior knowledge for the task, as a skill level self-assessment questionnaire indicates:

> "[T]he other is a self-self assessment, so before they go into this experiment, they have a questionnaire of, I don't know, twenty, twenty-five questions that ask them different things, from familiarity with [the] English language, whether they used crowdsourcing tools, whether, you know, what level of software engineering they are, whether they've done model verification, and so on and so forth." (interview 1)

It becomes very explicit that it is important that the crowd workers 'correctly' read the 'atomic task'. The language itself apparently was cause for concern if the reading would be 'correct'.

Ensuring the 'correct' reading is non-trivial, irrespective of trust in the workers. Even if the students command the language well, an initial attempt to use a 'controlled language' failed: "So we experimented with kind of a controlled language, where they ... should use a certain way to describe these defects, and it didn't work ... because they, they just didn't do it" (interview 1) What this shows is an 'unruly' crowd, deviating from the *script* of the task (Akrich, 1992). Consequently, this crowd needs to be tamed, e.g., through improved task design, as I will show in section 6.4.2.

In two interviews, the crowd didn't come up as a problem. Interviewee 2 showed novices that had just started with the crowdsourcing task a set of questions where they had designated one answer as 'correct'. Not always did the workers agree with these 'correct' answers. However, interviewee 2 took a clear stance on the question of whether this presents a problem:

> "The only thing where one could say there's right and wrong was this hate speech, and that's why these were always going to the board. But in the other cases, one has to admit that what insults you may leave me cold and the other way round. Thus, I find a certain amount of variation good, and one has to keep it."[12]

In this case, it was welcome for the crowd workers to interpret the data, within certain boundaries, according to their own standards. More fundamentally, in this case, the volunteers have been fully trusted, as interviewee 2 states when asked about this:

> "Interviewer: But basically, the volunteers have been assessed as trustworthy, that they have earnest interest to participate in the spirit of your work?
> Interviewee: precisely, precisely."[13]

Interviewee 5 develops a crowdsourcing platform that is directed towards small teams and use in teaching. There, it was either peers that collaborated and implemented a kind of peer-review system, or sometimes a supervisor-student relationship was in place, and it was up to the supervisor to eventually resolve any remaining disagreements and uncertainties:

> "[There was mutual control happening], i.e. it was always, there were always small teams, two to three people, who were working on, e.g., an atlas, and there was usually one the main expert and the others were typically PhD students. Thus this was more an informal way of control through a regular supervision relationship."[14] (interview 5)

Trust issues or the desire to closely supervise the workers (students and researchers in this case) never came up during this interview.

To sum up, each 'atomic task' comes with an inscribed script that the crowd workers should follow. How strictly they have to stick to a script varies between crowdsourcing projects, as my material shows. Besides this, it is important to ensure a 'correct' reading of the task. There are, broadly speaking, two reasons why this can fail: First, suspected bad intentions on the side of the crowd worker, or a badly designed task that does not faithfully transport the intended meaning and thus can easily be misread.

### 6.1.4. When aggregation fails

Crowdsourcing projects frequently involve the annotation of an individual data point by multiple people. This was also the case in most interviews, in some for all data points (interviews 1, 2, 4, and 5), but at least for a small subset of the data (interview 3). Interviewee 4 nicely describes the motivation for multiple annotations per data point, in their case each being a sentence:

> "But I do – normally one does this to be able to do majority voting, to stack three sentences and look, okay, two say this, one says that, so what the two say is right. But I'm not doing it for majority voting, but simply to get rid of a little variance,

> because if I ask only one [worker], per sentence, then it could be that it's a funny coincidence that they agree with the medical experts."[15]

As the crowd is usually laypersons or experts in training (students), there's the underlying assumption that they can make mistakes, something that experts do in rare cases only: "Of course these medical experts also make an oversight every now and then"[16] (interview 4).

The above quote also implies a relationship between the design of the 'atomic task' and the resulting agreement ("look, my tool is really very good"). But it's not only the result of the aggregation that is entangled with the design of the crowdsourcing process. The question of how aggregation is even possible is highly dependent on the design of the 'atomic task', as interview 1 nicely shows:

> "So you get a model, a text, an element and you have to say whether there is a defect or not, and of course, this is very hard to aggregate because if you have just text, you can't really aggregate these things, so if you have three different students just putting in natural language what they think and in the end you have nine hundred judgements, ah, overall this is really hard to aggregate – so we experimented with asking them to use a controlled natural language."

But even the usage of controlled language didn't work because the workers, i.e. students, didn't adhere to the controlled language, as already described in section 6.1.3. In section 6.4.2 I will describe how this problem was solved by interviewee 1.

In one project where the workers had to mark segments of an audio recording, interviewee 3 let most data points get annotated by only one person each. Partially this was because it would be hard to combine multiple annotations:

> "Besides, it's difficult with this task. It's not like I have an image and want to label what is in the image, and then can look at what the majority says; instead, it is, it is a music track where everyone marks five to ten boundaries, and it's not quite clear how one then combines these. Because what they do is identifying segments. To say this is a unit, that is a unit. This is not so; it's more difficult to combine the contradicting annotations by several annotators."[17]

In this case, the task was unfit for aggregation, or at least interviewee 3 didn't see a suitable way to make it work. Consequently, letting several crowd workers annotate one data point makes no sense: How, after all, should they be combined in the end?

To sum it up, if the process of crowdsourcing involves annotating each data point multiple times, all elements in the actor-network need to be well aligned, in particular, the data and the

design of the 'atomic task'. If this is not the case, how to make use of these multiple annotations causes uncertainty. Section 6.4.2 will discuss some strategies to tackle this problem, to make 'atomic tasks' fit for aggregation.

## 6.2. Doing the work: crowd workers

The so-called 'crowd', i.e. the people who actually solve the 'atomic tasks', are one crucial element of the crowdsourcing process. Going into this research, I wondered how the practitioners' imagination of the crowd impacts coping with uncertainties. This section will focus on three topics centred around the 'crowd' that came up during the interviews. First, implicitly the question came up when a 'crowd' is a 'crowd' and when it is considered not to be a 'proper' crowd. Second, in several interviews, the working conditions of the 'crowd' came up, impacting the practitioners' decisions. Third, and related to the question of working conditions, is how workers organise themselves. Finally, I will touch on the question of crowd demographics.

### 6.2.1. Crowd work and boundary work: Who is a crowd, who are experts?

Who is actually solving the tasks during crowdsourcing? Who are these people? In the case of interview 3, the distinction between crowd and experts never came up. But the majority of my interviewees was quick to engage in boundary work (Gieryn, 1995):

> "This is now a task [done] by very few people, and it is a bottleneck, so we try to use crowdsourcing mechanisms to expand it to [a] larger group of people, yeah, so we do some experiments where we use crowdsourcing technology, but for expert sourcing, yes, so our crowd is not really the laymen crowd, but they're students, there are sixty, seventy students" (interview 1)

While crowdsourcing *technology* is used, the tasks are not distributed to a layperson crowd but to more qualified people, which runs under the banner of 'expert sourcing'. At the same time, this quote implies a distinction between the 'real' experts – of whom there are only a few – and the experts taking part in the crowdsourcing. The term 'expert sourcing' was also used by interviewee 5:

> "Originally, the whole thing began because we wanted to transcribe maps, historic maps from a library collection ... they should be transcribed and geo-referenced; thus, that was really this expert sourcing task that was done then. That means it was people from libraries themselves who were familiar with the material, who created transcriptions of the location names and maps, and who ... assigned them

> to geographical points so that one could easily extract metadata from these maps. That was actually the main scope. And from this has been developed this tool and that gets used for various things, of which I'd say, only a small part is really crowdsourcing, it's rather a commenting function during courses or generally when historians work with sources and so forth."[18]

Here, teams of librarians and historians usually work on a shared corpus of sources, sometimes also using the tool during courses to let students work with the material. Even though it can also be used for crowdsourcing, this is not what it is usually used for. One of the reasons why this kind of task has not been crowdsourced to laypersons is the difficulty of the task:

> "...one has a huge map, with about eight, nine hundred location names on it, but these are for laypersons really hard to read. Me personally, I have seen many of them, but as computer scientist I can hardly read them, thus one really needs experience working with it"[19] (interview 5)

Interviewee 5, whose tool is also frequently used in training contexts, even uses the notion of 'top expert' to differentiate between senior academics and students:

> "[we] try ...to do this balancing act, where one says, we have experts, but we also want to involve people who are not top-experts but just students, typically, and offer a platform to focus this labour, more or less like crowdsourcing even if it's not directly crowdsourcing"[20]

The line between top experts and the students is implicitly drawn along the lines of academic degrees. The students in the case of interview 1 are also doing the work in the context of a university course where part of the curriculum is to learn the task they have to solve with crowdsourcing. In this teaching context, interviewee 1 also made double use of the existence of the expected answers (as described in section 6.1.1): "the other [motivation] is this learning analytics, so if you can involve the students and you really have everything very detailed and cut into small bits, and digitalised, you can give them immediately feedback". Thus, in this particular context, the crowdsourcing process is used to distribute tasks to experts while at the same time contributing to the formation of their expertise.

In interview 4, the roles between experts and laypersons were clearly assigned: The crowd is anonymous without any particular expertise, and almost any user on Amazon Mechanical Turk can participate. In contrast to this, medical experts define what the expected outcome of crowdsourcing is. At the same time, the work of interviewee 4 attempts at establishing a very particular relationship between the experts and the crowd: "it's basically only about figuring

out, can I get the workers to be almost as good as the medical experts, that is basically the challenge."[21] Interviewee 4's goal is to essentially level the disparity in expertise between crowd workers and medical experts for the task at hand. This should become possible with the help of carefully crafted support for the crowd workers.

A different approach was taken by interviewee 2: Instead of levelling expert judgement and the crowd's combined judgement, a cooperative distribution of work was devised: "[The NGO has] created a board, with researchers from universities, linguists, sociologists, and they came together as experts who then have the last say on the comments."[22] However, this was not done on all comments, but only in two scenarios. First, comments were sent to this board when three volunteers disagreed: "when not all submitted the same rating then this comment was forwarded to a central committee"[23]. Second, even if the volunteers agreed, in some cases this agreement was not deemed sufficient:

> "with two exceptions. If someone clicks 'hate speech' then the comment goes to the board. Because hate speech, it's delicate, I don't want to say it's hate speech, except if there's also experts who really confirm this. And the other [case] if someone says the comment is somehow ambiguous, or– then it also went to the board, thus these always went to the board who then had to decide is it really hate speech . . . , or did they simply not get it"[24]

What this interview describes, then, is a two-stage process where the crowd addresses a lot of clear cut cases and an expert board provides its expert judgement in three cases: Either the crowd cannot agree, or if someone thinks a comment is ambiguous, or, finally, even if the crowd agrees that it is hate speech. In the last case, a confirmation by the board is necessary.

A similar situation where a (more senior) expert confirms, or approves, the crowd's work is present in interview 5: "There was most of the time one the main expert, and others were PhD students, typically. Thus it was more informal control by means of a regular supervision relationship"[25]

To conclude, the relation between laypersons and experts is multi-faceted and varied across the interviews. Still, a few shared themes did come up. First, crowdsourcing hardly aims at involving (senior) experts but people with intermediate to no expert knowledge (interviews 1, 2, 3, 4, and to some extent interview 5). Second, one of the challenges of many crowdsourcing projects is providing support so that these people with less expertise can, not individually, but as many (the 'crowd'), achieve a similar quality to experienced experts (interviews 1, 2, 4). Finally, some crowdsourcing projects employ a cooperative model where less experienced people do a lot of the 'leg work' and (more senior) experts approve their work (interviews 2 and 5).

### 6.2.2. Working conditions

In some of the interviews, the working conditions of crowd workers came up, particularly the (assumed) working conditions of crowd workers on established commercial platforms such as Amazon Mechanical Turk:

> "[It is] a bit too complicated to set up [Amazon Mechanical Turk] and one does not quite know what you get, whether one gets good data with it and one exploits somehow people in precarious situations, that are all reasons that spoke against even trying it"[26] (interview 3)

At a later point in the interview, the respondent repeated the precarious situation as one reason not to use Amazon Mechanical Turk: "... and third, as I've said, because one exploits people in precarious situations that get very little money for somehow doing work for us"[27]. This means that there's a clear image of Amazon Mechanical Turk as exploitative workplace due to bad remuneration. This narrative was shared by interviewee 4, who imagines the crowd workers there as people "who maybe simply want to earn some money on the side because one cannot become a millionaire, and I don't know how well one can live from it if one does it full-time."[28]

This image of commercial crowdsourcing platforms as workplaces with low compensation is interesting because whoever puts up tasks on these platforms can themselves set the remuneration. Thus, setting high compensation would be possible. However, doing this would raise its own set of concerns:

> "A colleague told me that it's hard to set the right price. I have sentences, I can say for each sentence: if you annotate this, you get that amount of money. And the idea is if I pay more then I should get better quality. But what I've heard is that, basically, if one pays more, this makes it more attractive to [scammers], or for people who do it half-heartedly. Not very bad but fair, so that they get the money quickly. That means, one is not supposed to pay too much money, but on the other hand one must not pay too little money because they then think, no, that doesn't make sense, because for one cent I'm not going to do this, that's not worth it. That's something I found interesting, for example."[29] (interview 4)

Implicitly, there's an expectation to do the work on crowdsourcing platforms not primarily for the money but with a certain amount of passion and dedication for the task ("people who do it half-heartedly"). Getting the pricing correct to only attract workers with sufficient dedication appears to be a delicate balancing act, not offering too little compensation to seem offending, yet not too much to attract those without dedication.

In the case of interviewee 2, it was volunteers that did the crowd work. In that case, it was explicitly assumed that they do the work due to their own motivation and with dedication:

> "Interviewee: Honestly, if someone participates offhandedly, after (laughs) after one day he doesn't participate anymore because it is a very exhausting job, they have every day 100 comments or so, that they have to rate, and, yeah, that's somehow also psychologically a bit exhausting, because one ugly comment is sufficient to mess up your day
> Interviewer: Yes, yes
> Interviewee: Well, I have not even rated them myself, but simply during sorting and processing one sees some things, and–they, yes, they can .. cause some distress."[30]

In this case, dedication is also necessary due to the mental burden caused by the data to be rated during this crowdsourcing campaign.

### 6.2.3. When the crowd reacts: Worker organisation

Until now, I have only looked at the 'crowd' as a mostly passive workforce. However, throughout the interviews, and in particular interview 4, some moments of reaction, or even resistance to participate in the crowdsourcing came up.

First, the most obvious way of resisting the aforementioned expectation of doing the work with dedication is to solve the tasks as quickly as possible. This practice may lead to mistakes but can also involve intentionally clicking at random answer options. At least, this seems to be an assumption on the side of the person uploading the tasks to the platform, as described in section 6.1.3 where I discuss the crowd as a concern.

Second, workers on Amazon Mechanical Turk create their own communities:

> "And there are countless messaging boards, there's this Turknation or so, ... where the Turkers, basically the workers, communicate with each other, and they have their own community there, and they can also say, hey that one only puts online bullshit, we're not going to work for them any more. Thus, they are organised, and one has to keep this all in the back of your mind when you submit something to Mechanical Turk."[31] (interview 4)

As this quote illustrates, the crowd is reactive and has to be persuaded to participate. If the persuasion fails, workers could even boycott the person submitting the tasks to Amazon Mechanical Turk. Hence, the crowd can introduce sanctions, not only the task submitter, as I will discuss in section 6.4.4.

### 6.2.4. Crowd demographics

Three of my interviews had (mostly) students as 'crowd', some from computer science (interviews 1 and 3), some from the humanities (interview 5). The remaining interviews were NGO volunteers (interview 2) or workers on Amazon Mechanical Turk (interview 4). Here I want to detail some information about their worker demographics.

In the case of interview 2, the work was done by volunteers recruited by the NGO that initiated the crowdsourcing campaign:

> "Interviewee: I don't know more, we would have to ask [the NGO], but they divided that somehow regionally, they have chapters in each region and each region then finds within their region volunteers, and they really tried to make this distributed across the country, that it's not only rated centrally but that it's the whole country.
> Interviewer: Mhm, but that means they all were [NGO] members?
> Interviewee: Yes, members, or somehow recruited by them. Right now, I don't know if all volunteers are really members."[32]

What's of note here is the attention to regional distribution. While this was not stated during the interview, one possible reason for this regional distribution is that some debates on social media may be regionally specific. If it were only people in, e.g., the capital, then some of these regional issues could be evaluated differently. Another possible reason for the regional spread is a secondary goal of this crowdsourcing campaign: "and secondly, it was important to [the NGO] to involve the grass-roots level, that it's an action by the basis for the basis."[33] That it was volunteers that conducted the crowdsourcing work was beneficial as it was also a source of trust as described in section 6.1.3.

Interviewee 4, who used Amazon Mechanical Turk for crowdsourcing, had only vague ideas about the people conducting the crowd work. When asked about who the workers are, interviewee 4 had no clear image of the workers:

> "Well, in principle, anyone can register on the website .. I myself have no worker account. .. Difficult to say, but I know for certain, there are these qualifications, this means I could, for example, say that I only want people who have a Facebook account, I only want people with a High School degree, and so forth, that means, I think, that it's diverse, but it's presumably rather people who maybe simply want to earn some money on the side because one cannot become a millionaire, and I don't know how well one can live from it if one does it full-time."[34]

Interestingly, though, practitioners using Amazon Mechanical Turk can narrow down the po-

tential workers for their task, giving them some degree of control about the worker 'population' who can solve their task. Beyond that, as I have already described in section 6.2.2, there is a base assumption of Amazon Mechanical Turk workers looking for additional income, but not full time.

