Minds and Machines (2021) 31:563–593
https://doi.org/10.1007/s11023-021-09573-8

GENERAL ARTICLE

Linking Human And Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning

Thilo Hagendorff[1]

Received: 28 July 2020 / Accepted: 9 September 2021 / Published online: 26 September 2021

Abstract

Machine behavior that is based on learning algorithms can be significantly influenced by the exposure to data of different qualities. Up to now, those qualities are solely measured in technical terms, but not in ethical ones, despite the significant role of training and annotation data in supervised machine learning. This is the first study to fill this gap by describing new dimensions of data quality for supervised machine learning applications. Based on the rationale that different social and psychological backgrounds of individuals correlate in practice with different modes of human–computer-interaction, the paper describes from an ethical perspective how varying qualities of behavioral data that individuals leave behind while using digital technologies have socially relevant ramification for the development of machine learning applications. The specific objective of this study is to describe how training data can be selected according to ethical assessments of the behavior it originates from, establishing an innovative filter regime to transition from the big data rationale $n = all$ to a more selective way of processing data for training sets in machine learning. The overarching aim of this research is to promote methods for achieving beneficial machine learning applications that could be widely useful for industry as well as academia.

Keywords Artificial intelligence · Machine learning · Machine behavior · Technology ethics · Training data · Data quality

✉ Thilo Hagendorff
thilo.hagendorff@uni-tuebingen.de

[1] Cluster of Excellence "Machine Learning - New Perspectives for Science", University of Tuebingen, Tübingen, Germany

1 Introduction

When developing learning software, practitioners have additional ethical responsibilities beyond those of standard, non-learning software (Wolf et al., 2017). These responsibilities originate in the careful selection of inputs that build the very basis for the computational learning process itself. In a recent paper on machine behavior, Rahwan et al. stress that machine learning applications "cannot be fully understood without the integrated study of algorithms and the social environments in which algorithms operate" (Rahwan et al., 2019, p. 477). With regard to supervised machine learning, meaning artificial neural networks, support vector machines, naive Bayes classifiers, regression algorithms etc., those "social environments" can, among others, be understood as different training stimuli that shape the behavior of a machine. Machine behavior can be seen in analogy to the behavior of biological agents as an observable response to (internal or) external stimuli. Training data fed into supervised machine learning applications reflect, in case it is about behavioral data, people's (e.g., discriminative) behavior, so people's behavior has an indirect influence on machine (discriminative) behavior (Barocas & Selbst, 2016). This influence cannot be described as a direct relationship, meaning as an equivalence between people's behavior and machine behavior. However, particular traits of learning machines, that is how they solve for instance classification, prediction, or generation tasks, can originate in similar features that are part of behavioral data sets. Thus, when technology ethicists talk about "moral machines" (Wallach & Allen, 2009) in the context of machine learning applications, one also has to ask for "moral people" and "moral people's data", to put it simply. Of course, these "moral machines" are also the result of engineering or design choices, they are dependent on the selection of hyperparameters or specific wirings of artificial neural networks, and the like. But in general, today's machine learning techniques are dependent on human participation. In many cases, they harness human behavior that is digitized by various tracking methods. These machine learning methods do not create intelligence, but, taking up the figurative words of Mühlhoff, "capture" (2019, 1873) it by tracking human cognitive and behavioral abilities. Without the empirical aggregation of recordings of human behavior, many parts of machine learning would not work. An extensive infrastructure for "extracting" (Crawford, 2021, p. 15) valuable personal data or "capturing" human behavior in distributed networks via user-generated content, expressed or implicit relations between people, as well as behavioral traces (Olteanu et al., 2019) builds the bedrock for a computational capacity called "artificial intelligence" (Mühlhoff, 2019).

Here, I want to ask whether there are differences in the "quality" of human participation in artificial intelligence. To do this, one must further answer the question about what constitutes "good" influences or "good" behavioral datasets for supervised machine learning applications. In order to accomplish this, I will focus on identifying data sources reflecting behavior that is ethically sound, which in turn can be identified via scrutinizing particular states and traits of an individual that are to be described in more detail. With the help of a matrix of different

evaluation frameworks, a normative evaluation of different data sources can take place. To the best of my knowledge, such an approach has not yet been enlarged upon in the computer sciences. Hitherto, normatively oriented machine learning research is mainly concerned with fairness (Kearns & Roth, 2020) or preventing discrimination (Hagendorff, 2019c), robustness (Amodei et al. 2017), explainability (Mittelstadt et al., 2019), or preserving privacy (Dwork et al. 2006). Besides that, especially in the field of supervised machine learning, the question of what characterizes—from an ethical perspective—good data contexts remains largely unanswered. This is crucial, since morally sound machine learning applications are in many regards only as sophisticated as their "environmental influences" or training and supervision stimuli. This can be illustrated well with the example of labels in medical machine learning applications. When X-ray images or other outcomes of medical imaging techniques are annotated, this is not done via crowd- or outsourcing labeling procedures, but via expert labeling (Irvin et al., 2019). Why? Because experts are versed enough to provide the right decision-making routines that can help medical machine learning to work properly. So why not extend this tenet of selecting for expert data or for versed individuals and capturing only their cognitive and behavioral abilities in other areas of machine learning? This paper will elaborate on this idea in detail.

Fruitful research can emerge when the social sciences are combined with machine learning research, so that not only ethics, but also technology development can be advanced. Most research works in this area provide critique rather than engage constructively by creating positive ideas and visions on how to use machine learning technologies for the common good. This paper stands in line with and continues the "good data project" (Daly et al. 2019a), promoting tangible good and ethical data practices and frameworks instead of mainly criticizing what goes wrong with machine learning and big data applications. The following chapters elaborate on that in more detail. Chapter 2.1 describes the present approach which is to use as much behavioral data as possible for machine learning development. Although criteria for data quality exist to prefilter training stimuli, these criteria are solely oriented along technical dimensions, not ethical ones, as depicted in chapter 2.2. Chapter 3.1 then describes how human behavior is normatively classified in sociology and psychology, while chapter 3.2 describes how tracking technologies can be inspired by those classifications in order to single out datasets from certain subpopulations that are deemed to be the most competent or morally versed group for a particular task. Chapter 4 then investigates five particular applications of machine learning, namely autonomous cars, language generation, search engine ranking algorithms, social media filtering systems, and e-commerce recommendation systems, that can be made more beneficial by following the presented ideas. To that end, machine behavior objectives, behavioral data sources, tracked states and traits to assess the quality of those data, as well as quality training stimuli of each of the example applications are to be described. Subsequently, chapter 5 covers some points of discussion and responds to them defending the presented approach for beneficial machine learning. Finally, chapter 6 concludes and sums up the paper's arguments.

2 More Data are not Always Better: Defining New Dimensions of Data Quality

2.1 *The Idea of n = all and its Shortcomings*

Before deliberating on data quality dimensions, I want to recapitulate the tenets of big data. Big data meant the emancipation from small data studies, a paradigm scientific knowledge discovery relied on for hundreds of years. Big data lead to the success of today's machine learning systems, which are heralded as the new gold standard of knowledge discovery since they are necessary to understand increasingly complex collections of data, especially in the sciences (Jordan & Mitchell, 2015; Mjolsness & DeCoste, 2001). Broadly speaking, this trend caused some kind of "amnesia" on the value of small data (Kitchin & Lauriault, 2015). While small data are of narrow variety, have a limited volume, are generated to answer specific questions, and produced in controlled ways, big data are the exact opposite. The latter are large in volume since they are generated continuously as a by-product of digital technologies. As stated many times, big data strives to be exhaustive in scope, or, in other words, it follows the ideology of $n = all$. The formula $n = all$ encapsulates the idea that "more trumps better" (Mayer-Schönberger & Cukier, 2013, p. 13; Perrons & McAuley, 2015). Hence, big data are often indifferent towards predefined, specific queries or areas of interest in the context of which one wants to gather insights. Queries often repurpose data to gain insights into phenomena that have no or only indirect linkage to the original context of the data acquisition.

Machine learning techniques allow probabilistic inferences on unknown features. This is why current machine learning applications work under the motto "the more data they have, the better they get" (Domingos, 2015, p. xi). But when speaking about behavioral data, this claim may not be true. It seems that learning applications do not have to be programmed, they program themselves. But when they program themselves while being fed with as much behavioral data as possible in order to aim at higher grades of accuracy, they also become indifferent with respect to the orientation towards certain moral values. The ideology of $n = all$ leads to technical systems that utilize an endless stream of choices made by humans interacting with online platforms and digital devices—a practice once called "laissez-faire data collection" (Jo and Gebru 2019), narrowing down everything towards scores which represent averages of whole populations. But instead of simply recognizing patterns within datasets of a whole population, one could single out datasets—and hence training stimuli—from a certain subpopulation, namely the most competent, eligible, or morally versed one for a respective task, and find patterns only within this data context. Subsequently, only those patterns provide the basis for the generalization ability of a given model or learning algorithm. By diversifying or sampling various data contexts in the larger frame of big data, one can reintroduce ideas connected to the concept of small data in the current situation of an abundance of data. This abundance is so prevalent that measures to tailor training data sets for machine learning applications with

respect to certain fractions of large data sets does not necessarily mean to significantly restrict technical capabilities of these applications. Notwithstanding this, advanced machine learning methods are able to learn from small datasets via data augmentation, can generate synthetic data via GANs or variational autoencoders to artificially increase the amount of training stimuli, use transfer learning to use knowledge from an already learned task, utilize few shot learning mechanisms, etc. (Shorten & Khoshgoftaar, 2019; Zhuang et al. 2020; Mayer et al., 2018). In short, learning algorithms are not reliant on data abundance. This allows for a transition from $n = all$ to $n = X,\ where\ X \subsetneq all$. But the decisive question is what subgroups or fractions to choose, what data nuances to accentuate. In this context, the claim that "more trumps better" is transposed to "better trumps more". But what is better data? To answer this question, one has to look at data quality.

2.2 Data quality Dimensions and Ethics

A common saying in computer sciences is "garbage in, garbage out", referring to the importance of quality data in data-intensive applications like supervised machine learning. Surveys found that data quality attributes comprise literally hundreds of variables (Wang & Strong, 1996). Nevertheless, discourses on data quality define it in terms of its suitability for a business purpose and decision-making efficiency in companies (Samitsch, 2015), and are solely focused on particular technical dimensions like data cleanliness (how many errors do data sets contain?), data completeness (how exhaustive for a particular task are data sets?), data objectivity (what biases do data sets contain?), data consistency and reliability (how many discrepancies are contained in data sets?), data timeliness (how current are the data?), data veracity and exactitude (how accurate and precise are information in data sets?), data interpretability (how readable are data sets?), data cost-effectiveness (what are the costs of data collections?), and the like (Gudivada et al., 2017). Similarly, data quality problems are defined in terms of missing data, duplicate data, inconsistent data formats, incorrect values, spelling errors, etc. (Woodall et al., 2014). Current approaches to improve datasheets for datasets also do not include aspects that go beyond technical and organizational items like "Who funded the creation of the datasets?", "Are there any errors, sources of noise, or redundancies in the dataset?", "Will the datasets be updated?", and so on (Gebru et al., 2018). All those questions and differentiations make perfect sense when assessing data sets that do not contain data that relate to human behavior. But in case data sets relate to it, the discourse on data quality has to be extended.