## 6.3. Crowdsourcing platforms: The role of infrastructure

As described in section 6, crowdsourcing platforms provide the means to distribute 'atomic tasks' to individual crowd workers. As such, they act as an intermediary that defines the roles of the involved actors (Callon, 1991): People become crowd workers by conducting work over these platforms, and my interviewees become so-called initiators or requesters. Depending on the particular platform, these roles may also imply becoming contractor and customer, respectively. Crowdsourcing platforms can also be prime examples of information infrastructure (Star & Ruhleder, 1996; Bowker, Baker, Millerand, & Ribes, 2009) as I will show in section 6.3.1.

In this section, I want to investigate these platforms in more detail. First, I will provide insights into how my interviewees chose a crowdsourcing platform. Second, I will look into ways of appropriating the platform to their particular need. Finally, I will look at the user interface that presents the atomic task to the workers.

### 6.3.1. Choosing a platform

Given the central role of the platform (see figure 3), as an intermediary between the initiator and the workers, choosing the right platform is crucial to the whole crowdsourcing endeavour. Consequently, it is interesting to know which reasons motivated my interviewees' choice of a particular platform. As I will show, the reasons and choices varied considerably. Broadly, the motivations can be subsumed under either the concepts of efficiency or established standards, as I will show.

The most common way to achieve efficiency is what I'd call convenience, largely convenience through familiarity. Interviewee 1, for example, had prior experience for the platform Crowd Flower (now running under the name Figure Eight), and that was the primary reason to use it: "Yes, yes, so the, this is actually, a kind of a – it depends on what experiences you have, in the [working] group." Being familiar with a platform means less effort to learn the platform's intricacies, which, quite simply, makes it faster to get it to work. But there was another reason to stick with the familiar:

> "So I had [an account] in crowd flower, and then I just [stuck] with it, so the goal of the research was not so much to evaluate different platforms, I just knew it existed,

so we stuck with that one."

Since the evaluation of different platforms was not the primary focus of this interviewee's project, the convenience of using a familiar platform and already possessing an account was important. It was not only convenience out of familiarity but also organisational convenience, as interviewee 1 describes a reason why Crowd Flower was chosen at a previous workplace:

> "So, aah-I had already, aah figure eight, it was Crowd Flower before, account from [a previous workplace], and then I was familiar with that, and we just used it. Also, because when I was at [the previous workplace], I remember Amazon Mechanical Turk was kind of hard to get an account because you needed a US bank account and what not!"

Here, the requirement to have a US bank account was a major stumbling block to the adoption of Amazon Mechanical Turk. Bureaucratic doubts also came up in another interview:

> "And it's not really clear, how one pays that, whether the university can simply .. pay that, the platform, because now we have set up some contracts for work labour with the students. I don't know what one– if it is a problem to buy something on Amazon Mechanical Turk, from what budget this would be paid"[35] (interview 3)

What is clear is that the crowdsourcing platform has to fit (more or less) easily to the organisational and regulatory requirements, as exemplified by the payment woes Amazon Mechanical Turk posed to these interviewees. Put differently, established crowdsourcing platforms are *embedded* (Star & Ruhleder, 1996) in particular payment infrastructures. Choosing a commercial crowdsourcing platform thus implies choosing a payment infrastructure.

Interviewee 2 used R Shiny, a software package related to the statistical software R, to create a crowdsourcing platform from scratch. The reason was not that they had previous experience using R Shiny for crowdsourcing, but being experienced with the R ecosystem, including R shiny:

> "Well .. let's put it this way, the app comes from me, and because I'm not, I'm not really an app designer, it's a Shiny app because that's what I can do and [the NGO] didn't have the resources either to come up with something fancy, hence they took what I provided"[36]

In this case, it was efficient to stick to existing expertise instead of investigating alternatives. Another reason for using R Shiny was the degree of control this allowed interviewee 2, as well as the integration into a familiar workflow centred around the statistics software package R, another way of making the crowdsourcing process efficient:

> "Well, the nice thing about Shiny is, firstly, I decide how, what's going in, how it's going in, and I can program it myself that if ... certain topics show up, they click there, then it automatically opens another window where I can rate additional things. The second is, it's very easy for me to download the data, sample them and put them into the app because I do all this in R. And if I had to do this in some Surveymonkey or something, then this would be an extra step, how do I get the data, that I have now sampled in a format into the app, how do I then save the data from this other app, and how do I then get them so I can process and analyse them. Due to all being in R, at least the workflow is easier."[37] (interview 2)

In this case, crowdsourcing was only one part of the project, the others being data acquisition (i.e. the downloading of social media comments) and the subsequent analysis.

In a similar vein, interviewee 3 preferred the familiarity and used a web application that had been used at their research group previously:

> "Interviewer: Well, why again did you decide against classic crowdsourcing with the common platforms like Amazon Mechanical Turk, or, I don't know, Crowd Flower I also know, what else is there, Task Rabbit I think, there are various
> Interviewee: Mhm. Well, first one would have to look into it on a technical level, what one would need to set it up
> Interviewer: Mhm
> Interviewee: Probably not much more than what I already had, if I already have the web interface, but I never looked at it"[38]

Here it was, again, the availability of existing tools and the familiarity with them that made it an effective choice. Additionally, as already described in section 6.1.3, concerns over the 'crowd' and concerns over their impact on data quality were additional reasons not to use the established platforms.

A different take on the choice based on familiarity is to see it as an *installed base*: As Star and Ruhleder note, "[i]nfrastructure does not grow de novo" (1996, p. 113). Re-using existing accounts, software parts and building on an existing workflow all build on pre-existing artefacts and knowledge. This installed base can bring benefits but can also imply limitations, such as the app developed by interviewee 2 being framed as not 'fancy'.

A rather different reasoning was done by interviewee 4, whose work is about creating a better crowdsourcing process. To demonstrate that it is better, interviewee 4 wants to compare it to prior research:

> "In this case, it was again because, in the paper from 2018, they also used [Amazon] Mechanical Turk. I know there's also Figure Eight, that's also very good, I believe, and .. I think, if I–if I had a choice, I'd rather use Figure Eight. But in this case, it has to be as close as possible so that the results are very credible and convincing that I want to describe in my paper. That's why I said, in this case, okay, we use the same as they did, they used [Amazon] Mechanical Turk"[39]

In this case, it's not the interviewee's own prior work (or that of their team) that had a significant impact on platform choice but the practice of other researchers in their field. To be part of that particular community of crowdsourcing researchers involves adhering to their *conventions of practice*, even if this comes with limitations (Star & Ruhleder, 1996).

Finally, interviewee 5, who is developing a full-fledged crowdsourcing platform themselves, attempted to use other platforms for their projects but failed to make them work:

> "Well, I believe, the interesting thing is actually that it's always been slightly different than these typical [Amazon] Mechanical Turk, or also Galaxy Zoo, or Zooniverse projects. Well we also tried to work with Zooniverse, the platform, but never really succeeded because it never really fit, based on the infrastructure there"[40]

As noted before, an installed base can pose limitations. In this case, the limitations were too significant, prompting the development of an entirely new platform. Creating new platforms (as by interviewees 2, 3, and 5) maximizes the possible control over the platform. But any choice poses limitations and, as the latest quote implies, has to be appropriated to fit the crowdsourcing application at hand. In the next section, I will take a closer look at how the interviewees have done this.

To sum up, which crowdsourcing platform gets used heavily depends on the installed base in the organisational context (experience with certain platforms and availability of re-usable tools) that implies limitations but also enables efficient use of resources. Compatibility with other infrastructure (payment infrastructure) and adherence to conventions of a community of practice can likewise be important. At the same time, it is interesting to note that in most interviews there was, contrary to the title of this subsection, no real choice between crowdsourcing platforms.

### 6.3.2. Appropriating the platform

As described in the previous subsection, many of my interviewees could build on pre-existing tools, either because they used commercial, existing crowdsourcing platforms (Amazon Mechanical Turk and Figure Eight), or because there were tools available in the organisation that were

the starting point – without need to start from scratch. Despite the ability to build on existing tools, these tools had to be appropriated to work for the particular crowdsourcing tasks. In contrast to that, interviewees 2 and 5 created platforms that were built anew. In these cases, existing tools were found to be limiting – the purpose-built tools allow for more control. In either case, the tools had to be tailored to work successfully, with issues ranging from access control to task distribution and user interface design. In this subsection, I will dive into how the various tools were appropriated to fit the purpose.

Interviewees 1 and 2 both had to make sure that the tasks get distributed correctly. In both cases, the correct amount of tasks needed to be distributed to each worker. In interview 1, this was because it was part of a teaching exercise, and each student should get an equal amount of tasks. Usually, crowdsourcing platforms provide support to ensure that each task gets done a specific number of times. In this case, however, each worker (i.e. student) needed to get a sufficient number of tasks, with the number of students per task not being important.

Interviewee 2 also bundled tasks into sets and distributed these sets to the NGO volunteers acting as crowd workers: "...and secondly, all comments were rated by three volunteers; thus I partitioned them into small packages, and each package was sent to three different annotators ..."[41] While the reason why this was done hasn't been discussed during the interview, one reason could be to distribute the tasks uniformly among the volunteers, as it was important to the NGO to have volunteers across the country involved (as discussed in section 6.2.4).

Interviewee 4 also distributed tasks to three workers. Since the goal was to show that three different workers agreed (hopefully) with medical experts, it is important to prevent one task from being annotated multiple times by the same person. In this and interviewee 2's case, the correct distribution of tasks may seem like a small detail, but it is crucial for the claims to be made based on the produced data set. In ANT terms, the correct distribution of tasks is essential for the crowdsourcing platform to act as a faithful intermediary and not as a mediator that complicates and subverts the relations (Latour, 2005).

Interviewees 3 and 5 did not have such strict requirements: In interviewee 3, it was simply important that each data point got annotated once. In the case of interviewee 5, all crowd workers had access to all the data. In both cases, access control is crucial, i.e. that only the right people have access to the tasks, but the distribution is not sophisticated.

The second crucial element of the platform that needs appropriation is the user interface, i.e. how the 'atomic tasks' are presented to the crowd workers, both visually and from an interaction point of view. Interviewee 3 could re-purpose a web-based user interface previously developed for an industry project for annotation and visualisation purposes. Similarly, interviewee 4 could

build on pre-existing user interfaces:

> "Interviewee: Well, on [Amazon] Mechanical Turk, there is this design interface, where you can put in HTML, you can put in JavaScript, you can basically build your own homepage there. The homepage, how you create it, will be displayed to the workers, and, I proceeded as follows, the task that I have, mark text, in, for example, mark certain parts of a text, that's called Named Entity Recognition
> Interviewer: mhm
> Interviewee: And then I searched Named Entity Recognition for Mechanical Turk, and then I already found such an HTML template. I took that and modified it how I needed it, i.e. that sentences got displayed, that my data gets imported, and I simply removed from the old– from the other interface whatever I don't necessarily need, and this is how I adjusted it."[42]

Both interviewees could build on an *installed base*: User interfaces with very similar requirements were either available within the organisation (interview 3) or available publicly due to the popularity of Amazon Mechanical Turk. In both cases, there was only minimal work necessary to adapt the existing user interfaces to their use cases. While Amazon Mechanical Turk does have limits regarding what can be achieved with the user interface, these were not significant for interviewee 4. Indeed, the platform was very allowing, which is why interviewee 4 had no problems designing the interface as planned, contrary to their own expectations:

> "Interviewer: and were there any restrictions when building the interface that have hindered you, was there anything that you would have changed if you had full control?
> Interviewee: Nope, actually no. That was actually positively surprising. I thought that one cannot simply put in JavaScript and everything, but it worked, without problems"[43]

Quite different to this is the case of interviewee 2. The interviewee had to design a user interface from scratch, as they chose to use the statistical software R as underlying infrastructure (see section 6.3.1 why this technology was chosen):

> "Erm, I'm not enthusiastic about the app, eh, well, if there's someone who can develop a better user interface and so forth, would–I would do it immediately, erm .. but still, it's better than obtaining comments manually, and it, it provides standardisation, that means there's always the same form and I, there are things written that are important for [the NGO], eh, yeah. But Shiny is certainly not the best to use for this." [44]

Building on top of R as an installed base meant inheriting its strengths (convenient workflow without manual intervention), but also the associated limitations, namely an interface that the interviewee deemed not optimal.

Whereas interviewee 2 created a makeshift solution, interviewee 5 ended up developing an entirely new platform that continues to be used across several projects. In this case, it was the limitations of existing platforms that were too restrictive:

> "Interviewer: Mhm. Well, because you already mentioned Amazon Mechanical Turk, that's, of course, somehow, when one is talking about crowdsourcing, the–the big name or so. Were there reasons why one didn't use such existing platforms?
> Interviewee: Yes, for us, it was the content, erm, we actually never really saw a way to bring it into Amazon Mechanical Turk." [45]

In this case, appropriation has found its limits. As I will discuss in detail in section 6.4.3, one of the key reasons was the importance of contextual information and the difficulty to provide this contextual information on existing platforms.

## 6.4. Addressing uncertainty

In section 6.1 I identified sources of uncertainty during crowdsourcing: The question how much the data to be crowdsourced is a matter of fact or a matter of concern, the translation of the problem into 'atomic tasks', the 'crowd', and how to aggregate the resulting data (if necessary) all can cause uncertainty. In this section, I will focus on how my interviewees addressed these sources of uncertainty. Some strategies came up repeatedly throughout the interviews, whereas some were only used by few interviewees, or even only once. The means of addressing uncertainty range from training the crowd, finding a good task design, providing context for the tasks, supervising the crowd, to conducting pilot runs of the crowdsourcing process. In the remainder of this section, I will take a closer look at these strategies.

### 6.4.1. Training the crowd

As I've described in section 6.1.3, the crowd can become a concern. As elaborated, this can be due to a lack of trust towards (potentially anonymous) crowd workers, but it can also be due to concerns over the crowd's competence in solving the task. Training the crowd is one approach to build their competence. But training can also help to address unclear tasks by ensuring that the workers know exactly what it is actually what they're supposed to do, how to read the 'atomic tasks' they should solve properly. Finally, it can also help in cases where the data remains a matter of concern, where there's no clear 'right' and 'wrong' way of annotating the data.

In the case of interviews 1 and 5, crowdsourcing was frequently used in teaching, and the task was closely related to the content of the university courses. Hence, it was not laypeople, but actually students who were already trained to do the task: "We use crowdsourcing technology, but for expert sourcing, yes, so our crowd is not really the laymen crowd, but they're students, there are sixty, seventy students". Consequently, there was no specific training necessary for the crowdsourcing itself, as the university course was providing that training. Similarly, the platform developed by interviewee 5 was initially used by librarians who didn't need training:

> "...thus that was really this expert sourcing task that was done then. That means it was people from libraries themselves, who were familiar with the material, who created transcriptions of the location names and maps, ..."[46]

Later, it was frequently used in a teaching context where small projects with students were conducted on the platform.

In both cases, the 'crowd workers' solved tasks topical to their studies or area of professional activity. Hence, they could be considered experts that don't need training explicitly for crowdsourcing. (See also section 6.2.1 for more on the role of the workers' expertise.)

In other cases, when the crowd workers were laypersons, some form of training was used. One approach was guidelines and reading instructions as training devices, as was the case with interview 3, where singing should be detected in recordings of operas:

> "Interviewee: then they got access, exactly, and in the interface, there were additional instructions how it ought to be used, the
> Interviewer: mhm
> Interviewee: interface, and in the interface there was also a guideline what we define as singing, what to pay attention to"[47]

Even though the detection of singing is not a task that requires extensive expertise, it is still important to ensure that the crowd workers *read* the task correctly. But, as already discussed in section 6.1.2, these guidelines have to be of appropriate length and complexity, lest they themselves become confusing and, as a consequence, problematic. This, of course, creates a tension between being concise on the one hand and being sufficiently precise and detailed on the other hand.

However, guidelines can not always clarify all cases that can come up in the data, as interviewee 3 mentions with regards to a different project they conducted. In that project, it was about detecting boundaries within a recording when it changes, e.g., from talk to music:

> "Interviewee: ... and [there] are then also deviations, either things are shifted, or some transitions are simply missing, and, well, there were very detailed guidelines, how this was supposed to get annotated
> Interviewer: mhm
> Interviewee: by those who organised this, but it still leaves room for interpretation, and sometimes one says, ah yeah, that is a change, and someone else says, no no, this [all] belongs to one part." [48]

This is a case where there is no rigid definition of what to consider a change in the recording, where what to deem 'correct' is still *disputed* and a *matter of concern*. Or, put differently, there's no unified enactment of a change in the recording.

The most extensive training was done by interviewee 2. NGO volunteers who were new to crowdsourcing were initially working with data points that were already annotated previously, i.e. that were part of a 'ground truth'. This training took three to four days and was also used to let the volunteer crowd workers familiarise themselves with the app, "so they learn dealing with the app, how does it work, what can I do, what can I not do"[49]. The training was basically learning by doing. The comments used were tricky ones identified during a piloting phase (more on this in section 6.4.7).

During this test phase, it was also observed how much the volunteers agreed with each other and the previously assigned annotations. It's important to note that disagreement was not seen as a form of deficit due to interviewee 2 openly acknowledging that the data points remained open for dispute.