Besides the data quality discourse, a further discourse addresses the construction of digital persons via data traces from volatile and non-volatile data acquisitions, sensors of all kinds, surveillance measures, social media platforms, and the like. Personal data from different sources and domains are linked together in a form of a "dense rhizomatic assemblage" (Kitchin & Dodge, 2011, p. 90). Terms like "data subjects", "data derivatives" (Amoore, 2011), "data double" (Los, 2006), "shadow order" (Bogard, 1996), "digital persona" (Clarke, 1994), "dividuals" (Deleuze, 1992), or "data doubles" (Lyon, 2003) are used to describe the comprehensive

compilation of personal data, the creation of increasingly detailed and fine-grained digital footprints of individuals, which are then later processed in machine learning applications, which in turn have various (and in some cases negative) ramifications for society (Calvo et al., 2020; Eubanks, 2018; O'Neil, 2016). In this process, various "filters" mediate the translation from an individual's original behavior to eventual computer outputs. Those filters take effect through the selection of certain sensors, data cleansing processes, feature extraction, software libraries, data visualizations, etc. This is why the literature on critical data studies claims that something like "raw data" does not exist (Gitelman, 2013). There are, metaphorically speaking, always bottlenecks, strainers, gates, intentionally or non-intentionally regulating the "permeability" for data at different stages of the computational processing of reality. But those filters do not have an ethical dimension. They do not lead to an ethically motivated selection and sorting out of different data contexts with varying ethical data qualities. To define what I mean by ethical data qualities, one has to analyze how data quality is affected by certain personality traits or modes of behavior of individuals, and how those traits or states can be assessed from an ethical point of view. Eventually, finding quality data shall not primarily serve the pursuit of an improved marketability, but of socially accepted, beneficial machine learning applications.

3 Human Behavior and its Digital Records

3.1 Classifying Human Behavior

Typically, behavioral data are the result of tracking online activities of all kinds, meaning user-generated content, expressed or implicit relations between people, or behavioral traces (Olteanu et al., 2019). Different modes of behavior eventuate in different data contexts. Individuals leave different data traces behind depending on their emotional state, educational background, intelligence, wealth, age, moral maturity, and the like. In order to sort those traits and to classify human behavior and stages of development, one can draw on well-established theories in psychology and sociology. Within the framework of these theories, the aim is to distinguish different modes of behavior or stages of development according to empirical findings. As a general rule, behavior or personality development is understood to be largely a product of one's social environments. Those environments are classified, for instance, with the help of theories of social stratification (Bourdieu, 1984; Erikson et al., 1979; Grusky, 2019; Schulze, 1996; Vester, 2001). A person's milieu, meaning, simplistically speaking, upper, middle, or lower classes, determines their habitus, which in turn determines parts of their behavioral routines and vice versa. Individuals occupy a certain position in "social space" which is the result of a contested distribution of resources, meaning economic, cultural, social, or symbolic capital (Bourdieu, 1989). The position an individual occupies in social space is in large parts "hereditary" and can be affected by social injustices. Nevertheless, the amount of capital a person can concentrate on her- or himself has a structuring power on

many areas of life, meaning that it organizes a person's taste, language, estate, political orientation, or, to say it more generally, his or her dispositions.

Further, these dispositions also structure and have an impact on the way a person uses digital technologies, and influence what kind of data are tracked by these technologies. By using terms like "media-based inequalities", "digital divide" or "digital inequality", several studies show the strong influence a user's socioeconomic status has on media or Internet usage patterns (McCloud et al., 2016; Zillien & Hargittai, 2009; boyd 2012; Hargittai, 2008; Mossberger et al., 2003). Individuals with a higher socioeconomic status are more likely to engage in online activities that enhance their social position, have status-specific interests, interact more frequently with e.g. political or economic news or health information, have higher levels of computer literacy, use less often chat platforms or social networking sites, and so forth. All in all, the position of an individual in social space heavily influences his or her ways of using digital technologies and hence the kind of behavioral data that are digitally recorded—with the respective consequences for biases, scopes, representative statuses, or ethical quality dimensions of data sets.

While behavior is in many respects an outcome of the respective social environment, class, milieu, or social position, the same holds true for personality development, which is widely dependent on the circumstances of socialization. Developmental psychological theories either postulate that personality development follows a continuous process or that it passes through discontinuous stages, where logical reasoning is learned, moral senses are developed, social norms are adopted, emotional intelligence is acquired, stereotypes are negotiated, role models are changed, self-reflection is learned, values are internalized, personal crises are overcome, and the like (Erikson, 1980; Kohlberg et al., 1983; Loevinger, 1997). In order for an individual's socialization to succeed, it requires, among other things, a certain range of beneficial influences from a social environment, which can be separated from harmful influences. To scrutinize these influences is the objective of developmental psychology. The discipline focuses on long-term progressions with regard to the experiences and the behavior of individuals in order to find patterns and regularities that are crucial for the development of intellectually and emotionally sound and mature individuals (Lerner 2015). A succeeding development is measured by aspects such as problem-solving abilities, emotional intelligence, cognitive development, prosocial behavior, mental health, educational success, etc. As soon as such norms for a successful personal development are defined, one can roughly differentiate between positive and negative environmental influences. The latter can affect health, gross and fine motor skills, socio-emotional development, the speed of information processing, self-concepts, knowledge, or language behavior and range from alcohol to stress during pregnancy, residential areas with high crime rates, low educational levels, emotional, physical or sexual abuse, as well as a neglectful parenting style (Sullivan and Knutson 2000; Spera 2005).

According to theories from developmental psychology, "higher" forms of personality development lead to other behavior patterns than "lower" ones (Hart et al., 1997; Paul B. Baltes et al. 1978; Kohlberg et al., 1983). Normally, more cognitive-moral growth leads to more socially desirable or acceptable behavior. Philosophical theories about ideal moral acting, ranging from Kant's categorical imperative (Kant,

1977), Habermas' discursive will-formation (Habermas, 1987), or Rawls' theory of social contract (Rawls, 1999), imply that individuals possess fully developed cognitive capacities. In this context, one can assume that personality or character development may strive towards the target values and rationality standards of these models. In order to measure the "proximity" of a person's character to certain target values, differential psychologists have developed various tools for personality assessment to understand and predict behavior in different social contexts. Amongst personality assessment tools are the widely used Five-Factor Model (John et al., 2008; McCrae & John, 1992), non-scientific tests like the Myers-Briggs type indicator (Myers and Myers 1995), or less known methods like the Multidimensional Personality Questionnaire (Tellegen & Waller, 2008), "PerformanSe" (Patel, 2006), and many more. All these tools have specific weaknesses, they ignore the fact that personality can be in a state of flux, and that it may be unclear what personality characteristics mean in terms of behavioral manifestations in certain situations. But apart from that, they more or less reliably measure traits like motivation, extraversion, emotional stability, openness, conformism, rationality, impulsivity, dynamism, anxieties, social activity, and the like.

The mentioned theories and tools have a tacit consensus about certain ethical target values or "attraction poles" (Sloterdijk, 2009). This can be exemplarily elucidated with regard to the Five-Factor Model (John et al., 2008). All five personality dimensions have an attraction pole, meaning that all dimensions can be spanned between two poles whereas one pole is designated as the favored one. Typically, more complexity in an individual's mental and experiential life is better than less (openness), more impulse control that facilitates goal-directed behavior is better than less (conscientiousness), more social activity and positive emotionality is better than less (extraversion), a more prosocial and communal orientation is better than less (agreeableness), and more emotional stability is better than less (neuroticism). In the background of personality assessment tools, developmental psychological, or social milieu theories, tacit normative presuppositions exist that structure attraction poles of all kinds. However, making these presuppositions and polarizations explicit may not be equated with attempts to classify humans as such. The mere idea of classifying humans provokes strong moral intuitions to refuse such practices. But besides these moral intuitions, the application of classification or scoring systems on humans is common industry and government practice in many countries (Engelmann et al., 2019). People are classified with respect to their financial situation, their social reputation, their risks of conducting certain actions, their personality, etc. Moreover, they are classified along geodemographic segmentations, purchasing histories, lifestyle types, and the like. However, the circumstance that the application of ranking systems on people corresponds to the status quo of the digital economy does not mean that the related practices are morally correct. Quite the opposite may be true (Zuboff, 2015). In the context of this paper, though, I do not propose to apply digital ranking practices to individuals as such, but to particular types of behavior or particular personality traits. They can be measured alongside certain dimensions, which have a more or less strong normative alignment. Hence, in the following chapter, I want to elaborate on how personality traits and different types of human behavior can be digitally measured and classified.

3.2 Tracking Human–Computer Interaction

Dataveillance (Clarke, 1988; van Dijck, 2014), in other words practices of recording and analyzing digitally mediated behavior, has at least three complications or downsides. First, it is a morally contested practice, causing negative "chilling effects" of all kinds (Schneier, 2015). Second, monitoring human–computer interactions or online behavior does not yield data that corresponds to real attributes but it constructs them (Haggerty & Ericson, 2000). And third, one can only infer personal attributes with the right data bases. Obviously, missing data strictly limits the scope of information one can gain, albeit the possibility of statistically inferring information on unknown features exists. Before these three complications or downsides are discussed in more detail in chapter 5, it is important to see that in practice, various tools are used to track personal traits and states (Matz & Netzer, 2017). Single- as well as multimodal approaches that combine several psychological attribute recognition methods, and that can detect involuntary (e.g. physiological), semi-voluntary (e.g. facial expressions), as well as voluntary (e.g. key presses) signals are used (D'Mello & Kory, 2015; Zeng et al., 2009).

Some instances are listed hereafter: By analyzing clickstreams, browsing histories, or search queries, inferences on users' demographic information can be made (Acar et al., 2014; Bi et al., 2013; Hu et al., 2007). Affective computing serves to detect emotions, mostly through text, voice, face, or posture processing techniques (Picard, 1997). Written text can be investigated in order to detect mental illnesses, to conduct sentiment or personality analysis via differential language analysis, natural language processing, and other machine learning methods (Guntuku et al., 2017; Pang & Lee, 2008; Schwartz et al., 2013). Moreover, digital images, for instance social media profile pictures, can also be used to reveal personality attributes (Segalin et al., 2017). Various sensors—especially the ones in smartphones and other wearable devices—are used to track physiological signals, movements, activity levels, mobility patterns, face-to-face encounters, and the like in order to infer internal states and personal attributes (Harari et al., 2017; Kwapisz et al., 2011). User input via display touching behavior, mouse movements, or keyboard strokes can also be used to infer personality traits (Khan et al., 2008). Many other applications could be added.