To sum up, training seems to be deemed necessary if the task is not aligned with pre-existing expertise of the workers. The main goal of training is to ensure that the 'atomic tasks' work as immutable mobiles by making sure they get read appropriately by establishing a shared understanding between initiator and crowd workers. This highlights that, as Michael (2016) notes, the power of an immutable mobile depends on the readers' appropriate reading, "in ways desired by their producers" (p. 60).

### 6.4.2. Task design

As discussed in section 6.1.2, the task itself can be a considerable source of uncertainty. In this section, I will show how my interviewees used good task design to address uncertainty. Two intertwined processes became apparent in most interviews: Task formalisation and task decomposition. However, this did not apply to all interviews, with interviewee 5 deviating from this pattern.

Interviewee 1 walked me nicely through their iterative task design process that, broadly speaking, consisted of three stages. The first challenge was to find a suitable way to split up the task of verifying a conceptual model into smaller tasks, i.e. the task decomposition: "we wanted to see, how you can split [up] this task that is very monolithic". This was done with the help of software engineering experts who provided input on how this decomposition could work in a meaningful way: "[F]or them a good way to split it up was to focus on different elements, right? So concepts, or relations, or or what not ... So we got these insight ... and then we built a task ..."

This lead to a first version of the task design where the crowd workers had to check whether a shown model corresponded with a textual description. The answer then is a binary 'yes' or 'no' choice. As interviewee 1 later explains, it was also important to describe the defect. Initially, it was entirely up to the crowd workers how to (textually) describe the defect if there was one. But this very open form of answering the task caused issues when aggregating the results, as already described in section 6.1.4.

Hence, the task design was changed in the next iteration: "so we experimented with asking them to use a controlled natural language, yeah, so kind of a schema, ... So there were different defect types, typologies". But, as it turns out, this didn't work out either. This time it was due to the non-compliance of the workers, who simply didn't do as asked, who read the 'atomic task' not as intended, as already described in section 6.1.3. Thus, the 'atomic task' didn't work as a faithful intermediary but as a complicating mediator.

This led to the next iteration, which also was the final task design and user interface:

> "And then in a second version, we changed the interface to [a] more guided interface, where they should–they had to answer a set of questions, yeah, and based on what they answered it was like a decision tree, you ended up to certain defect types."

Through iterations, interviewee 1 decomposed a complex, open-ended question into a series of single-choice questions that made the results viable for aggregation and turned the 'atomic task' into a proper immutable mobile.

Interviewee 2 similarly used task decomposition. They created an interface that asked the crowd workers a sequence of questions to come up with the final classification of the social media postings, distinguishing between positive or neutral comments on the one hand and negative comments on the other. If the crowd workers chose 'negative', further questions had to be answered:

> "... then, ah, comes the choice, negative but not problematic, because I can, ah, express myself negatively without somehow insulting someone. Erm, then, negative

> and problematic, i.e., somehow insulting or discriminating, erm, – then hate speech, where we provided examples as . . . defined by the EU; thus we also made it according to this definition, erm, planned and trained the people who had to rate it. And then there was a fifth choice, and that is, erm, ambiguous, . . . " [50]

Like interviewee 1's approach, the final task design involves conditional branching of single choice questions, resulting in an unambiguous classification. Of note here is that this unambiguity of choices was achieved by explicitly acknowledging ambiguous content by providing a dedicated 'ambiguous' option (more on this in section 6.4.5).

Task decomposition was also important in interview 4, where three pieces of information got marked in abstracts of medical studies. Instead of doing this in one task, interviewee 4 split it up in two ways:

> "Nah, it's again; it ought to be as simple as possible. That means I ask one group of workers only–only what are–what are the participants of the study. Then I ask a different group of workers only what is the intervention. Thus, I basically create three subtasks, and they always mark up only the intervention, intervention, intervention, and the others mark up only the outcomes, which means there's no switching between; it's basically three–three separate tasks." [51]

The first split is to let only one piece of information get annotated at a time, and also to let one crowd worker only annotate one type of information so they can focus on this subtask. The second split is that of the abstract into individual sentences.

A consequence of this dual task decomposition is that this results in many small tasks to mark up the three pieces of information in one abstract, about 25–30 per abstract based on the numbers given by interviewee 4. This appears to be a worthwhile trade-off, as interviewee 4 identifies this strategy as one key improvement over prior research:

> "They did it differently, because they displayed the workers the full abstract and said, 'look, that's the abstract, mark up all, all participants, all that participated in this study', that means they didn't do it on sentences, that is, for example, one, that is in my opinion in any case an improvement, . . . "[52]

Splitting up longer texts into individual sentences has another effect. Because the crowd workers on Amazon Mechanical Task get paid per task, the tasks must be of similar difficulty or take a similar amount of time, according to interviewee 4:

> ". . . because the workers can look at each, at each task before they start with it. That means when a worker sees, okay, that abstract has 50 sentences, I won't bother, I

certainly won't do it, I'll immediately click reject. That means this is not really fair if you take the full abstract. Sentences are always about the same length; thus, this problem is, I think, more minimal." [53]

Making sure that all tasks require roughly the same effort to solve is perceived to contribute to fairness. If this was not the case, interviewee 4 feared that some tasks requiring more effort wouldn't get solved.

The platform developed by interviewee 5 deviates substantially from the approaches chosen by all other interviewees. In this case, it's also about annotating images and text. Once these documents get uploaded to the platform, users (crowd workers) can "enrich them semantically, that means they can mark people, they can mark locations, text in images and they can, ah, simply create comments on it."[54] While the other interviewees tried to reduce the crowd workers' leeway of solving the task, the platform developed by interviewee 5 puts minimal restrictions on the crowd workers:

"Interviewee: ... thus it was always teamwork, the people have uploaded individual packages and then distributed the work and then started to transcribe, erm, and that quite simply as a team, or in pairs, i.e., there was no formal task process, but the people have simply used the tool for transcription
Interviewer: mhm
Interviewee: controlled each other, erm, data evaluation, like in the sense of a map visualisation of the points that were resolved and so forth. But it was no formal, how it's [done] on Amazon Mechanical Turk or so, that it's split up into simple tasks, that was not happening. Thus it was strongly the work of historians, simply tool-supported." [55]

There was no clear assignment of (small) tasks to workers, and there were no pre-defined categories, no single-choice interfaces for the annotations, etc. Instead, the whole process is centred around creating annotations and commenting on these annotations with some additional support for coordination:

"So, if one says one is doing a transcription but has some doubts, then one can write a comment, and there are discussion threads below, that was the one thing, erm, the second was the resolution as contemporary geocoordinates, there it was also that one could not always identify it, and there we introduced a flagging system where one says, okay, I cannot resolve it, ah, and then this popped up yellow, that means one did always immediately see where the problems lie, and then there were, at least in

> the first platform, some reasons why it was flagged yellow, where one could filter, if one says I think this is a location, but I cannot find it; I think it's a location that doesn't exist at all because maybe he made an error, or because it was mythical locations on these maps, and there was a small taxonomy of possibilities why this task failed" [56]

Even though there was a supporting structure in the form of discussion threads and the flagging system in place, there was no formal process for using these. It was entirely up to the workers' convention how to use the capabilities of the platform. I also want to highlight that in this case, uncertainty has been openly acknowledged as a possibility: Flagging an annotation indicates that the crowd worker is uncertain how to evaluate an annotation, including different types of uncertainty.

Additionally, on this platform, cooperation and dialogue is a central concept: It's not a competition among workers for tasks (and consequently money in the case of Amazon Mechanical Turk), nor are workers' annotations compared to each other for aggregation or with a 'ground truth'. Instead, each annotation is a collaborative effort and ambiguities, anomalies and disagreements are resolved through direct communication and cooperation among workers. Eventually, it may be that a senior expert makes a final decision, as discussed in section 6.2.1, but the guiding principle is one of cooperation.

To a large part, interviewee 5's work is the antithesis to the work of the other cases, where the original problem gets decomposed into small tasks that get distributed to crowd workers who then work in isolation on individual tasks. This case highlights that task decomposition is dependent on data and a task that allows this process. This also has implications on how context is introduced (see next section) and how the crowd's work is eventually used to create a final data set (see section 6.4.6).

### 6.4.3. Establishing context

A recurring issue among the interviews was the importance of context. This context is important for each data point to be understood correctly. In some interviews, context was mentioned explicitly; in others, it was present implicitly.

The importance of context was most explicitly addressed in the case of interview 5. It was the failure to adequately provide the necessary context that led to the development of the platform. While the platform that got developed differs considerably from typical crowdsourcing platforms, these platforms were the starting point:

> "Well, one has to imagine there's a huge map, with about eight, nine hundred location

> names on it, but these are for laypersons really hard to read. Me personally, I have seen have many of them, but as a computer scientist, I can hardly read them; thus, one really needs experience working with it, and the people then simply transcribed it and need the surrounding context. I.e. that one says to cut it into small snippets and tell [someone] 'transcribe this for me' that doesn't really work in this context. Based on the content as well as based on the required expertise, I believe." [57]

Hence, this particular application scenario resisted the strategies of task decomposition. And it was not only Amazon Mechanical Turk that was insufficient, but also Galaxy Zoo and Zooniverse, two platforms that are more targeted at scientific crowdsourcing. These platforms often get used to crowdsource, e.g., deep sky images. Crowd workers then classified if small parts of these images showed galaxies and which kind of galaxy. In these cases, volunteers (often laypeople) who were interested in astronomy participated.

Interviewee 5 also tried to work with these platforms, but even working with people running Zooniverse, who are experienced with crowdsourcing, didn't lead to a viable solution "[b]ecause Zooniverse also strongly depends on the basis to fragment everything into micro-tasks, there are usually small images that are either labelled or transcribed, and that simply didn't work for us."[58] The principle of splitting up a task into small, 'atomic' tasks is not, it seems, a universal solution.

Interviewee 4 also had to address the issue of context. After all, interviewee 4 split up the medical studies' titles and abstracts into individual sentences. This also meant a loss of context that had to be addressed:

> "... I post each of these individual sentences to Mechanical Turk and for the workers to also have a certain context I basically show them the full–the full abstract and I highlight, 'look the sentence that you have to annotate, it appears here in the abstract.' That is important, for example, if a sentence contains the abbreviation 'AD' for Alzheimer Disease, but most of the time, there are explanations of what the abbreviations mean. And that's why it can be important to show the full abstract so that the worker has the full context where the sentence occurs if he wants to." [59]

The situation, it seems, is quite similar to the one explained by interviewee 5: The sentence without the full abstract can not always be understood – similar to one individual location name without the surrounding map. It appears that it is the media and the amount of context necessary that makes the difference: A huge map cannot be as easily displayed as context as an abstract of about eight to nine sentences.

As described in section 6.3.2, interviewee 2 also provided context information available in the task interface that was important to the NGO to be displayed alongside the 'atomic task' itself – as a form of context. However, there's a distinction to be made here: Whereas in the case of interviewee 2, the context information appears to be static, to be the same for each 'atomic task', in the two cases discussed above it is other data points that are relevant context: One name of the map can often only be transcribed successfully if other, neighbouring names and their relation to the name under scrutiny are also visible. Similarly, the sentences have to be evaluated in relation to other sentences in a particular abstract – not only more general context shared across all tasks. Put differently, taken out of the context of other data points, they lose their meaning.

In line with training, guidelines and task design, providing this kind of context is another means to ensure that the crowd workers *read* the data and the task correctly. How much context is necessary to achieve this depends on the specific project, the kind of data, and the task to be solved.

### 6.4.4. Supervising the crowd

In section 6.1.3 I have shown how the crowd can become a source of uncertainty. Training the crowd (section 6.4.1) can reduce this uncertainty if the concern is the crowd workers' competence to solve the task. A different approach is to supervise the crowd and sanction them or exclude them from the crowdsourcing process if they don't perform as expected.

This is particularly tempting if the crowd are anonymous people where trust is minimal and has to be established. For interviewee 3, addressing an anonymous crowd via Amazon Mechanical Turk was off-putting because they didn't know if the quality of the resulting annotations would be sufficient, contributing to the decision against using this platform. They instead chose to work with students to whom personal contact existed.

Only interviewee 4 actually crowdsourced to an anonymous crowd, using Amazon Mechanical Turk. During the interview, the respondent outlined a few ways to supervise the crowd. One is to monitor and validate the work they've done: "In the worst case one can always click 'does not contain', submit, 'does not contain', submit, but I can, I see this afterwards, of course, I have two days, and can say, this worker gets no money, for example."[60]

A different form of monitoring is to look at a random sample and check if the answers are meaningful:

> "Interviewee: Or, if it's super obvious with one sentence, I say it's medical data, it's not so easy, but there are sometimes examples where it's very obvious what is

> correct
>
> Interviewer: mhm
>
> Interviewee: and if something like this occurs increasingly often, that he always says, no, wrong, wrong, and then–then I see okay, he's not taking it seriously, and then I'd, e.g., reject him and he would get no money, the task will be put online again automatically, after I have rejected it so that I hopefully get better results from a different worker" [61]

This behaviour of not properly doing the task was also called 'scamming' by interviewee 4. By monitoring the crowd worker's annotations, the initiator of the crowdsourcing can identify annotations that are too homogeneous, too monotonous, suggesting 'scamming' behaviour. Together with the possibility of withholding payment, the crowd workers can be disciplined to behave orderly.

A different, more efficient approach is to select 'good' workers based on an initial test run, as explained by interviewee 4:

> "What's good practise is to conduct a small, a small run, and take only those workers who are very good, that means you look very closely for this small run, let's say ten sentences, and let each sentence get annotated by 20 workers. I then thoroughly look through the sentences for these 20 workers and look 'this worker is good, that worker is good' and then I may find, e.g., that 18 work very well and two don't, then I can define that I conduct the task–I put more sentences online, but it's only the 18 who are allowed to participate. That means the two are no longer allowed to contribute and no one else, either." [62]

Here, the incentive to work on more tasks (and earn more money) is a slightly different form of disciplining. This is also where a so-called 'ground truth' comes in handy (discussed in section 6.1.1). This process can be easily automated if data from this 'gold standard' is used to compare if the crowd workers provide the expected results: "I can evaluate automatically, and I can see automatically, this worker has an 80% accuracy, and then I could kick out everyone who has less than 50%"[63]. A similar system is what interviewee 3 had on their mind if they had used Amazon Mechanical Turk (which they eventually didn't do).

To sum up, there are various means to supervise the crowd if it is not (fully) trusted. But they have limits: Either they are laborious, as they involve manual work, or they require at least a minimal data set where the 'correct' answer is already known, i.e. a 'ground truth'.

But there's also another issue at stake here. Drawing on Mol's ontological politics (1999), by comparing the crowd workers' results with the known 'correct' results and based on that enacting

them as 'correct' or 'incorrect', the decision if the crowd workers' results are 'correct' has been shifted to the creation of this 'ground truth'. From this point and at this stage, this 'ground truth' seems like a 'fact', and the selection of good workers appears like a mere technicality. In addition to that, by rejecting work that does not align with the 'ground truth', this enactment as 'incorrect' *interferes* (Mol, 1999) with a loss of earnings, even though work was done.

### 6.4.5. Involving experts

The previous section was about the means to supervise the workers with the underlying assumption that the workers are either not trustworthy or incompetent and have to be monitored and disciplined. A different approach involves some sort of expert who is supposed to have better knowledge of the topic. In this case, the data and the task are seen as the culprits – whereas the workers are trusted and assumed to do as good a work as they can. In my material, this kind of expert judgement came up twice. In interview 2, this was clearly regulated, whereas in interview 5 this was an informal process.

The project where interviewee 2 was part of had an expert board installed that should handle tricky cases, i.e. social media posts that were either delicate or difficult for the volunteers. There were clear rules when this board was supposed to get involved: Each comment was rated by three volunteers. Throughout most of the project, one rule was that if the three volunteers did not fully agree, the board had to make a final decision. But during the final weeks of the project, there was a change. As already discussed in section 6.2.1 in this final phase, there were only two cases when the board got involved: If the volunteers clicked on 'hate speech' and if they clicked on 'ambiguous'. It was then the task of the expert board to confirm if it's hate speech or to decide how to rate the comment. In all other cases, majority voting was used.

The situation is slightly different in interview 5, where students and their supervisor frequently cooperate on a set of documents, as described in section 6.2.1. In line with what I've written in section 6.4.2, and contrary to interview 2, there are no clear rules when the supervisor gets involved; instead, this is done informally and flexibly.

While I did not cover this during my interview, I suppose that the board involved in interview 2 did come to a conclusion in a cooperative way, similar to the discussion structure of the platform developed by interviewee 5. In both cases, it's openly acknowledged that the 'correct' result can be disputed, that what is 'correct' is a *matter of concern* (Latour, 2005). This is in stark contrast to the mechanisms presented in the previous subsection. This stance is also different in that it problematises the data and the task to solve as difficult and uncertain, and not the 'crowd', indicating an underlying assumption to trust the crowd workers.

### 6.4.6. Leveraging the power of the many: Aggregation

In section 6.2.1 I have already noted that often the crowd as many, together with a clever crowdsourcing process, should rival the work of an individual expert. To achieve this, one individual task is solved by several crowd workers. This, in turn, needs an aggregation method for combining the individual workers' solutions into one solution.