All in all, the baseline is clear. Tracking technologies for digitally mediated behavior can in many cases successfully measure a broad spectrum of psychological traits, affective states, and personal attributes. Many tracking applications specifically aim at measuring the six basic emotions (sadness, fear, anger, disgust, joy, surprise), while in practice, though, these basic emotions can be observed only seldom, instead engagement, confusion, boredom, curiosity, frustration, and happiness are more frequent affective states in human–computer-interaction (D'Mello, 2013). But tracking technologies can also measure more complex attributes like age, gender, sexual orientation, occupation, mean income, ethnicity, religious views, political attitudes, personality traits, intelligence, pregnancies, use of addictive substances, job performance, parental separation, and many more (Kosinski et al., 2013, 2014). Here, the connection to psychological as well as sociological approaches to classify human behavior can be made. While many classical theories from the humanities

approach social structures and individual-related traits and distinctions on a very high and abstract level, digital behavior tracking technologies only capture "microscopic" behavior traces in data sets. But the former and the latter can be combined in order to transition from the $n = all$ ideology to $n = X,\ where\ X \subsetneq all$, which means to single out datasets from certain subpopulations that are deemed to be the most competent or morally versed group for a particular task.

Tracking technologies, even if they are themselves ethically contested, can be the bedrock for an ethically motivated selection of different data contexts with varying ethical data qualities that can then lead to beneficial machine learning applications and machine behavior. When recognizing that a person's dispositions structure and influence the way she or he uses digital technologies, then methods to detect digitally mediated behavior can, in turn, infer those dispositions when analyzing data traces. That means that "higher" states of personality or moral development, socially desirable or acceptable behavior, distinct cognitive abilities, emotional stability, rationality standards or, in general, the "proximity" of a person's dispositions to certain socially accepted and ethically defined target values can be measured. Problem-solving abilities, emotional intelligence, cognitive development, prosocial behavior, educational status, mental health—all those assessment dimensions have attraction poles that are used in many social contexts to rank human behavior and to assess whether particular individuals can be put in charge, or are competent enough or eligible for certain tasks. This principle is, at least when adopting a meritocratic perspective (Young, 1994), effective in many social institutions. From here on, I want to make the transition to prefiltering training stimuli for machine learning applications according to certain individual attributes, states, and traits.

4 Beneficial Machine Learning: Putting the Approach into Practice

In the context of (supervised) machine learning development, there are three ways in which hereditary and "environmental" information can be inscribed into algorithms (Rahwan et al., 2019, p. 480): they can be incorporated into applications by programmers making design choices in algorithms (Brey, 2010; Friedman and Nissenbaum 1996), by particular training stimuli (i.e. data), or by a machine's own "experiences". Taking up the humanistic differentiation between hereditary factors and environmental influences that shape an individual, one can stress that machine learning applications also combine both; the former through algorithm design, and the latter through training stimuli, where both factors interfere with each other. Training stimuli, in other words a set of examples used for learning, are used to fit and tune the architecture, parameters, or weights of a classifier. Training data sets are supposed to allow artificial neural networks to generalize from the sample of the training data set to potentially every other case, meaning that the network has the best possible performance on any new data. In this context, training stimuli must be distinguished from validation and test or holdout sets, where the former serve the purpose of tuning the architecture of a classifier, and the latter of measuring the performance of a trained classifier. Supervised machine learning predicts a categorical or continuous value Y in the form of target variables or labels given an input X

in the form of a set of variables by training a function F, where $F(X) = Y$ (Ghani & Schierholz, 2017). Here, the set of variables can represent behavioral data, but one also has to keep in mind that the same holds true for labels. Labels are most often the result of manual clickwork (Irani, 2016), but they can have the form of genuine behavioral data, too. For instance, this is the case when labels for video data of an autonomous vehicle's surroundings are generated by capturing the driver's behavior (Tsutsui et al., 2018). This way, human behavior also becomes machine behavior via labels, not just via training data itself.

Concepts that follow the idea of an ethically motivated selection or limitation of particular training data or labels in order to influence the process of developing certain machine behaviors are non-existent up to this point. One single exception is Davidow and Malone's cursory concept of "starving AI" or of "putting artificial intelligence on a data diet" (Davidow & Malone, 2020). The idea is to ensure trustworthy artificial intelligence not by controlling it, but by putting it in "virtual prisons", meaning that the applications are disallowed to use whatever training stimuli they can get for learning. The authors follow a rather metaphorical approach, but, in a nutshell, they rudimentarily capture an idea similar to the one presented in this paper, namely developing beneficial machine learning applications by filtering training stimuli according to ethical considerations. My line of argument starts at the assumption that a person's social background, educational level, personality, intelligence, etc. shape his or her way of using digital devices. Moreover, these devices are equipped with sophisticated tracking technologies that can in many cases accurately measure and infer the user's personal attributes, traits, or states. Depending on these measurements, data traces the respective user produces, that is data traces that provided the basis to the measurements itself as well as data traces that are situated in the same context as the measurement, are assessed from an ethical perspective. Thereafter, this assessment enables data scientists to relinquish the idea of using as much relevant data as possible that represent averages of whole populations. Instead, they single out quality data that are representing behavior of specific subpopulations which are deemed to be especially competent, eligible, or morally versed for a particular task. This limitation, that stands in contrast to the credo that the bigger the data the better the machine learning models, serves to tailor training data in a way that machine behavior can be steered into a direction that promotes its beneficence.

4.1 Concrete use Cases

To further elaborate on that, I want to sketch out use cases that exemplarily illustrate the process that is outlined above. For that end, I investigated five machine learning driven applications and demonstrate how beneficial machine behavior objectives can be achieved by selecting certain quality data contexts for model training. The investigation shall delineate how the paper's ideas can be put from theory into practice. I will focus only on applications that are widely used, like e-commerce recommendation systems, search engine ranking algorithms, or autopilots in self-driving cars, and describe how these applications can be amended by following a stringent approach for quality training data selection along particular ethical considerations.

4.1.1 Self-Driving Cars

4.1.1.1 Machine Behavior Objectives Autonomous vehicles are supposed to guarantee as much safety as possible (Koopman & Wagner, 2017). Avoiding crashes with self-driving cars (Xu et al., 2019) is paramount to advance their deployment. Achieving this goal has many dimensions, but it certainly encompasses safe machine behavior, meaning a car complies with safe overtaking maneuver, following, emergency stop, cornering, or line choice rules.

4.1.1.2 Data Sources Autonomous cars must infer from past traffic situations to new ones. Thus, training data, meaning video recordings and further sensor data of all kind representing countless hours of driving, as well as annotations for these data are of utmost importance. Because it is rather expensive and in some cases notoriously difficult to acquire enough annotation data, in some autonomous vehicles, label collection happens via measuring behavioral cues from human driving behavior, e.g. acceleration, deceleration, steering, etc., in manual mode or during autopilot disagreement (Eady, 2019). The labels are then linked to the respective footage of the vehicle's surroundings. Additionally, data traces from actual driving can be combined with customer data as well as further behavioral data from third party organizations like data-brokers.

4.1.1.3 Tracked States and Traits There are certain individual characteristics like gender, age, driving experience, distraction, attention, reaction time, visual function, sensation seeking, impulsivity, etc. that predict risky driving behavior (Anstey et al., 2005; Fergusson et al., 2003; Wayne & Miller, 2018). According to accident statistics and empirical investigations, individuals who cause fatal as well as non-fatal car crashes tend to be male, of young age, have high levels of aggressiveness, sensation seeking, and impulsivity as well as some other traits like lower levels of income, poor mental health status, higher levels of neuroticism, possibly raised blood alcohol concentration, lower driving experience, and show various forms of antisocial behavior or higher levels of social deviance (Abdoli et al., 2015; Čubranić-Dobrodolac et al., 2017; Hyman, 1968; Vaughn et al., 2011; Wang et al., 2019; West & Hall, 1997). Many, if not all of these traits can be digitally detected at some degree of accuracy. Those characteristics as well as additional cues like engine speed, pedal pressure, improper following, speaker volume, driver body posture, gestures, head movement, verbal outbursts, etc. can be digitally tracked in order to predict a driver's safety level (van Ly et al., 2013).

4.1.1.4 Quality Data Contexts As soon as the above-mentioned traits are digitally tracked and recorded, the driving behavior data that is related to the respective driver can be excluded or downgraded from the data set that is used to train the models that determine the machine behavior during autopilot. Traffic psychologists can help machine learning practitioners to further establish tools to classify data that represent decent driving behavior. In short, quality data contexts arise from drivers who possess decent driving experience, have a good reaction

time, tend to be female, have low levels of aggressiveness, sensation seeking and impulsivity, show active head movement in traffic, distinguish oneself in few or no verbal outbursts, proper following behavior, or modest acceleration behavior, to name just a few attributes.

4.1.2 Language Generation

4.1.2.1 Machine Behavior Objectives Chatbots as well as speech assistants of all kinds are supposed to produce appropriate, sufficiently eloquent language that does not violate social norms, discriminate against certain groups of people or perpetuate biases that are incorporated into training data (all of which is especially precarious in open domain conversations) (Köbis et al., 2021; West et al., 2019; Danaher, 2018; Silvervarg et al., 2012; Sheng et al., 2019; Bolukbasi et al. 2016).

4.1.2.2 Data Sources Natural language generation is based on finding statistical patterns in text corpuses (Solaiman et al., 2019), which then allows a machine learning model, among other things, to predict the next word in a sentence based on previous words. To learn those patterns, the chosen text corpuses can be digitized books, forum posts, news articles, communication data, Wikipedia articles, websites, blogs, scientific papers, and many more.

4.1.2.3 Tracked States and Traits States and traits that can be tracked in order to assess text data quality may range from an author's educational background or occupation, intelligence, the characteristics of his or her keyboard strokes or display touching behavior (backspacing etc.), the time between writing and posting, and in particular by assessing the used publication platform, filtering intermediates, review processes, and the language skills themselves.

4.1.2.4 Quality Data Contexts Especially text data that is not produced by professionals, meaning journalists, writers, scientists, etc., but by lay persons is expected to be of lower quality. Text data that is not editorially controlled and therefore did not undergo any kind of review or filtering intermediate may be interspersed with orthographic mistakes, poor syntax, smaller word pools, slang, invectives, strong biases, etc. Quality data contexts are to be assessed in dependence on the respective purpose of an application for natural language generation. Texts from the public domain may be suited to improve a chatbot's realism, hence its ability to produce convincing, authentic, and human-like everyday language. On the other hand, these texts can be infiltrated with aggressive, discriminatory, or offensive phrases (Wolf et al., 2017). To avoid these and other pitfalls, the selection of text corpuses that are used to train conversational robots should not follow the bigger-is-better-approach like many commercially developed chatbots do (Bender et al., 2021). Instead, the selection of corpuses can be narrowed down to digital writings that underwent a firm quality check through publishers, peer reviews, or media agencies, that is embedded in a sophisticated web of citations or links, or that stem from individuals with high levels of language skills. Moreover, language proficiency can be determined by assessing the

structure, continuity, errors, vocabulary richness, length of sentence, changes made to text, etc.