I have already discussed in section 6.1.4 that aggregation is highly dependent on the task design to work – something that was prevalent in my interviews, with interview 5 being an exception. But if the task is designed in a way to make quantitative aggregation possible, there are different ways how the aggregation can be done. Majority voting is one of the easiest ways to do this and a popular choice: "[T]his was just an initial [approach] because ah it's the simplest" (interview 1). It was also used by interviewee 2 during stretches of their project (see section 6.4.5 for more details). If majority voting gets used, three workers per task is a handy choice: "Why three? Aaah – well, because with three one can think about a tie breaker, that means ... we could switch to the majority wins"[64] (interview 2). Letting three workers annotate a single data point is a minimum to calculate a 'majority' mathematically.

Interviewee 4 didn't attempt to establish a final data set but was more interested in the process itself, as already discussed in section 6.1.4. Despite this, the choice here was also to let three people solve one 'atomic task', and I can imagine that in the future, majority voting with three crowd workers would be used to create data sets with the crowdsourcing process since the interviewee also mentioned this: "normally one does this to be able to do majority voting"[65].

During interview 1, other approaches to aggregation were mentioned. They tend not to work based on the agreement among a fixed number of crowd workers but adopt more sophisticated approaches. These approaches are discussed in the scientific discourse as "truth inference" (interview 1). If one adopts one of these approaches, one may collect more than three results if the first three showed disagreement, or one compares workers across several tasks and weighs each worker's result differently, based on some measurement that is supposed to indicate the "quality of the workers" (interview 1). One such approach was mentioned specifically, the so-called Crowd Truth method (Aroyo & Welty, 2013). More sophisticated methods use even more complicated calculations, involving "probabilistic graphic models ... using statistics" (interview 1). While it was interesting to learn that there's active research on more sophisticated methods for aggregation, it's even more interesting how widespread the use of majority voting seems to be among practitioners.

Sophisticated or not, it is not always necessary to do this aggregation explicitly. As interviewee 3 says, with tongue in cheek, one can also train machine learning algorithms with 'alternate facts'

(see section 6.4.2 for more). It's then up to the machine learning algorithm to make the most use of the training data. This can mean that, if the whole data set is large enough, there's no need for multiple annotators per data point, or one can use diverging annotations for the same data point directly for training and hope that this helps the machine learning:

> "And through that it happens that, if they agree the [neural] net gets two times the same [signal], and if they disagree it gets two different things and will become more uncertain, or learns from both a little bit" [66] (interview 3)

The hope was justified because "it worked well" (interview 3), as the trained neural network performed better when confronted with contradicting results than when they were omitted. As interviewee 3 noted, it is probably that the information over the crowd workers' disagreement is itself valuable – something that is not incorporated in majority voting (Crowd Truth (Aroyo & Welty, 2013) attempts to use this information, but as mentioned, is not used by any of my interviewees).

Conducting aggregation explicitly allows calculating new measures. Giving an 'atomic task' to several crowd workers enables the calculation of their agreement, which, in turn, can be a means to quantify the certainty of an annotation: It can be argued that the more people agree, the more certain their annotation is the 'correct' one. Based on these new inscriptions (quantified agreement), new claims can be made:

> "And if I ask three people, and all three agree with the medical experts, then I can say, look, my tool is really very good, because all three agree on the sentence, three different Mechanical Workers, and all three say exactly the same thing as the medical expert."[67] (interview 4)

Being able to show high agreement statistics allows for stronger, more credible statements to be made, in this case about the quality of the crowdsourcing approach, in other cases, this may support any claims made from analysing the annotated data, or even claims about the quality of the annotated data set itself. Such claims cannot be made directly in the case of interviewee 3, where the machine learning algorithm implicitly combines the results.

In the case of interviewee 2, relating agreement to certainty is also highly visible. In that case, social media postings got forwarded to a board if the crowd workers disagreed. This was not done if all crowd workers agreed, implying sufficient certainty that the common annotation is appropriate.

To sum up, the aggregation has two goals: First, to integrate several crowd workers' annotations, and second, to show how much they agree, which in turn allows new claims and addresses

uncertainty. But, as I have pointed out in section 6.1.4, the 'atomic task' has to be designed appropriately so that aggregation can work. In section 7.4 I will take up this issue and discuss its potential wider consequences.

### 6.4.7. Stabilizing crowdsourcing: The piloting phase

In the previous sections, I've been describing means to address uncertainty: Properly training the workers, creating a good task design, providing sufficient context information, supervising the crowd, and involving experts. But how can one find out what a good task design is, what amount of context is necessary, how to best train and support the workers? This is where the 'piloting phase' enters the picture.

'Piloting phases' were mentioned by several interviewees. In the case of interview 1, it involved several iterations of the task design to make sure that the task was well understood by the crowd workers and that the results could be aggregated. In section 6.4.2 I already described how the task design evolved through these iterations. Interviewee 2 conducted a very similar pilot phase:

> "Well, actually, that might be interesting; at first, they wanted to differentiate, erm, positive, neutral, negative, and then differentiate the negatives even further. Then – eh, but we did a pilot, and a few times simply tested in small groups, and we found that the differentiation between positive and neutral is even harder than somehow between neutral and negative; thus we simply merged them both." [68]

The piloting phase is a combination of adjusting the task design and understanding how crowd workers can cope with the task, and both get improved iteratively during the piloting phase. The data that got annotated during this phase was used to evaluate the prototypical task designs and got re-used for the training of the volunteers, as described in 6.4.1.

Interviewee 4 also planned to run small tests before actually conducting the crowdsourcing on Amazon Mechanical Turk with anonymous crowd workers:

> "what I'll also do is to discuss with colleagues how they assess the interface, if it makes sense to show the abstract, or not, and, yes, – well it's not yet–it's still written in the stars if it's even a good or a bad idea" [69]

The piloting phase, in this case, is done with colleagues, which allows a closer interaction and faster iteration.

While the three pilot phases had slightly different emphases, they all calibrate the individual parts of the crowdsourcing process, to stabilise the crowdsourcing process. From the earliest iterations, as described by interviewees 1 and 2, it's clear that these initial crowdsourcing processes failed or were at least very likely to fail.

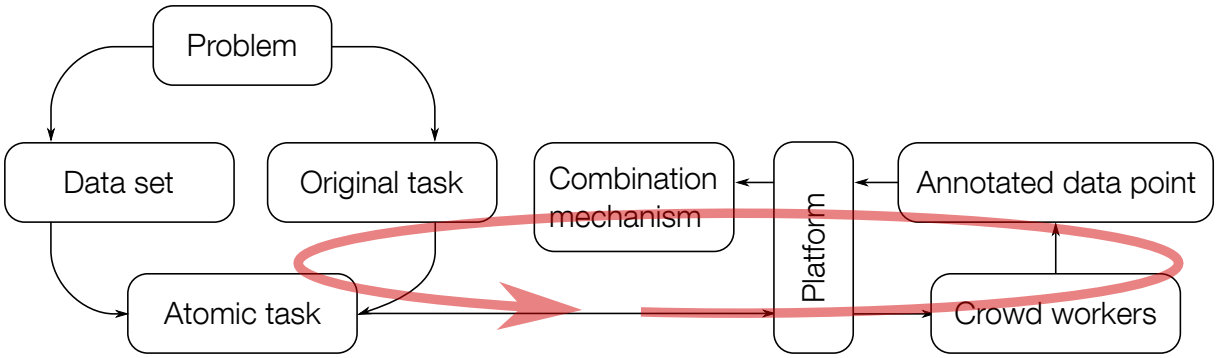


Figure 4: The piloting phase involves iteratively adjusting and aligning the 'atomic task', how the annotations should be performed by the crowd workers and the combination mechanism.

Looking at this process from an ANT perspective, it becomes clear that this piloting phase is when the network builder actively constructs the actor-network through many translations. In most cases, there are two key relations that seem particularly important yet also tricky.

First, as I noted several times throughout this thesis, the 'atomic tasks' have to be made into faithful intermediaries, into proper immutable mobiles. If this fails, crowd workers may annotate data points, but the results may be worthless as they solved the task differently than intended. That's why my interviewees went to great length to ensure that the crowd workers correctly read the 'atomic tasks': Crowd workers were provided concise instructions and guidelines on how to solve the task, they underwent a short training phase, got provided relevant context information, and were guided by a refined user interface, all to ensure the correct reading of the 'atomic task'. These elements were configured and re-configured in an iterative fashion.

Second, for all but interview 5, it was important to successfully enrol a final mechanism that then had to process the crowd workers' annotations. In most cases, this was an aggregation mechanism that produced one final, combined annotation per data point, or, as in the case of interviewee 3, it was a machine learning algorithm that could directly process the annotated data. As I have discussed in section 6.1.4, this does usually not work on the first try but requires trials of strength. In these trials of strength, the task design gets adapted so that it can establish a stable relationship with the aggregation mechanism.

Revisiting the diagram schematically depicting the crowdsourcing process, it becomes clear that the piloting phase involves a large part of crowdsourcing (figure 4). It is noteworthy that agency during the piloting phase is temporally shifting between different actors(Pickering, 1995): At first, my interviewees created an 'atomic task' that seemed reasonable. Then this got distributed to crowd workers (even if it was only colleagues) who now were the active agent.

Eventually, the combination mechanism combined the resulting annotations (e.g. aggregation through majority voting or a machine learning algorithm). At this point, the agency shifts back to my interviewees, who re-adjust the 'atomic tasks' and a new iteration starts. At the same time, my interviews confirmed what Pickering observed: Often, the assumptions and intentions of my interviewees got readjusted: The failing controlled language (because the students didn't follow the instructions properly) readjusted interviewee 1's expectations how to set up a working 'atomic task'. Similarly, the piloting phase showed that it was hard to differentiate positive from neutral social media posting reliably – readjusting the NGO's and interviewee 2's intentions to only differentiate negative postings from non-negative ones.

# 7. Discussion

Until now, I have confronted you, dear reader, with intricate details from my empirical results. What, then, to make of this more analytically, thinking the issues that arose more broadly? In this section, I want to sort my findings and tie up some lines of thought that I opened up before.

## 7.1. The importance of context and its relation to biases

Before I expand on the lines of thought I opened in the previous section, I want to return to my empirical material, my interviews. I am fortunate that I could interview interviewee 5 that develops a crowdsourcing platform to annotate, e.g., historical maps. This case is different to the other four interviews. I believe one could debate whether this even is a case of crowdsourcing. Not only because it's experts that participate as crowd workers (which is the case in another case, too), but also because it deviates considerably from the structure I've outlined in section 6 and figure 3: There are no 'atomic tasks', there's no clear method to combine the annotations from multiple crowd workers, two core elements present in the other cases. The most important contribution of this interview is that it brought the role of context into the spotlight: In this case, the necessary context of other data was the crucial element that made other crowdsourcing approaches fail.

After I have conducted this fifth interview, it suddenly became clear that the establishment of context in various ways was central to the work of all interviewees. I have already discussed in-depth how context got introduced into the 'atomic tasks' (section 6.4.3). These 'atomic tasks' have to work as *immutable mobiles* and retain their meaning when travelling from the practitioner to the crowd workers. In addition to the context information provided with the 'atomic task', I consider the training of workers in its various forms, as discussed in section 6.4.1, as another form of establishing context beyond individual 'atomic tasks'.

Given the importance of context, one must wonder what the consequences are for the resulting data set. First, my findings confirm that data sets are always contextual (boyd & Crawford, 2012) and, that they are produced from a particular vantage point. This stance was also shared by one interviewee who noted that the data gets annotated within a particular historical setting – and a machine learning algorithm trained on that data may fail to work properly at a different time. Second, I'd like to relate my findings to those of Jaton (2021b): Jaton argues that 'ground truths' (and algorithms created subsequently upon them) are the result of a particular problematisation process, and he uses the term 'biasing' for this process. In his analysis "biases are not the consequences of algorithms but, perhaps, are one of the things that make them come into existence" (p. 307). This is a definition of bias that is quite different to the kind of bias

problematised and criticised in much of the fairness research (see section 2.2.2 for more on this). Understanding biasing in Jaton's way, I'd argue that the context that gets established also plays a crucial role in biasing the resulting data set – and given the effort put into establishing context, it's clear that these biases are on the one hand necessary to obtain a coherent data set (Jasanoff, 2017), and on the other hand, they are actively produced.

Of course, the issue of necessarily biased data sets seems obvious when seen through the lens of ANT's multiplicity: I relate the notion of 'unbiased' datasets to the notion of a singular reality (Latour, 2005; Mol, 1999) and attempts to fully de-bias data sets as a striving to uncover this singular and essential reality. Instead, I argue, each data set enacts a particular reality, just the way that we never see 'the car', but always a particular car (Michael, 2016). This aligns well with Selbst et al.'s (2019) notion of 'abstraction trap', but while they mentioned the social context where machine learning systems will be used could not be abstracted away, I argue that it's impossible to consider data sets as abstracted from the context embedded in them through the creation process.

Consequently, I concur with Jaton (2021a) that biases should not be automatically perceived as negative. However, and again in line with Jaton (2021b), it is important to attend to the possible consequences of this kind of biasing: If the result systematically treats or evaluates people differently along the lines of ethnicity, gender, dis/ability, etc., it has to be addressed. How context gets established in the crowdsourcing process can be a useful starting point to investigate how this kind of harmful biased consequence could have been embedded into the data set.

This is because my research also underlines Jaton's (Jaton, 2021a) assertion that in the process of generating 'ground truths' (and, I'd argue, data sets more broadly) is hardly a clean, linear path, but one riddled with hesitations, exploration, and, I'd add, experimentation (see, e.g., sections 6.4.2 and 6.4.7). The data set resulting from crowdsourcing could be otherwise if the problem would be formalised differently, if the 'atomic tasks' were defined differently, if the aggregation would be done differently, if the context would be provided differently, etc. As such, not only the biasing of the 'ground truth' could be different, but its consequential biases could be different, too.

## 7.2. Piloting, iterations, and shifting agency

However, where Jaton (2021a) sees, drawing on the pragmatist philosopher William James, "genuine options" (p. 8) among which to decide, I'd keep with Mol (1999) that it's often not a choice or decision which reality to enact. To illustrate this point, let me return to the

platform that the interviewees used for crowdsourcing. As discussed in section 6.3.1, none of my interviewees chose between two (or more) options: Instead, it was a familiar platform that got used again, a platform that had been used by competing researchers and had to be chosen again to make my interviewee's results as comparable as possible, new platforms were created with technologies familiar to my interviewee, or because there were no existing platforms that achieved what was needed. In these cases, I got the impression that, if the project was to continue, it was not a matter of choice but perceived as a matter of necessity to use these platforms.

Where I do see 'moments of hesitation' and investigations of the "fragilities and uncertainties of a genuine option" (Jaton, 2021a, p. 7) is the case of the task design. In most of the studied cases, initial task designs were created by my interviewees that they considered meaningful. However, interviews 1 and 2 showed that the final task design changed in more or less dramatic ways. Relating to Mol's (1999) question where choice is located and connecting to Jaton's conceptualisation as collective exploration, I'd locate this choice as distributed, as temporally shifting among several actors (Pickering, 1995). As discussed in section 6.4.7, these final designs of the 'atomic tasks' are the result of resistance of the data (being, e.g., difficult to interpret), the crowd workers (not following the task instructions), and the combination mechanisms (e.g., when it's not possible to aggregate the results) leading to the final design of the 'atomic task'. And, as Pickering notes, this process adjusts not only the 'atomic task' but also the practitioners' intentions.

While this was not covered during my interviews, a similar temporal shifting of agency is likely present at other parts of the crowdsourcing process schematically illustrated in figure 4. I did not cover the collection of the data set to be crowdsourced. Still, I think it's likely that the problem formulation is likewise re-adjusted in light of the data set that gets compiled – probably in a similarly iterative process that I described in section 6.4.7. Similarly, I do wonder how the expert board overseeing 'tricky' cases described by interviewee 2 (cf. sections 6.2.1 & 6.4.5) was shaped in their practice by the data they were confronted with and the (preliminary) categorisation conducted by the NGO volunteers, i.e. how these laypeople re-adjusted the experts' judgements.

## 7.3. Making tasks amenable to calculation

A different topic that became clearly visible is the tight entanglement of the combination method and the 'atomic task' design. In these cases, it was important to break down the task into small 'micro-tasks', as Berry (2019) calls this kind of task. He notes that these micro-tasks are a form of "discretization of human activity" (p. 50). As interview 5 showed, there are limits to this

discretisation if the context is too encompassing to be maintained when discretising the tasks, as I've noted in 6.4.3.

Several of my interviewees were not content in the discretisation of the work, in splitting up a large task into several small tasks, but making the task amenable to calculation was a key consideration for three of my interviewees (1, 2, & 4). For this to work, discretisation of work is a prerequisite and at the same time intensifies this discretisation: It was achieved by moving away from open-ended questions to a series of discrete options to choose among (interviews 1 and 2) or multiple sub-tasks that can be easier combined later (interview 4).

At this point, let me detour to my main research question, how do get uncertainties addressed in crowdsourcing processes? One key strategy is to give the same task to multiple crowd workers. In my interviews, there were largely two reasons why this was done. First, because the 'crowd' was seen in some way as deficient compared to experts, as more error-prone. Second, the data may be acknowledged as a matter of concern, as disputable, and by letting several people work on one data point, it's easy to identify those that are undisputed between the workers. In both cases, the uncertainty gets spread across several people, a common mechanism to cope with uncertainty (Nowotny et al., 2008).