4.1.3 Search Engines

4.1.3.1 Machine Behavior Objectives Modern search engines like Google Search, Bing, Yandex, etc. use a plethora of signals to rank search results, make autocomplete suggestions, predict users' intentions, evaluate websites, and so on. The main machine behavior objective is to ensure that rankings and content fit to the anticipated needs of the users. This, in turn, is supposed to cause a lock-in-effect and bind users to the respective search engine, eventually raising the likelihood of contact with advertisements. Despite this well-established machine behavior objective, one can name several other objectives that could determine the architecture of search algorithms like content quality, expertise, and trustworthiness (in general search engines), equal opportunity (in people search engines), sustainability (in product search engines), and many more.

4.1.3.2 Data Sources Search engines use diverse tracking techniques, harnessing the large amount of different human–computer-interactions. Each list of search results shown to users nudges them to become behavioral data contributors for further calibration and model training by clicking on links, mousing over items, using the back button, scrolling through pages, entering terms in search bars, interacting with ads, spending time on a page, and many more. Besides such behavioral data, search engines can analyze main and supplementary contents of websites, the amount of internal and backlinks to a website, labels, page load speed, aggregated views, end-user device specifications, duplicates, and so on.

4.1.3.3 Tracked States and Traits Many of today's relevant search engines are embedded in broader online platforms that allow for a comprehensive user classification. By collecting and analyzing data on search terms, visited websites, clicked ads, user location, keyboard strokes, mouse movements, interaction speed, product or profile views, and the like, it becomes possible to probabilistically infer a variety of different personal states and traits. Among them are a person's gender, age, occupation, residence, religion, political views, favorite brands, personality, intelligence, literacy, and many more (Bi et al., 2013). These states and traits can then be used to assess a user's "signal quality".

4.1.3.4 Quality Data Contexts Professional general search engines do have page quality rating systems in place (Underwood, 2015). They are used to recognize the purpose of a website. Beneficial pages that are supposed to help users and are created by individuals with high expertise, authoritativeness, and trustworthiness receive the highest ranking. Pages that contain hate or misinformation, encourage harm to others, or have a deceptive intent receive the lowest rating. However, these page quality ratings do not solely determine the search results. They are accompanied by machine learning techniques that "digest" user behavior in order to re-train

the search algorithm. This user behavior can also be assigned to varying "quality" stages. The clickstream habits of a person who, for instance, regularly uses politically extremist search terms, visits websites of low quality, has numerous typos, etc. should be less considered for shaping the search algorithm. On the other hand, clickstream habits that give evidence of ethically desired traits could preferably be used to optimize ranking algorithms.

4.1.4 Social Media

4.1.4.1 Machine Behavior Objectives Recommendation systems on social media platforms come in all shapes and sizes. They are used to filter posts, friends, images, videos, music, news, search results, and many more. Hitherto, the main goal of these systems is to increase user engagement in order to bind them to the respective platform. This, in turn, shall raise the likelihood of advertisement contact and click-through-rates (Eyal & Hoover, 2014; Hagendorff, 2019b; Kuss & Griffiths, 2017). Taking social responsibility seriously, platforms could rearrange their objectives towards values of a vital and fair public discourse, truth, and information quality. This means to change the methods for algorithmic measurement and determination of information relevance. Fake news, hate speech, extremist content, etc. may cause the strongest user engagement, but the engagement quantity should not determine the subsequent dissemination and recommendation of the respective content. Instead, engagement quality should determine data quality and help to build responsible machine recommendation behavior.

4.1.4.2 Data Sources Social media platforms can track a plethora of user signals. Amongst the more obvious ones are clickstreams, search queries, demographic or profile information, reactions to posts, duration of post views, scroll behavior, networks of friends, comments, and many more. All these data traces are used to determine the relevance of posts, videos, images, tweets, friend suggestions, etc. in order to operate the platforms' recommendation systems.

4.1.4.3 Tracked States and Traits Tracing back to a differentiation from behavioral economics (Kahneman, 2012), one can distinguish system-1- and system-2-interactions. System-1 comprises fast, emotional, effortless, cognitively simple thinking processes that are prone to biases and mistakes, whereas system-2 covers slow, rational, and deliberate thinking processes. Those two modes of thinking do also influence the way digital platforms are used (Lischka & Stöcker, 2017). Amongst other factors, social media platforms could measure whether users operate with a platform on a more irrational, bias prone, impulsive system-1 mode. This mode allows for rather quick and impulsive actions, resulting in a stream of unreflected human–computer-interactions. Impulsive, system-1 user behavior could be tracked by things like reaction or comment speed, the susceptibility to nudging techniques, and scrolling or reading behavior. The platforms could also use further inferences to educational levels, intelligence, psychological traits and states like anxieties or

radical political or religious views to assess user behavior that is connected to the respective attributes.