But the interviewees using this mechanism didn't stop there. Instead, discretisation gets a prominent role again. In combination with giving the task to multiple crowd workers, making the task suitable for calculation allows to quantify the uncertainty: It allows to calculate the agreement among workers. However, potential disagreement can be attributed differently: Interviewee 4 wants to show high agreement values as a testament to a good task design and works on a problem that should be solved by the 'crowd' instead of 'experts'. In both cases, disagreement problematises not only the task design but the 'crowd workers', too – especially those deviating from the majority (more on this later). Interviewee 2 used agreement measures in two ways: First, to identify disputed data points that were considered as tricky and forwarded to an expert board and, second, to automatically decide on the 'correct' categorisation if under time pressure. However, in this case, disagreement was not identified with suspicion towards the crowd but as an indication of tricky data.

As discussed in section 6.4.6, being able to show these agreement numbers (that hopefully show high agreement) can also be a way to make new, strong claims. Being able to point to numbers makes these claims stronger, partly because "[t]he confounding of calculation with rational thinking implies that whatever cannot be reduced to number[s] is illusion or metaphysics." (Berry, 2019, p. 47)

## 7.4. Silencing the deviant

The practice of aggregation opens up another issue, that of silencing. All interviewees that used aggregation mechanisms used majority voting or hinted at its future usage. Other mentioned approaches also roughly follow the assumption that the majority is 'correct'. In this case, it is obvious that these practices silence the minorities, the deviant, those that disagree, even if this silencing is 'justified' by calculations of agreement, highlighting how 'strong' the majority is. The aggregated value acts, thus, as a spokesperson for the majority (Callon, 1986). This also begs the question, among whom there's a dis/agreement. Here, it was only interviewee 2 who explicitly showed sensitivity for the representativeness of the crowd workers, as discussed in section 6.2.4.

What aggregation makes possible is to keep uncertainties contained (Nowotny et al., 2008), to keep them from spreading further. If a certain level of agreement is reached, if the different 'atomic tasks' can be successfully aggregated, then closure can be reached, and from here on, this decision can be put beyond doubt (cf. Amoore (2019)).

This silencing can be particularly problematic if crowdsourcing is used as a form of participatory process (e.g. as part of citizen science): In these cases, not only get those who disagree with the majority marginalised but the silencing is justified by referencing the majority. This implicitly claims the decision to be a result of a democratic process, even though there was no room for deliberation, for debate.

Framing majority voting as democratic also interferes (Mol, 1999) with the question of population: Who is the population that participated in this 'democratic' process? How many have been included in the 'voting' process? Who got included, and who got excluded?

## 7.5. Data: A matter of fact or a matter of concern?

The usage and meaning of an existing 'ground truth' differed significantly. Interviews 1 and 4 asserted that these annotations were the 'correct' annotations, that crowd workers should agree with this 'ground truth'. Contrary to this, interviewees 2 and 3 were less definite about the correctness: They both used it as a form of quality control of the whole crowdsourcing process, without equalling deviations from the 'ground truth' as 'incorrect'. Instead, interviewee 1 openly welcomed variance and disagreement, acknowledging the disputedness of the data. Interviewee 3 did compare the result of the crowd workers with parts of the 'ground truth' and, if there was a disagreement they couldn't explain, got into a dialogue with the crowd worker. In both cases, the interviewees treated their data as matters of concern, whereas those that designate certain answers as 'correct' or 'incorrect' treat their data as matters of fact. Or, put differently, it's also

a question whether the data get enacted as disputed or as clear-cut cases beyond doubt.

The question of how the data gets enacted, whether it's enacted as doubtful or beyond doubt, becomes highly political in certain crowdsourcing situations. The 'ground truth' with 'correct' answers can also get used to select which crowd workers can participate in the crowdsourcing process, and it can be used to supervise them (cf. 6.4.4). This can be done by checking how strongly a particular crowd worker agrees with the 'ground truth' when solving test tasks based on it, but also with the majority of crowd workers. As noted by Irani & Silberman (Irani & Silberman, 2013), based on this dis/agreement, initiators of the crowdsourcing can decide not to remunerate these workers – at least on Amazon Mechanical Turk.

What this shows are two things. First, there are clearly unequal power relations between different humans involved in crowdsourcing (D'Ignazio & Klein, 2020; Miceli et al., 2021), with commercial crowdsourcing platforms contributing to this inequality. Second, the question of how the data is enacted (as doubtful/beyond doubt) interferes (Mol, 1999) with the enactment of a 'good' and trustworthy worker worthy of payment.

## 7.6. In/visibility of human work

Due to the focus of my thesis, I didn't interview crowd workers. Hence I can contribute little to the debates about crowd workers' working conditions. However, I do want to address the topic of crowdsourcing as an "unending stream of labour-power on demand in a similar fashion to an electricity or water supply" (Berry, 2019, p. 49). As my study shows, one cannot tap crowd workers the same way one would turn on the tap or plug in a charger. While crowd workers may be interchangeable in the sense that there are always other crowd workers at the ready, even this thought is questionable, as concerns over 'scammers' were voiced by those considering established commercial crowdsourcing platforms (section 6.1.3). But more to the point, the work expected from crowd workers is not as interchangeable as tap water or electricity. All my interviewees had to go to great lengths to make the crowdsourcing process stable.

Once the crowdsourcing process has been set up, it may well be that the sheer number of crowd workers available on commercial platforms allows operations to scale on demand (cf. section 6.1.1). But I do wonder whether this likening of crowdsourcing to electricity follows the narrative brought forward by "algorithm-related organizations, especially the most powerful ones, [who] are frankly reluctant to make hesitations and uncertainties visible and thus favour the modernist path of inevitable mastery" (Jaton, 2021a, pp. 7–8). It should thus be questioned whether this characterisation is not falling for marketing speak of entrepreneurs like Amazon's Jeff Bezos and technologists trying to praise their latest developments.

My study investigates the construction site of crowdsourcing. This 'making of' provides a view on crowdsourcing "that is sufficiently different from the official one. Not only does it lead you backstage and introduce you to the skills and knacks of practitioners, it also provides a rare glimpse of what it is for a thing to emerge out of inexistence" (Latour, 2005, p. 89). Particularly the variety of cases covered by my interviews illustrate that "things *could be different*" (p. 89, emphasis in the original).

Given the effort put into producing data sets, attempts to document data sets better than is currently the case (Gebru et al., 2020; Miceli et al., 2021) can not only lead to a better accounting of biases embedded in the datasets. They can also make more visible the work involved, not only of crowd workers (Irani & Silberman, 2013; Berry, 2019; Newlands, 2021), but also on the side of researchers, developers and other practitioners who initiate crowdsourcing processes. This, in turn, would also heighten the sensibility how much human labour is necessary to make 'technical solutions' work.

There's another hope frequently attached to crowdsourcing: That the 'crowd' could equal or even surpass the work of an expert (Brabham, 2013). This idea was also the motivator for two interviewees (1 & 4), where the 'crowd' should rival software engineering experts and medical experts, respectively. Even if this endeavour was successful, it would not be 'the crowd' that achieves this, but the crowd and a combination mechanism, i.e. an aggregation algorithm, with – as just noted – considerable work needed to stabilise this process. Similar to the director of the Daresbury laboratory, it is the distributed agency between the initiator of the crowdsourcing, the crowd workers, well designed 'atomic tasks', the platform and the aggregation algorithm that makes the performance of the crowd possible (Law, 1994).

## 7.7. A dialogue-oriented alternative

Let us go back one last time to interview 5, where there was no 'atomic task' and no aggregation taking place. How can we make sense of this case and its work processes? This case is not only special and different in the amount of context necessary and thus provided, but it is also different in that crowd workers openly and directly cooperate to solve the tasks. In the other cases, the crowd workers work, as Berry (2019) noted, insulated from each other: No crowd worker could see the work done by their peers. This is different in interview 5, where workers create annotations and could add comments to their own annotations and those of others. They could even use a flagging system pro-actively notifying other crowd workers, but also a supervisor, of doubts and uncertainties. This approach allows for deliberation, for the revision of assessments and embraces a hesitant and reflexive way of working.

As noted above, the combination mechanism works as a mediator for the other, more 'typical' crowdsourcing approaches. The contrast is even stronger if a crowdsourcing approach is used where the majority voting decides over the remuneration, as is possible on Amazon Mechanical Turk. Here, the workers are brought into a situation of competition, with the remuneration of one worker depending on the result of other workers.

The contrast between the collaborative, deliberative work process of interview 5 and the competition and individualisation of the other interviews immediately reminded me of Hacking's (1990) study of the jury systems in France and England: Just as in my cases, France used majority voting, whereas, in England, the decision was to be made in consensus – and in Scotland, uncertainty was acknowledged by the third option of 'not proven'.

# 8. Conclusion

The ongoing digital transformation has lead to an ever-growing amount of data to be produced, circulated and consumed. For an increasing number of topics, people turn to this data to solve problems, to address questions. After all, data gets hailed as *the* new, valuable resource, and taking advantage of data is often framed as beneficial, sometimes even as a necessity to remain competitive, e.g., as a business.

Numerous scholars in STS and neighbouring disciplines have pointed out that data is not 'simply out there' in its 'raw' form but gets produced for a specific purpose and always needs interpretation (Bowker, 2005; boyd & Crawford, 2012). Crowdsourcing of data (sets) plays a dual role in the life of data: On the one hand, it is a means to make sense of data produced elsewhere. On the other hand, it is itself a production of (meta)data about this data: Data points get annotated and labelled with additional information to make it useful to solve a particular problem.

Crowdsourcing promises to be an efficient approach to use the huge amounts of data produced nowadays: By outsourcing to a large 'crowd', it comes attached with hopes to at once be a scalable process and produce results that rival experts. Commercial platforms even provide wrappers that promise access to the 'crowd' as if it was a piece of software.

Practitioners turn to crowdsourcing to make data sets useful. The data set itself at this point is uncertain; practitioners cannot readily use it to address the problem. Certainty has to be established: How much offensive content is present in a data set of social media postings? Where in a data set of recordings does speech occur? Where in a medical study's abstract are the patients' conditions mentioned? Crowdsourcing promises to be a suitable way to address these uncertainties, to solve the problem.

But as practitioners open the black box of crowdsourcing, new sources of uncertainty present themselves: How easily can the data be interpreted? How can one know if the 'crowd' did good work and annotate 'correctly'? How can the problem be reformulated and formalised to be fit for crowdsourcing? Who are the people that form the 'crowd'? Can they be trusted?

In my research, I could identify a dominant way of addressing these issues. At the same time, I encountered limits to this approach. There are two crucial elements to this approach that is shared by many crowdsourcing applications: First, a task gets solved by several crowd workers. To make use of these multiple results, they have to be aggregated, i.e. combined, to create one final annotation per data point. Second is a so-called 'atomic task', i.e., reformulation, formalisation, and discretisation of the original problem into small micro-tasks. These tasks often have the form of single-choice questions. This way, they become amenable to calculation,

making it possible to calculate agreement measures and easily aggregate the results, e.g. through majority voting. This approach allows practitioners to spread the uncertainty across several people, and at the same time, make the uncertainty calculable.

Often practitioners have available or produce themselves a (smallish) so-called 'ground truth', i.e. a data set where the expected annotation is already known. This 'ground truth' can play an important role in several ways. First, Practitioners can use it to address uncertainties about the crowd. It can be used to initially select crowd workers depending on their agreement with this 'crowd truth'. They can also use it to intersperse test tasks and monitor how frequent crowd workers agree with the 'ground truth'. Depending on their agreement, it is possible to exclude them from further participation and even deny remuneration for their work.

A 'ground truth' can also be used to train crowd workers. This can be necessary because the 'atomic task' does not only have to fit an aggregation method. It also has to work as immutable mobile (Latour, 1987) and faithfully transport the intended meaning. Training can help to ensure that the crowd workers read the task as expected. Of course, instructions and guidelines are useful, too. However, it's uncertain what achieves the best results even there: If the instructions are too extensive, crowd workers may simply ignore them.

Context information is often required to solve a task successfully. This can, to some extent, be provided in the described approach of task formalisation and decomposition into 'atomic tasks', which in turn is required for aggregation mechanisms. But, as my study shows, not all data can be compartmentalised into 'atomic tasks' that can be worked on in isolation from other data points. For certain problems, for certain tasks, data points have to be interpreted and evaluated in light of other data points and their relation to them. For this kind of data and task, the dominant approach to crowdsourcing easily fails. In these cases, alternative ways of crowdsourcing are needed: One way is to enable a more cooperative and deliberative way of collaboration among crowd workers instead of isolating the crowd workers from each other and putting them into a competitive situation, as is often the case on commercial platforms and to refrein from discretisation of the task.

To sum up the above, not only is crowdsourcing itself a means to address uncertainty about a data set, but during the process, uncertainty shifts – from the task formalisation and decomposition to the 'crowd', to the aggregation mechanism. This, however, is not a strictly linear process but a back-and-forth where the different elements of crowdsourcing get readjusted and reconfigured, and uncertainties about these elements are gradually addressed.

How the practitioners decide among these possible ways of conducting crowdsourcing is, however, often heavily dependent on an installed base, either in their (scientific) community or

within their organisation: Building on top of an installed base allows to set up the crowdsourcing efficiently, but frequently needs appropriation to make it work for the given application. The installed base also comes with limitations, sometimes requiring the creation of entirely new platforms due to, e.g., the form of collaboration and data not being supported.

What my study shows is that there is tremendous effort necessary to make crowdsourcing 'work'. Thus, crowdsourcing involves considerable human labour not only on the side of the crowd but also on the side of those that initiate the crowdsourcing. This is a form of labour that is often (made) invisible when discussing data, data analytics and 'AI' applications that often build on top of the resulting data sets (Newlands, 2021). The necessary labour of practitioners is invisible even in critical accounts that compare crowdsourcing to commodity infrastructure such as tap water (Berry, 2019).

The result of crowdsourcing is another data set: A data set augmented with additional metadata. As I discuss in this thesis, this additional data gets produced with a certain context in mind that is not simply 'in' the original data but has to be provided *additionally*. This begs the question if the data produced by crowdsourcing can be meaningful without additional context information. Is the context inscribed into the annotations? If so, can the data set be re-used in a meaningful way beyond the original problem formulation? If not, how can the crowdsourcing process be documented appropriately to make this restriction clear (cf. Gebru et al. (2020); Miceli et al. (2021))?

Finally, my study shows that a deliberative, cooperative and dialogue-oriented crowdsourcing process is possible in some cases. This is in stark contrast to the mathematical aggregation mechanisms dominantly used in crowdsourcing. The latter approaches tend to favour the work of a 'majority' at the cost of a deviating minority. This, on the one hand, silences the minority, eradicating their contribution from the final data set, but also can lead to the deviant workers getting paid less. Could crowdsourcing strategies that embrace this disagreement as a legitimate contribution work for a broad spectrum of tasks? Could strategies work that value disagreement as a productive source of uncertainty, that see it as "a starting point for an exploration intended to transform and enrich the world in which we decide to live" (Callon et al., 2009, p. 257)? I think this would be an interesting space for further research.

## References


Adams, J., & Brückner, H. (2015). Wikipedia, sociology, and the promise and pitfalls of Big Data. *Big Data & Society*, *2*(2), 1–5.

Akrich, M. (1992). The de-scription of technical objects. In W. E. Bijker & J. Law (Eds.), *Shaping Technology/Building Society* (pp. 205–224). Cambridge, MA: The MIT Press.

Amoore, L. (2019). Doubt and the algorithm: On the partial accounts of machine learning. *Theory, Culture & Society*, *0*(0), 1–23.

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93-117.

Aroyo, L., & Welty, C. (2013). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci'13*.

Ashton, P., Weber, R., & Zook, M. (2017). The cloud, the crowd, and the city: How new data practices reconfigure urban governance? *Big Data & Society*, *4*(1), 1–5. doi: 10.1177/2053951717706718

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671–732.

BBC News. (2018, December 19). Facebook's data-sharing deals exposed. *BBC News*. Retrieved July 25, 2021, from https://www.bbc.com/news/technology-46618582

Bechmann, A., & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, *6*(1), 1–11. doi: 10.1177/2053951718819569

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., ... Panovich, K. (2015). Soylent: A word processor with a crowd inside. *Communications of the ACM*, *58*(8), 85–94. doi: 10.1145/2791285

Berry, D. M. (2019). Against infrasomatization: Towards a critical theory of algorithms. In D. Bigo, E. Isin, & E. Ruppert (Eds.), *Data politics: Worlds, subjects, rights* (pp. 43–63). London/New York: Routledge.

Bloor, D. (1991 [1976]). The strong programme in the sociology of knowledge. In *Knowledge and social imagery* (pp. 3–23). Chicago: The University of Chicago Press.

Bowker, G. C. (2005). *Memory practices in the sciences.* Cambridge, Massachusetts: MIT Press.

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2009). Toward information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of internet research* (pp. 97–117). Dordrecht: Springer.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences.* Cambridge, Massachusetts: MIT Press.

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Brabham, D. C. (2013). *Crowdsourcing.* Cambridge, Massachusetts: MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of machine learning research* (Vol. 81, pp. 1–15). New York: PMLR. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In *Power, action and belief: A new sociology of knowledge?* (pp. 196–233). London: Routledge & Kegan Paul.

Callon, M. (1991). Techno-economic networks and irreversibility. In J. Law (Ed.), *A sociology of monsters* (pp. 132–161). London and New York: Routledge.

Callon, M., Lascoumes, P., & Barthe, Y. (2009). *Acting in an uncertain world: An essay on technical democracy* (G. Burchell, Trans.). Cambridge, MA: MIT Press.

Callon, M., & Latour, B. (1981). Unscrewing the big leviathan: How actors macrostructure reality and how sociologists help them to do so. In K. Knorr-Cetina & A. Cicourel (Eds.), *Advances in social theory and methodology* (pp. 277–303). Boston, London and Henley: Routledge & Kegan Paul.

Cha, A. E. (2015, May 19). Health and data: Can digital fitness monitors revolutionise our lives? *The Guardian*. Retrieved July 25, 2021, from https://www.theguardian.com/society/2015/may/19/digital-fitness-technology-data-heath-medicine

Charmaz, K. (2016). *Constructing grounded theory. a practical guide through qualitative analysis.* London: Sage.

Clarke, A. E. (2005). *Situational Analysis. Grounded Theory after the postmodern turn.* Thousand Oaks/London/New Delhi: Sage.

DBpedia Association. (2021). *About DBpedia.* Retrieved June 11, 2021, from https://www.dbpedia.org/about/

Desrosières, A. (1998). *The politics of large numbers: A history of statistical reasoning.* Cambridge, Mass: Harvard University Press.

Dewey, J. (1946). *The public and its problems: An essay in political inquiry.* Chicago: Gateway Books.

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society*,

*2*(2), 1–6.

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism.* Cambridge, MA: The MIT Press. Retrieved from https://data-feminism.mitpress.mit.edu/

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, *54*(4), 86–96.

Douglas, M. (1966/1969). *Purity and danger: An analysis of concepts of pollution and taboo.* London: Routledge & Kegan Paul.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* New York: St. Martin's Press.

Finlay, L. (2012). Five lenses for the reflexive interviewer. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (2nd ed., pp. 317–331). Thousand Oaks: SAGE.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2020). Datasheets for datasets. *arXiv:1803.09010 [cs]*. Retrieved 2021-06-06, from http://arxiv.org/abs/1803.09010

Geertz, C. (1973). *The interpretation of culture.* New York: Basic Books.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020, January). Garbage in, garbage out?: Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 325–336). Barcelona Spain: ACM. doi: 10.1145/3351095.3372862

Gieryn, T. F. (1995). Boundaries of science. In S. Jasanoff, G. E. Markle, J. C. Petersen, & T. Pinch (Eds.), *Handbook of science and technology studies* (pp. 393–443). London: Sage.

Gillespie, T. (2016). Digital keywords: A vocabulary of information society and culture. In B. Peters (Ed.), (pp. 1–16). Retrieved from http://culturedigitally.org/wp-content/uploads/2016/07/Gillespie-2016-Algorithm-Digital-Keywords-Peters-ed.pdf

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass.* New York: Houghton Mifflin Harcourt.

Hacking, I. (1990). *The taming of chance.* Cambridge: Cambridge University Press.

Haraway, D. (2001). Situated knowledges. the Science question in feminism and the privilege of partial perspective. In M. Lederman & I. Bartsch (Eds.), *The gender and science reader*

(pp. 169–188). London: Routledge.

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv:1805.03677*. Retrieved from https://arxiv.org/abs/1805.03677

Howe, J. (2006, jun 1). The rise of crowdsourcing. *WIRED*. Retrieved June 11, 2021, from https://www.wired.com/2006/06/crowds/

Iliadis, A., & Russo, F. (2016, December). Critical data studies: An introduction. *Big Data & Society*, *3*(2), 1–7.

Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 611–620). Paris: ACM.

Jasanoff, S. (2017, December). Virtual, visible, and actionable: Data assemblages and the sightlines of justice. *Big Data & Society*, *4*(2), 205395171772447. doi: 10.1177/2053951717724477

Jaton, F. (2021a, January). Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society*, *8*(1), 1–15. doi: 10.1177/20539517211013569

Jaton, F. (2021b). *The constitution of algorithms: Ground-truthing, programming, formulating*. The MIT Press. doi: 10.7551/mitpress/12517.001.0001

Jensen, E. A., & Laurie, A. C. (2016). *Doing real research: A practical guide to social research*. London: Sage.

Kazimzade, G., & Miceli, M. (2020, February). Biased priorities, biased outcomes: Three recommendations for ethics-oriented data annotation practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 71–71). New York NY USA: ACM. doi: 10.1145/3375627.3375809

Kitchin, R. (2014). *The data revolution*. London: Sage.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29.

Kolly, M.-J., & Schmid, S. (2021, April 19). Sie ist hübsch. er ist stark. er ist lehrer. sie ist kindergärtnerin. *Republik*. Retrieved July 25, 2021, from https://www.republik.ch/2021/04/19/sie-ist-huebsch-er-ist-stark-er-ist-lehrer-sie-ist-kindergaertnerin

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger

(Eds.), *Proceedings of the 25th international conference on neural information processing systems - volume 1* (pp. 1097–1105). Red Hook, NY, United States: Curran Associates Inc.

Larson, C. (2018, August 20). Who needs democracy when you have data? *MIT Technology Review*. Retrieved June 12, 2021, from https://www.technologyreview.com/2018/08/20/240293/who-needs-democracy-when-you-have-data/

Latour, B. (1983). Give me a laboratory and i will raise the world. In K. Knorr Cetina & M. Mulkay (Eds.), *Science observed* (pp. 141–170). Thousand Oaks: Sage.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Cambridge, Mass: Harvard University Press.

Latour, B. (1990). Drawing things together. In M. Lynch & S. Woolgar (Eds.), *Representations in scientific practice* (pp. 19–68). Cambridge: The MIT Press.

Latour, B. (2005). *Reassembling the Social.* New York: Oxford University Press.

Latour, B., Woolgar, S., & Salk, J. (1986). *Laboratory life: The construction of scientific facts.* Princeton, New Jersey: Princeton University Press.

Law, J. (1986). On the methods of long-distance control: Vessels, navigation and the Portuguese route to India. In J. Law (Ed.), *Power, action and belief: A new sociology of knowledge?* (pp. 234–263). London: Routledge & Kegan Paul.

Law, J. (1994). *Organizing modernity.* Oxford, UK/Cambridge, Mass.: Blackwell.

Lee, T. B. (2010, May 16). Autopilot was active when a Tesla crashed into a truck, killing driver. *Ars Technica*. Retrieved July 25, 2021, from https://arstechnica.com/cars/2019/05/feds-autopilot-was-active-during-deadly-march-tesla-crash/

Lupton, D. (1999). *Risk (key ideas).* New York: Routledge.

MacKenzie, D. (1998). The certainty trough. In R. Williams, W. Faulkner, & J. Fleck (Eds.), *Exploring expertise: Issues and perspectives* (pp. 325–329). London: Palgrave Macmillan. doi: 10.1007/978-1-349-13693-3_15

Metz, R. (2016, March 24). Why microsoft accidentally unleashed a neo-nazi sexbot. *MIT Technology Review*. Retrieved July 25, 2021, from https://www.technologyreview.com/2016/03/24/161424/why-microsoft-accidentally-unleashed-a-neo-nazi-sexbot/

Miceli, M., Yang, T., Naudts, L., Schuessler, M., Serbanescu, D., & Hanna, A. (2021, March). Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 161–172). Virtual Event Canada: ACM. doi: 10.1145/3442188.3445880

Michael, M. (2016). *Actor-network theory: Trials, trails and translations* (1st ed.). Thousand

Oaks, CA: SAGE.

MIT Technology Review. (2013, January). A more perfect union. *MIT Technology Review*. Retrieved June 12, 2021, from https://www.technologyreview.com/magazines/a-more-perfect-union/

MIT Technology Review Insights. (2020, November 19). The promise of the fourth industrial revolution. *MIT Technology Review*. Retrieved July 25, 2021, from https://www.technologyreview.com/2020/11/19/1012165/the-promise-of-the-fourth-industrial-revolution/

Mol, A. (1999). Ontological politics: A word and some questions. *The Sociological Review*, *47*(1), 74–89.

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., ... Erickson, T. (2019, May). How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Glasgow: ACM. doi: 10.1145/3290605.3300356

Newlands, G. (2021). Lifting the curtain: Strategic visibility of human labour in AI-as-a-Service. *Big Data & Society*, *8*(1), 1–15. doi: 10.1177/20539517211016026

Newlands, G., & Lutz, C. (2020). Crowdwork and the mobile underclass: Barriers to participation in India and the United States. *New Media & Society*, 1–21. doi: 10.1177/1461444820901847

Nowotny, H. (2014). Engaging with the political imaginaries of science: Near misses and future targets. *Public Understanding of Science*, *23*(1), 16–20. doi: 10.1177/0963662513476220

Nowotny, H., Scott, P., & Gibbons, M. (2008). *The new production of knowledge: The dynamics of science and research in contemporary societies* (Reprinted ed.). London: Sage.

Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., & Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation*, *11*(11).

Pickering, A. (1995). *The mangle of practice: time, agency, and science*. Chicago, Ill.: Univ. of Chicago Press.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science*, *14*(3), 399–441.

Rapley, T. (2007). Interviews. In C. Seale, G. Gobo, J. F. Gubrium, & D. Silverman (Eds.), *Qualitative research practice* (pp. 15–33). London/Thousand Oaks: Sage.

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and

the epistemologies of Big Data. *Big Data & Society*, *3*(1), 1–6.

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., Milland, K., & Clickhappier. (2015). We are Dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1621–1630). Seoul: ACM. doi: 10.1145/2702123.2702508

Scott, M. (2018, March 18). Politicians worldwide raise questions about cambridge analytica's use of facebook data. *Politico*. Retrieved July 25, 2021, from https://www.politico.eu/article/a-time-of-reckoning-for-facebook-and-politics-online/

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 1–12.

Selbst, A. D., boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). Atlanta, GA: ACM. doi: 10.1145/3287560.3287598

Silverman, D. (2000). *Doing Qualitative Research*. London: Sage.

Spradley, J. (2002). Interviews. In M. E. Pogrebin (Ed.), *Qualitative approaches to criminal justice: Perspectives from the field* (pp. 44–53). London: SAGE.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134.

Suchman, L. (1995). Making work visible. *Communications of the ACM*, *38*(9), 56–64. doi: 10.1145/223248.223263

The Economist. (2017, May 6). The world's most valuable resource is no longer oil, but data. *The Economist*. Retrieved June 12, 2021, from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Vardi, M. Y. (2012, March). What is an algorithm? *Communications of the ACM*, *55*(3), 5–5.

Vertesi, J., & Ribes, D. (2019). Introduction. In *digitalSTS: A field guide for Science & Technology Studies* (pp. 1–10). Princeton & Oxford: Princeton University Press.

Wakabayashi, D. (2018, March 19). Self-driving uber car kills pedestrian in arizona, where robots roam. *The New York Times*. Retrieved July 25, 2021, from https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

Wikipedia. (2021, June 18). *reCAPTCHA*. Retrieved July 25, 2021, from https://en.wikipedia.org/wiki/ReCAPTCHA

Ziewitz, M. (2016). Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values*, *41*(1), 3-16.

Zou, J., & Schiebinger, L. (2018). Design AI so that it's fair. *Nature*, *559*(7714), 324–326.

# Appendices

## A. Original quotes

[1]orig. “die Daten, was ich von den crowdsourcing workers annotieren lasse, die sind schon annotiert, von medizinischen Experten, das heißt ich weiß schon ganz genau, von jedem einzelnen Beispiel, was [die richtige] Antwort wäre”

[2]orig. sie “[haben] gesehen, [dass] eigentlich sehr viele politiker sind geübt . . . aussagen zu treffen [die] nicht eindeutig . . . beleidigend sind, aber die wirklich sehr knapp an der linie sind”

[3]orig. “es muss, jeden-jedenfalls von meiner Sicht aus, es muss ständig, immer auch ein Mensch kontrollieren, und es müssen auch ständig Menschen weiter bewerten, ah, weil das was dieses Jahr aktuell ist und das was man-ah mit dem algorithmus dieses Jahr feststellt wird in zwei Jahren nicht mehr stimmen, man muss das immer nachjustieren.” (interview 2)

[4]orig. “das sind Aufgaben, die relativ einfach sind für auch nicht-Experten. Also es ging zum Beispiel drum, in einer Opern-Aufnahme zu erkennen, wann jemand singt, und wann nicht, oder in einer . . . Radio-Aufzeichnung zu erkennen, wann Musik gespielt wird, wann gesprochen wird und wann nicht”

[5]orig. “wir hatten . . . einmal ein Training, danach eine Pilotphase, wo wir eigentlich für vier Wochen Daten gesammelt und bewertet haben, und haben diese Daten genommen um zu schauen, okay, welche Fragen machen Sinn, wie ist es für unseren Freiwilligen gelaufen, was war für sie leicht, was war schwierig, was ist da, ahm, was ist da irgendwie falsch gelaufen, um dann mit dieser Information die ganzen Apps und Algorithmen, das Datensammeln und das Training für die-für die ah Freiwilligen neu zu gestalten”

[6]orig. Bei “dieser Studie, die vor zwei Jahren durchgeführt wurde, da haben die, damit sie diesen [Amazon] Turk workers beibringen, wie das funktioniert, so ein Dokument hingeknallt, das hat zehn Seiten, und da steht ganz genau drin beschrieben, wie dieser Task funktionieren sollte, . . . also drei Seiten für die Participants [der medizinischen Studie], drei Seiten für die Intervention, wie man das annotiert. Und natürlich lest sich das keiner durch, schon, vor allem wahrscheinlich nicht die Turk workers. Und diese zehn Seiten enthalten sehr viele Beispiele - das zeigt einen Satz an, und genau ein bestimmter Teil in diesem Satz ist markiert und das soll eben zeigen, so funktioniert das.”

[7]orig. “Und natürlich lest sich das keiner durch, schon, vor allem wahrscheinlich nicht die Turk workers. Und diese zehn Seiten enthalten sehr viele Beispiele - das zeigt einen Satz an, und genau ein bestimmter Teil in diesem Satz ist markiert und das soll eben zeigen, so funktioniert das. Und das große Problem, das ich eben gesehen hab, ist dass du dieses riesige Dokument hast, mit den zehn Seiten, du musst dich immer durchblättern und eben herausfinden, welches Beispiel passt jetzt gut”

[8]orig. “man will man [die crowd worker] nicht überfordern, nicht zu viele Informationen geben, weil die Beschreibungen sollen sehr kurz, präzise, genau sein, und man will sie quasi nicht mit Informationen überladen”

[9]orig. “Interviewte*r: es war erstens nicht klar, wie gut die Qualität der Daten ist, die man da bekommt
Interviewer: Mhm, was meinst du damit?
Interviewte*r: Dass es Leute gibt, die einfach schnell irgendwie [durchklicken] um geld zu verdienen und man muss sich dann irgendein system überlegen, wie so ein eingangs-check oder auch zwischendurch beispiele einstreuen, von denen man die Antwort weiß, um zu überprüfen, ob die noch aufpassen”

[10]orig. “Interviewer: wie hast du da gewusst, ob [die Daten] eigentlich gut annotiert ist oder nicht? Hast du da irgendwas...?
Interviewee: Nein, wusste ich nicht. Ich glaube ich hab mir vielleicht anfangs, ist ein bisschen länger her, ich glaub ich hab mir vielleicht angekuckt, was die bei diesen hundert Beispielen angeklickt haben, und die habe ich selber auch annotiert und geschaut obs da gravierende Abweichungen gibt. Es kann auch sein, dass ich einen annotator oder annotatorin angeschrieben hab und gefragt hab hey, wie ist das hier hätte ich eigentlich das erwartet, aber sonst, ich glaub vielleicht habe ich stichprobenartig mal reingeschaut . . . aber es schien alles vernünftig zu sein”

[11]orig. “Und, so möchte ich zum Beispiel rausfiltern, dass jetzt irgendwelche S[c]ammer daran arbeiten, die sich einfach nur schnell durchklicken.”

[12]orig. “Das einzige, wo man sagen könnte, es gibt richtig oder falsch war halt diese Hate Speech und deswegen sind die immer an die Tafel gegangen. Aber in den anderen Fällen muss man schon irgendwie zugeben, dass das was dich beleidigt mich vielleicht kalt lässt und umgekehrt. Also eine bestimmte Variation finde ich gut und die-die muss man da lassen.” (interview 2)

[13]orig. “Interviewer: Aber quasi die [Freiwilligen] sind als vertrauenswürdig eingestuft worden, dass die da wirklich ernsthaftes Interesse haben, im Sinne eurer Arbeit mitzumachen Interviewte*r: genau, genau” (interview 2)

[14]orig. “[Gegenseitige Kontrolle fand] Informell [statt] eigentlich, also es waren immer, es waren kleine Teams, also zwei bis drei Personen, die, zum Beispiel am Atlas gearbeitet haben, und da war meistens einer der Haupt-Experte, und andere waren eben PhD-Studenten typischerweise. Also das mehr informelle Kontrolle durch ein normales Betreuungsverhältnis gewesen”

[15]orig. “Aber ich mach – das macht man normal, damit man majority voting machen kann, damit man drei Sätze übereinander legt und schaut, okay, zwei sagen das, einer sagt das, also ist das, was die zwei sagen richtig. Aber ich machs nicht für majority voting, sondern einfach, damit man ein bissl Varianz rauskriegt, weil wenn ich jetzt nur einen [worker] fragen würde, pro Satz, dann könnts ja sein, dass es ein lustiger Zufall ist, dass der grade mit diesen medizinischen Experten übereinstimmt.”

[16]orig. “Natürlich machen diese medizin-medizinischen Experten auch ab und zu Flüchtigkeitsfehler”

[17]orig. “Außerdem ist es schwierig bei dieser Aufgabe. es ist ja jetzt nicht so, ich hab ein Bild und möchte labeln was da drin vorkommt, und kann dann schauen was sagt die Mehrheit, sondern es sind halt, es ist ein Musikstück, wo jeder fünf bis zehn Grenzen einzeichnet und es ist nicht so klar, wie man das dann kombiniert. Weil was die eigentlich machen ist halt Segmente bestimmen. Zu sagen, das ist eine Einheit, das ist eine Einheit. Das ist nicht so, lässt sich glaub ich schwieriger kombinieren die, ah, widersprechenden Annotationen von mehreren Annotatoren.”

[18]orig. “Ursprünglich begonnen hat das ganze, weil wir Landkarten transkribieren wollten, also historische Landkarten aus einer Bibliothekssammlung, . . . die sollten transkribiert und georeferenziert werden, also das war so wirklich dieser Expert-Sourcing Task der damals gemacht wurde. Das heißt wo Leute aus dem Bibliotheksbereich selbst, die mit der Materie vertraut waren, ah, Transkriptionen von den Ortsnamen und den Karten gemacht haben, und die auch . . . Geographischen Punkten zugeordnet haben, so dass man einfach Metadaten extrahiert aus diesen Karten. Das war eigentlich der Hauptaufgabenbereich. Und daraus ist eben das Tool entstanden und das wird für unterschiedlichste Dinge eingesetzt, von denen ich jetzt sagen würde, ist nur ein kleiner Teil wirklich crowd-sourcing, es geht eher um Kommentar-Funktion im-im Unterricht oder generell wenn Historiker mit Quellen arbeiten und so weiter.” (interview 5)

[19]orig. “. . . man hat eine riesengroße Landkarte, da sind, ja so um die acht-, neunhundert Ortsnamen drauf,

die sind aber wirklich für einen Laien schwer lesbar. Also ich persönlich hab viele von denen gesehen, aber als Informatiker kann ichs kaum lesen, na also man braucht wirklich Erfahrung, damit umzugehen" (interview 5)

[20]orig. "[wir] versuchen ... halt diesen Spagat zu finden, wo man sagt, man hat Experten, aber man will trotzdem Leute einbinden, die jetzt nicht top-Experten sind, sondern einfach eben Studenten typischerweise und eine Plattform bieten, die diese Arbeitskraft irgendwie bündeln kann, mehr oder weniger wie crowdsourcing, auch wenns nicht direkt crowdsourcing ist" (interview 5)

[21]orig. "es geht eigentlich nur darum zu schauen, kann ich die workers dazu bringen, dass sie fast so gut sind wie diese medizinischen Experten, das ist quasi die Aufgabe." (interview 4)

[22]orig. "[Die NGO hat] eine Tafel erstellt, mit auch Forschern aus den Universitäten, Lunguisten, Soziologen, und die haben sich zusammengetan als Experten, die dann das letzte Wort zu den Kommentaren abgeben würden"

[23]orig. "wenn nicht alle die gleiche Bewertung abgegeben haben, dann wurde dieser Kommentar an ein zentrales Kommittee weitergeleitet" (interview 2)

[24]orig. "mit zwei Ausnahmen. Wenn irgendwer auf 'Hate Speech' drückt, geht der Kommentar zum Tafel. Weil hate speech, es ist heikel, ich will nicht sagen, es ist hate speech, außer es gibt auch Experten die das auch wirklich bestätigen. Und das andere, wenn jemand sagt, kommentar ist irgendwie zweideutig, oder- dann ging es an die Tafel, also die gingen immer an die Tafel ..., die musste dann entscheiden, ist es wirklich hate speech ..., oder haben die's einfach nicht verstanden" (interview 2)

[25]orig. "da war meistens einer der Haupt-Experte, ahm, und andere waren eben PhD-Studenten typischerweise. Also das mehr informelle Kontrolle durch ein normales Betreuungsverhältnis gewesen" (interview 5)

[26]orig. "[Es ist] ein bisschen zu kompliziert, [Amazon Mechanical Turk] aufzusetzen und man weiß dann nicht, genau, was man davon hat, ob man gute Daten bekommt darüber und man nutzt irgendwie Leute in prekären Situationen aus, das sind alles Gründe, die dagegen gesprochen haben, das zu probieren überhaupt" (interview 3)

[27]orig. "... und drittens, wie ich gesagt hab, weil man damit Menschen in prekären Situationen ausnutzt, die dann sehr wenig Geld dafür bekommen irgendwie Arbeit für uns zu machen" (interview 3)

[28]orig. "die vielleicht sich jetzt einfach schnell ein bissl was dazu verdienen wollen, weil Millionär kann man nicht werden und ich weiß nicht, wie gut man davon leben kann, wenn man das hauptberuflich macht." (interview 4)

[29]orig. "Eine Kollegin hat mir gsagt, dass es schwer ist, einen richtigen Preis zu setzen. Ich hab ja Sätze, ich kann bei jedem Satz sagen, wenn ihr mir den annotierts, kriegts ihr so und so viel Geld. Und, die Idee ist ja, wenn ich mehr Geld bezahle, nachher sollt ich bessere Qualität kriegen. Aber was ich gehört hab ist quasi, wenn man mehr Geld zahlt, nachhern ist man noch attraktiver für [Scammer], oder für Leute, die das so ein bissl halbherzig machen. Nicht sehr schlecht aber so mittelmäßig, damit sie schnell ans Geld herankommen. Das heißt, man darf nicht zu viel Geld zahlen, aber auf der anderen Seite darf man auch nicht zuwenig Geld zahlen, weil sich die dann denken, na, das macht keinen Sinn, weil für einen Cent mach ich das nicht, das ist [es] mir nicht wert. Also das hab ich zum Beispiel interessant gefunden." (interview 4)

[30]orig. "Interviewte*r: Ehrlich gesagt, wenn jemand nur so mitmacht, nach .. (lacht) nach einem Tag macht er nicht mit, weil das ist schon eine sehr anstrengende Arbeit, die haben jeden Tag so 100 Kommentare bekommen, die sie bewerten mussten, und, ja das ist auch irgendwie psychisch ein bisschen anstrengend, weil es reicht auch ein unschöner Kommentar, um dir den ganzen Tag zu vermiesen
Interviewer: ja, ja
Interviewte*r: Also, ich hab sie nicht einmal selber bewertet, aber einfach im durchsortieren und berechnen sieht man ein paar sachen und-die .. ja, die können einen schon ein bisschen .. mitnehmen." (interview 2)

[31]orig. “Und es gibt unzählige Foren, es gibt ja dieses Turknation oder so, . . . wo die Turkers, quasi die Arbeiter, miteinander kommunizieren, und dass die da auch eine eigene community haben und die können nachher auch sagen, hey der eine stellt nur Schwachsinnstask online, für den arbeiten wir nicht mehr. Also die sind da auch organisiert untereinander, und das muss man alles, wenn man bei Mechanical Turk was reinstellt, im Hinterkopf behalten.” (interview 4)

[32]orig. “Interviewte*r: Mehr weiß ich nicht, dann müssten wir [die NGO] fragen, aber die haben das irgendwie regional aufgeteilt, sie haben in jeder Region ihren eigenen Chapter und die Region [finden dann] innerhalb ihrer Region Freiwillige und sie haben das wirklich versucht, über das ganze Land zu verteilen, dass nicht nur zentral irgendwo bewertet wurde, sondern dass wirklich das ganze Land mit bewertet hat. Interviewer: mhm, aber das heißt, das waren alles [NGO-]Mitglieder? Interviewte*r: Ja, Mitglieder, oder von ihnen irgendwie rekrutiert. Ich weiß jetzt nicht, ob die Freiwilligen alle wirklich Mitglieder sind.” (interview 2)

[33]orig. “und zweitens war es besonders für [die NGO] wichtig, dass sie ihre Basis mit einbeziehen, dass es eine Aktion ist von der Basis für die Basis.” (interview 2)

[34]orig. “Naja, auf der Seite kann sich grundsätzlich glaub ich jeder Anmelden, .. ich hab jetzt selbst keinen Account für die workers. .. Schwer zu sagen, aber ich weiß auf jeden Fall, es gibt ja diese Qualifikationen, das heißt ich könnte zum Beispiel angeben, ich möcht nur Leute, die einen Facebook-Account haben, ich möchte nur Leute, die einen High School-Abschluss haben und so weiter, das heißt ich glaub, dass das breit gefächert ist, aber vermutlich eher bei den Leuten, die vielleicht sich jetzt einfach schnell ein bissl was dazu verdienen wollen, weil Millionär kann man nicht werden und ich weiß nicht, wie gut man davon leben kann, wenn man das hauptberuflich macht. Aber welche Schichten da jetzt im Detail dabei sind, das, das könnt ich nicht sagen.” (interview 4)

[35]orig. “Und was mir eigentlich auch nicht klar ist, wie-wie das bezahlt wird, ob die Uni das einfach .. zahlen kann, die Plattform, weil jetzt haben wir halt so Werkverträge aufgesetzt mit den Studierenden. Ich weiß nicht was man– ob das ein Problem ist, ah, was bei Amazon Mechanical Turk zu kaufen, aus was für einem Budget man das machen würde” (interview 3)

[36]orig. “also .. sagen wir mal so, die App die kommt von mir, und da ich keine, da ich eigentlich nicht App designerin bin, ahm, ist es eine shiny App, ah, weil das ist was ich machen kann und [die NGO] hatte auch nicht die Ressourcen, um irgendwas fancy aufzustellen, haben sie übernommen, was ich zur Verfügung gestellt habe.” (interview 2)

[37]orig. “Also das-das schöne an Shiny ist, erstens einmal, ich entscheide, wie, was geht da rein, wie kommt es, und, ich kann das selber programmieren, dass . . . wenn bestimmte Themen auftauchen, wird dort drauf geklickt, dann macht automatisch ein anderes Fenster auf, wo ich weitere Sachen, ahm, ah bewerten kann. Ahm, das zweite ist, – es ist dadurch für mich sehr leicht, die Daten runter zu laden, zu samplen, und in die App reinzuschieben, weil das mach ich alles in R. Und müsste ich das irgendwie in einen Surveymonkey oder irgendwas machen, dann-dann wäre dann ein extra Schritt, wie kriege ich die Daten, die ich jetzt gesamplet habe in einem Format in diese App rein, wie speichere ich dann die Daten aus dieser anderen App, und wie bekomme ich sie wieder, damit ich sie bearbeiten kann und analysieren kann. Ahm, also dadurch, dass es alles in-in R ist, ist-ist-ist wenigstens der Workflow etwas leichter.” (interview 2)

[38]orig. “Interviewer: Also warum nochmal hast du dich gegen klassisches crowdsourcing mit den herkömmlichen Plattformen wie Amazon Mechanical Turk, oder ich weiß nicht, Crowd Flower ah-kenn ich noch, was gibt’s noch, Task Rabbit glaub ich, da gibt’s ja verschiedene
Interviewte*r: Mhm. Naja, erstens müsste man sich technisch mal damit beschäftigen, was man braucht um das aufzusetzen

Interviewer: Mhm.

Interviewte*r: Wahrscheinlich eh nicht viel mehr als das was ich dann schon hatte, wenn ich sowieso ein webinterface hab, aber .. eh, ich habs mir nie angeschaut" (interview 3)

[39]orig. "In dem Fall wars wieder weil die eben auch in diesem Paper 2018 Mechanical Turk verwendet haben. Ich weiß es gibt noch das Figure Eight, das ist glaub ich auch sehr gut, und .. ich glaube, wenn ich-wenn ich die Wahl hätte würde ich lieber dieses figure eight nehmen. Aber in dem Fall muss es eben so ähnlich wie möglich sein, damit die Ergebnisse eben sehr-sehr glaubhaft und überzeugend sind, die ich in meinem Paper beschreiben will. Drum hab ich in dem Fall eben gesagt, okay wir nehmen das selbe wie die, die haben halt Mechanical Turk verwendet" (interview 4)

[40]orig. "Also ich glaub, es, das spannende ist eh eigentlich dieses, dass es irgendwie ein bissl anders immer war, als so diese klassischen Äm-Mechanical Turk oder auch, ahm, galaxy zoo oder zooniverse Projekte. Also wir haben auch versucht, mit zooniverse, mit der Plattform zu arbeiten, haben das irgendwie nie wirklich hin gekriegt, weils doch irgendwie ned ganz gepasst hat, von den-von der infrastruktur, die da war" (interview 5)

[41]orig. "... und zweitens, ah, wurden alle Kommentare von drei verschiedenen Freiwilligen bewertet, also ich hab die-die in so kleine Packages aufgeteilt und jedes Package ging halt an drei verschiedene Bewerter ..." (interview 2)

[42]orig. "Interviewte*r: Also bei mechanical turk haben wir quasi dieses Design Interface, da kannst du HTML reinhauen, JavaScript reinhauen, da kannst du dir quasi eine eigenen Homepage drinnen basteln. Die Homepage, so wie du sie bastelst, so wird sie dann auch diesen Workers angezeigt, und, da bin ich so vorgegangen, dieser Aufgabe die ich habe, Text markieren, in zum Beispiel, a-in einem Text bestimmte Teile zu markieren, das nennt sich named entity recognition

Interviewer: mhm

Interviewte*r: Und nachher hab mal gesucht, named entity recognition für-für mechanical turk und nachher hab ich eben so eine HTML-Vorlage schon gfunden. Die hab ich dann genommen und modifiziert so wie ichs eben brauche, eben dass Sätze angezeigt werden, dass meine Daten eingespielt werden, und hab aus dem alten-aus dem anderen Interface eben rausghaut was ich nicht unbedingt brauche und so hab ich das angepasst" (interview 4)

[43]orig. "Interviewer: und gab's jetzt irgendwelche Einschränkungen beim Interface-Bau, die dich-die dich gehindert haben, ä-gibt's irgendwas, was du geändert hättest wenn-wenn du jetzt quasi vollkommene Kontrolle drüber gehabt hättest?

Interviewte*r: Na eigentlich nicht. Also das hat mich eigentlich positiv überrascht. Ich dachte, dass man, dass man da nicht einfach so sein JavaScript reinhauen kann und alles, aber es hat funktioniert, ohne Probleme" (interview 3)

[44]orig. "ahm, ich bin nicht von der App so begeistert, ah, also, wenn sich jemand findet, der eine bessere User Interface und so weiter entwickeln kann, würde ich-würde ich sofort machen, ahm .. aber trotzdem es ist besser als manuell Kommentare runterzunehmen, und es, es gibt ein-eine-eine Standardisierung, also es gibt immer das gleiche Formular und ich, es sind schon sachen dort aufgeschrieben, die für [die NGO] wichtig sind, ah, ja. Aber shiny ist bestimmt nicht das beste, was man für sowas benutzt." (interview 2)

[45]orig. "Interviewer: Mhm. Gut, weil Sie schon Amazon Mechanical Turk jetzt erwähnt haben, das ist natürlich irgendwie so, wenn man von crowdsourcing jetzt redet, so die–der große Name oder so. Gabs da Gründe, warum man nicht so bestehende Plattformen verwendet hat?

Interviewte*r: Ja für uns wars der content, ahm, wir haben eigentlich ned wirklich gesehen, wie man das in Amazon Mechanical Turk so reinbringen würde." (interview 5)

[46]orig. "... also das war so wirklich dieser Expert-Sourcing Task der damals gemacht wurde, das heißt wo Leute aus dem Bibliotheksbereich selbst, die mit den-mit-die-mit-die mit der Materie vertraut waren, ah, Transkriptionen von den Ortsnamen und den Karten gemacht haben, ..." (interview 5)

[47]orig. "Interviewte*r: dann haben sie einen Zugang bekommen, genau, und in dem Interface war noch dazu eine Anleitung, wie mans benutzen soll das
Interviewer: Mhm
Interviewte*r: interface, und im interface war auch eine guideline, was wir definieren als Gesang, und worauf sie achten sollen" (interview 3)

[48]orig. "Interviewte*r: ... und [da] gibt's dann halt auch Abweichungen, entweder sind Sachen ein bisschen verschoben, oder manche Übergänge fehlen einfach, und halt, also, da gibts gabs schon sehr detaillierte Guidelines dazu, wie das annotiert werden sollte
Interviewer: mhm
Interviewte*r: von denen, die das organisiert haben, aber es lässt trotzdem Raum für Interpretationen, und manchmal sagt man, aja, das ist eine Änderung und jemand anderes sagt, nein nein, das [alles] gehört zu einem Teil" (interview 3)

[49]orig. "damit sie lernen umzugehen mit der App, wie funktioniert das, was kann ich tun, was kann ich nicht tun" (interview 2)

[50]orig. "... dann, ah, kommt die wahl, negativ, aber nicht problematisch, weil ich kann, ah, mich auch kritisch äußern, ohne irgendwie jemanden zu beleidigen. Ahm, dann, Negativ und problematisch, also, irgendwie beleidigend oder diskriminierend, ahm, – dann hate speech, wobei, wir haben auch, ahm, Beispiele von hate speech, die geben, die ... von der EU wird das definiert, also haben wir das auch gemäß der Definition ahm, ahm, so vorgesehen und trainiert die leute, die das bewerten müssen. und dann gab's noch eine fünft wahl und das ist, ahm, ambiguous, ..." (interview 2)

[51]orig. "Na es ist ja wieder, es soll so simpel wie möglich sein. Das heißt, ich frage eine Gruppe von Workern nur–nur was sind–was sind die Teilnehmer der Studie. Dann eine andere Gruppe von Workern frage ich nur, was ist die Intervention. Also ich erstell quasi drei Unteraufgaben und die markieren einfach immer nur andauernd die Intervention, Intervention, Intervention, und die anderen markieren nur die Outcomes, das heißt da gibts kein hin und her switchen, es ist quasi drei–drei eigene Tasks." (interview 4)

[52]orig. "Die habens anders gemacht, die haben nämlich denen workers das komplette Abstract gezeigt und haben gesagt 'schauts her, das ist das Abstract, markiert mir alle, alle Participants, alle die was in dieser Studie teilgenommen haben', das heißt die haben das nicht für Sätze gemacht, das ist zum Beispiel eine, das ist meiner Meinung nach eine Verbesserung auf jeden Fall, ..." (interview 4)

[53]orig. "... weil die workers können sich ja jeden, jeden Task anschauen bevor sie damit starten. Das heißt wenn ein worker sieht, okay, das Abstract hat 50 sätze, das tu ich mir nicht an, das mache ich bestimmt nicht, da geh ich dann auf reject. Das heißt, das ist nicht wirklich fair, wenn du das ganze Abstract nimmst. Sätze sind ungefähr immer gleich lang, das ist nachher glaub ich dieses Problem minimierter. " (interview 4)

[54]orig. "semantisch anreichern können, das heißt sie können Personen auszeichnen, sie können Orte auszeichnen, Text im Bild und können, ah, einfach Kommentare dazu machen." (interview 5)

[55]orig. "Interviewte*r: ... also es ist immer in Teams gearbeitet worden, die Leute haben sich einzelne Pakete hochgeladen und dann die Arbeit verteilt und dann an-angefangen zu transkribieren, ahm, und das zum Teil halt einfach i-im Team oder zu zweit einfach gemacht, also sozusagen, es hat jetzt keinen formellen Task-ablauf gegeben, sondern die Leute haben einfach das Tool zum transkribieren benutzt

Interviewer: mhm

Interviewte*r: sich gegenseitig kontrolliert, ahm, Datenauswertungen gemacht, also so im Sinne von einer Kartendarstellung von den Punkten die aufgelöst worden sind und so weiter. Aber es war jetzt keine formelle, also wie man das jetzt bei Amazon Mechanical Turk [macht] oder so, dass das in ganz einfache Tasks gesplittet wird, das hat's da eigentlich nicht gegeben. Also es war sehr stark die Arbeit von Historikern einfach Tool-unterstützt." (interview 5)

[56]orig. "Also wenn man sagt, man macht eine Transkription, man hat aber seine Zweifel dran, kann man einen Kommentar reinschreiben und dann gibt's noch so diskussionsthreads drunter, das war das eine, ahm, das zweite war die Auflösung nach modernen Geokoordinaten, auch da war die-der Punkt, dass mans nicht immer identifizieren konnte, und da haben wir dann auch so ein Flagging-System eingeführt, wo man sagt, okay, ich kanns nicht auflösen, ah, und dann ist das auch im Gelb aufgepoppt, das heißt man hats auch immer gleich gesehen, wo waren denn die Probleme, und dann hat's, zumindest in der ersten Plattform, so Gründe gegeben warum man's auf Gelb gesetzt hat, wo man das dann weiter auseinanderfiltern hat kön-können, wenn man gesagt hat, ich glaube es ist ein Ort, aber ich finde ihn nicht, ich glaube, es ist ein Ort der überhaupt nicht existiert, weil der vielleicht einen Fehler gemacht hat, oder weils mythische Orte sind auf so Karten, und so eine ganze Kla-ah-eine mini-Taxonomie an Möglichkeiten woran sozusagen dieser Task gescheitert ist, hat's gegeben" (interview 5)

[57]orig. "Also man muss sich so vorstellen, man hat eine riesengroße Landkarte, da sind, ja so um die achtneunhundert Ortsnamen drauf, die sind aber wirklich für einen Laien schwer lesbar, also ich persönlich hab viele von denen gesehen, aber als Informatiker kann ichs kaum lesen, na also man braucht wirklich Erfahrung, damit umzugehen, und die Leute haben das dann einfach transkribiert und brauchen also sozusagen den Kontext rum. Also dass man sagt, man jetzt das alles in kleine Schnitzel zerschneiden und [wem] sagen, transcribe mir das, 'transcribe mir das', das funktioniert in dem Kontext eigentlich nicht. Sowohl von der–von den Inhalten her, als auch von der erforderlichen Expertise glaub ich." (interview 5)

[58]orig. "Also zooniverse funktioniert auch sehr stark auf der Basis, man-man zerlegt das alles in micro-tasks, man hat meistens kleine Bildchen, die man entweder irgendwie labelt oder transkribiert, und das hat bei uns einfach ned wirklich funktioniert" (interview 5)

[59]orig. "...jeden diesen einzelnen Sätze poste ich dann eben auf Mechanical Turk und damit die-die workers noch einen bestimmten Kontext haben zeige ich ihnen auch das komplette, den kompletten Abstract-Text quasi an und ich markiere darin schaut's her, der Satz den du annotieren sollst, der-der kommt in diesem Abstract hier vor. Das ist zum Beispiel wichtig, wenn in einem Satz kommt zum Beispiel vor, eine Abkürzung AD, das wäre Alzheimer Disease und der Mechanical Turk Worker hat natürlich wenn er nur den einen Satz sieht keine Ahnung, dass AD für Alzheimer Disease steht, aber meistens am Anfang im Abstract wird das, wird das halt erklärt wofür die Abkürzungen stehen. Und drum kanns auch wichtig sein, dass eben dieses ganze Abstract angezeigt wird, damit der Worker, falls er möchte den vollen Kontext hat in dem dieser Satz auftritt." (interview 4)

[60]orig. "Im schlimmsten Fall kann einer immer klicken "enthält nicht" absenden, "enthält nicht", absenden, aber ich kann, ich seh das natürlich nachher, ich hab zwei Tage Zeit, und kann sagen, dem Worker geb ich kein Geld, zum Beispiel." (interview 4)

[61]orig. "Interviewte*r: Oder, wenns bei einem Satz super offensichtlich ist, ich sag einmal es sind medizinische Daten, es ist nicht so leicht, aber es gibt manche Beispiele,wos sehr offensichtlich ist, was das richtige ist

Interviewer: mhm

Interviewte*r: und wenn sowas immer häufiger vorkommt, dass der immer sagt, na, falsch, falsch, und dann-dann seh ich okay, der nimmt das nicht wirklich ernst, und dann würd ich den zum Beispiel rejecten und der würde

nachher kein Geld bekommen, der Task wird automatisch wieder online gestellt, nachdem ich ihn rejected hab, damit ich da dann hoffentlich bessere Ergebnisse bekomme, von einem anderen Worker" (interview 4)

[62]orig. "Was eine best practice ist, ist du machst einen kleinen, einen kleinen durchlauf, und holst dir nur die workers, die sehr gut sind, das heißt du schaust dir für diesen kleinen Durchlauf, sagen wir zehn Sätze, und jeden Satz lass ich von 20 workers annotieren. Dann geh ich diese zehn Sätze für diese 20 Worker sehr genau durch und schau, der worker ist gut, der worker ist gut und nachher find ich zum Beispiel, dass 18 sehr gut arbeiten und zwei nicht sehr gut, dann kann ich zum Beispiel definieren, ich mach denn Task–ich geb jetzt wieder mehr Sätze online, aber es dürfen nur diese 18 mitarbeiten. Das heißt, die zwei dürfen nicht mehr mitarbeiten, und alle anderen auch nicht." (interview 4)

[63]orig. "das kann ich automatisch auswerten und ich kann automatisch sehen, der Worker hat eine 80%ige Accuracy und nachher könnt ich alle rauskicken, die weniger als 50% haben" (interview 4)

[64]orig. "Warum drei? Aaah – naja, da mit drei kann man schon einen tie breaker überlegen, das heißt . . . könnte man auch umstellen auf die Mehrheit gewinnt" (interview 2)

[65]orig. "das macht man normal, damit man majority voting machen kann" (interview 4)

[66]orig. "Und dadurch passiert das schon, dass wenn die sich einig sind kriegt das [neuronale] Netz an der Stelle zwei Mal das gleiche [Signal] vorgegeben und wenn sie sich uneinig sind kriegt es zwei mal unterschiedliche Sachen vorgegeben und wird sich halt auch unsicherer, oder lernt von beiden ein bisschen" (interview 3)

[67]orig. "Und wenn ich drei Leute frage, und alle drei stimmen mit den medizinischen Experten überein, dann kann ich sagen, schauts her, mein Tool ist wirklich sehr gut, weil die alle drei für diesen einen Satz übereinstimmen, drei unterschiedliche Mechanical Workers, und alle drei sagen genau das selbe wie dieser medizinische Experte." (interview 4)

[68]orig. "also eigentlich, das ist vielleicht interessant, zuerst wollten sie differenzieren, ahm, positiv, neutral, negativ, und dann die negativen weiter, ahm, differenzieren. dann – eh wir haben aber ein pilot gemacht, und ein paar mal einfach in kleineren gruppen ausgetestet, und, ah, wir haben dann festgestellt, differenzierung zwischen positiv und neutral ist noch schwieriger als irgendwie neutral/negativ, also haben wir die zwei einfach zusammengetan." (interview 2)

[69]orig. "und was ich noch mache, ich besprech mich mit Kollegen, wie sie das Interface empfinden, obs für sie Sinn macht, dieses Abstract zu zeigen, oder nicht, und, ja – also es ist noch nicht-es steht noch ein bissl in den Sternen, ob das jetzt überhaupt eine gute oder schlechte Idee ist" (interview 4)

## B. Interview guide

1. Please give me an overview of your project.
2. Please describe how a machine learning project works, how do you decide what data you need, how the data should be labelled and how do you decide if you need crowdsourcing?
3. How do you decide if you need crowdsourcing?
4. Would you say your project is a typical project or is it somehow special?
5. How do you decide which crowd-sourcing platform to use? Do you need to configure it in some way?
6. Are there any limitations imposed by the crowdsourcing platforms – what would you change if you could?
7. Can you tell me, who are the people labelling your data?
8. Can you explain what needs to be done to use the results from the crowd? Do you need some form of cleaning or processing, quality checks and so forth?
9. Can it happen that parts of the data are ambiguous? In which way can they be ambiguous? How do you deal with ambiguity?
10. Can you tell me how do you know if the data set is good, if it is usable? How do you know it will work for your application?
11. How could you fix bad data sets?

# C. Informed consent form

## Informed Consent

**Informed consent to participate in an interview and subsequent analysis for research purposes**

| | |
|---|---|
| **Purpose of interview** | Master thesis “Ambiguous supervision” (working title) on how ambiguities in the context of crowd sourcing of data sets for machine learning and knowledge bases are dealt with in practice. |
| **Responsible person** | _____ |
| **Funding** | none |
| **Further involved** | Univ.-Prof. Ulrike Felt (supervisor) |

I agree to participate in an interview for the Master thesis project “Ambiguous Supervision”. It addresses how ambiguities in the context of crowd sourcing of data sets for machine learning and knowledge bases are dealt with in practice.
The interview will be discussed in anonymized form with other students and supervising staff.
My participation is voluntary. If I want, I can call off the interview at any time.
I agree to the interview being recorded. All recordings are being stored on encrypted computers, treated confidentially and protected from access of third parties. If I call off the interview, any recordings will be deleted.
I am aware that quotes may be made public as part of the master thesis and potential further scientific publications. All quotes will be anonymized and my identity will not be revealed.

Name of participant: _____

Place, Date: _____ , _____

| | |
|---|---|
| _____ | _____ |
| Participant | Responsible person |

## D. English abstract

Crowdsourcing is a common approach to annotate a data set to be analysed directly or used for 'Artificial Intelligence' ('AI') applications. An initiator distributes tasks to crowd workers, who then annotate the data point. Turning to crowdsourcing exposes the initiator to multiple sources of uncertainty: How the task should be designed, who is part of the crowd, how to best make use of the annotations, and how to know if the crowd's work is any good are causes for concern. From a Science and Technology Studies perspective, this study investigates how practitioners that crowdsource data sets address uncertainties during this process. Adopting the stance of Actor-Network Theory, this thesis analyses the actual, messy practice of building a stable actor-network that is crowdsourcing. To achieve this, I conducted qualitative interviews with practitioners and analysed them using Situational Analysis.

In this study, I identify strategies to address uncertainty shared among most approaches to crowdsourcing. Among them is the decomposition of the problem into small 'atomic tasks'. They often involve single-choice questions, which makes them amenable to calculation. If this is the case, several crowd workers can annotate each data point, and their results get combined through a mathematical aggregation mechanism. This allows the initiator to spread uncertainty across crowd workers and make it quantifiable. Finding a suitable task design for this approach is difficult and involves extensive iterative experimentation where agency shifts between the initiator, crowd, and aggregation mechanism. These aggregation mechanisms often privilege a majority while silencing crowd workers that deviate from it.

The 'atomic tasks' have to act as faithful intermediaries. To make this possible, context information is crucial. I show in this thesis that, depending on the amount and type of context necessary, this puts a limit to task decomposition. The importance of context raises the question of how this context gets inscribed in the annotations and how this, in turn, contributes to biased data sets and potential 'AI' applications building on top of them.

As the 'crowd' can be anonymous or seen as deficient compared to experts, initiators use different forms of supervision to monitor their work. An existing, small 'ground truth' data set with known results plays an important role here: It can be used to select 'good' workers upfront, or it gets used for test tasks that help evaluate the workers. Workers can also be compared to their peers. The initiator then can discipline them by excluding them and denying payment.

My study shows that it is not straightforward to make crowdsourcing work but takes tremendous effort, labour that often remains invisible and hidden. At the same time, I show how epistemic approaches, whether the initiators consider the data as disputed and how this gets acknowledged, informs the structure of crowdsourcing processes.

## E. German abstract

Crowdsourcing ist ein etablierter Ansatz, um Datensätze zu annotieren, sodass sie analysiert werden können oder die Basis für Anwendungen von 'Künstlicher Intelligenz' ('KI') bilden. Ein*e Auftraggeber*in verteilt dabei Aufgaben an sogenannte Crowd Worker die Datenpunkte annotieren. Crowdsourcing konfrontiert die Auftraggeber*innen mit multiplen Quellen von Ungewissheit: Wie soll die Aufgabe gestaltet werden, wie können Annotationen am besten genutzt werden, und wie weiß man, ob die Crowd gute Arbeit leistet? All dies verursacht Ungewissheit. Diese Studie untersucht aus der Perspektive der Wissenschafts- und Technikforschung wie Praktiker*innen Ungewissheiten im Rahmen von Crowdsourcing-Prozesses umgehen. Dabei betrachte ich diese Frage aus der Sicht der Actor-Network Theory und analysiere die tatsächliche, chaotische Praxis des Crowdsourcings. Ich führte qualitative Interviews mit Praktiker*innen durch und analysierte diese mit Situational Analysis.

In dieser Studie identifiziere ich verschiedene Strategien zum Umgang mit Ungewissheit. Diese Strategien werden von den meisten Crowdsourcing-Ansätzen verwendet. Darunter ist der Ansatz, das ursprüngliche Problem in kleine 'atomare Aufgaben' zu zergliedern. Diese Aufgaben haben oft die Form von Single Choice Fragen, die es auch erlauben, rechnerisch die Ergebnisse zu verarbeiten. In diesem Fall ist es möglich, jeden Datenpunkt durch mehrere Crowd Worker annotieren zu lassen und die Ergebnisse durch Aggregation zu kombinieren. Dadurch kann der Auftraggeber die Ungewissheit über mehrere Personen verteilen und quantifizieren. Es ist aber schwierig, die Aufgaben adäquat zu gestalten, es benötigt oft mehrere Iterationen im Zuge derer die Handlungsmacht zwischen Auftraggeber, Crowd und Aggregationsmechanismus zirkuliert. Dabei privilegieren die Aggregationsmechanismen oft eine Mehrheit und bringen jene, die davon abweichen zum Verstummen.

Die 'atomaren Aufgaben' müssen als getreue Vermittler agieren. Um das zu erreichen ist Kontextinformation wichtig. I zeige in dieser Arbeit, dass je nach Umfang des notwendigen Kontexts damit dem Ansatz der Aufgaben-Zergliederung eine Grenze gesetzt ist. Die Wichtigkeit von Kontext führt auch zur Frage, wie dieser Kontext in die Datensätze eingebettet wird und in weiterer Folge in potentielle, darauf aufbauende 'KI'-Anwendung.

Auftraggeber*innen verwenden verschiedene Möglichkeiten, um die Crowd Worker zu beaufsichtigen, da die Crowd anonym sein kann oder als defizitär im Vergleich zu Expert*innen betrachtet werden kann. Ein kleiner 'Ground Truth'-Datensatz bei dem die Resultate bekannt sind spielt hier eine wichtige Rolle. Dieser Datensatz kann dazu verwendet werden, um anfangs 'gute' Crowd Worker auszuwählen, oder um Testaufgaben zu erstellen, mit denen die Crowd Worker evaluiert werden können. Die Crowd Worker können auch mit ihren Kolleg*innen ver-

glichen werden. Im Anschluss können Auftraggeber*innen Crowd Worker ausschließen und die Bezahlung verweigern.

Diese Studie zeigt, dass Crowdsourcing nicht 'einfach' gemacht werden kann, sondern grossen Aufwand bedarf. Diese Arbeit bleibt oft unsichtbar und versteckt. Gleichzeitig zeige ich wie erkenntnistheoretische Einstellungen, ob Auftraggeber*innen Daten als strittig anerkennen, die Gestaltung des Crowdsourcing-Prozesses beeinflussen.