4.1.4.4 Quality Data Contexts Cognitive heuristics that are part of system-1-interactions influence the way individuals interact with social media. Hence, biases are technically perpetuated via recommendation systems (Stieglitz & Dang-Xuan, 2012, 2013). Quality data can thus be scraped from contexts where user generated data do mainly represent system-2-human–computer-interactions. This way, recommendation systems can be trained on behavioral data that represents fewer biases and impulsive reactions. Instead of negatively affecting public discourse by helping the spreading of content that is mostly suited to cause emotional arousal and impulsive reactions, platforms help to automatically disseminate content that is less "toxic" for public discourse. In this context, a special focus can be laid on so-called "superusers", who are not just very active users with high levels of engagement, but who also disproportionately spread misinformation. According to one rare source, internal committees at Facebook urged to lower recommendation scores for content posted by "superusers" on the far right or far left of the political spectrum. Content from moderate users, in turn, would receive higher scores (Ng, 2020). This advice was turned down by Facebook's leadership. However, it perfectly corresponds to the idea that is advocated here. When recommendation systems are trained on behavioral data from individuals with higher education, who do not represent political or religious extremes, and who normally interact with quality, i.e. journalistically or scientifically verified, trustworthy information, it is to be expected that those systems automatically spread content that comes with various benefits for the public, instead of harms, leading to a situation where everyone is better off.

4.1.5 Online Shopping

4.1.5.1 Machine Behavior Objectives E-commerce platforms where people can buy goods and services use various methods to promote purchasing behavior. They use shopping search engines, product recommendation systems, product reviews, dynamic pricing, cross selling, customer analytics tools, conversion rate optimization, conversion funnels, varying payment options, specific user interface designs, etc. in order to maximize revenues. Therein lies the entrenched main machine behavior objective. However, via tweaking the underlying machine learning algorithms, the machine behavior objective can be diversified, comprising not just the pursuit of economic values, but also values of sustainability or public health. Especially the biggest online stores like Amazon, eBay, Walmart, Jingdong, Alibaba, and others would cause a significant impact by just slightly changing their machine learning models towards the mentioned values, taking their corporate social responsibility seriously.

4.1.5.2 Data Sources No different than search engines or social media platforms, online retailers can collect a broad variety of user data. They can analyze and track

the number of transactions, all kinds of product and customer data, the conversion rate, product impressions, average order values, product detail views, adding or removing of products from the shopping basket, withdrawals from checkout process, customer lifetime values, traffic sources, details of users' devices, and many more.

4.1.5.3 Tracked States and Traits Using the many data sources e-commerce platforms can gather as a basis, they can implement specific automated mechanisms for customer segmentation. Typically, differentiations for types of customers are made purely from a sales perspective, distinguishing between loyal, impulsive, novice, etc. customers. Notwithstanding that, customers can be segmented along criteria like health- or eco-consciousness by analyzing their product views, shopping behavior, product reviews, search terms, personality, socio-demographic factors, and the like.

4.1.5.4 Quality Data Contexts Sticking to the aforementioned values of sustainability and public health, e-commerce platforms could use all available data from health- as well as eco-conscious customers and use specifically those data to train models for product recommendations, dynamic pricing, or ranking algorithms for their search engines, to name just three major setting options. Such measures could significantly foster the extent to which e-commerce platforms promote more sustainable and healthier consumer behavior.

5 Discussion

In the following section, I want to gather some major points for discussion to the suggested approach to achieve beneficial machine learning applications. The approach relies on several practices like tracking, profiling, ranking, or filtering that have been applied in contexts of technology misuse for illegitimate ends more than just a few times (Brundage et al. 2018; Crawford et al. 2019; Crawford, 2021; Engelmann et al., 2019). Nevertheless, the mentioned methods cannot be simply discarded. Rather, they can be used for purposes that are in line with the common good and ethical values. The following subchapters will elaborate on that and take up the seven most pressing concerns that can be raised when putting the paper's ideas into practice.

5.1 Ranking Behavior

First of all, I want to address some ethical concerns that are connected with the idea of ranking human behavior, as described in chapter 3. As a cautionary tale how such ranking practices fail one could point at digital marketing, where marketers consider customers to be either "targets" or "waste" (Turow, 2012). People's tastes, demographic profiles, beliefs, income levels, and many more are digitally measured for the purpose of discriminating them for commercial goals and to find the customers

that are deemed most valuable. Reputation silos are constructed around people who statistically seem similar, but this practice often circumvents ethoses around equal opportunity, justice, or transparency. However, the idea of ranking or sorting human behavior along certain dimensions or hierarchies is not to be refused altogether. One can put the concept in another light when reframing it according to assessment criteria from ethics. Hence, besides arguing for a repurposing of existing social sorting structures in marketing for other socially accepted ends, one can in general say that society embraces various practices of behavior assessment under the term "ethics". Here, one transitions from saying that some social positions, dispositions, or stages of development are better than others to classifying behavior as morally right or wrong. In the context of this paper, though, the argument is somewhat more specific: Currently, digital tracking systems that measure users' social as well as psychological traits and states are used mainly to support marketing decisions, to foster customer relationship management, to personalize the marketing mix to individuals, and to support many other commercial purposes (Wedel & Kannan, 2016). I argue that the rich toolset that is already established for marketing purposes can be repurposed to assess the ethical quality of data contexts in order to develop beneficial machine learning applications. Currently, tracking methods are not deployed to evaluate digital behavior from an ethical point of view, which would be essential to assess ethical data quality dimensions. With that said, this assessment does not necessarily require in-depth knowledge about ethical theories that are developed in philosophic discourses. According to the intuitionist approach in moral psychology, ethical judgements are driven primarily by one's intuitions (Haidt, 2001). In view of that and apart from complicated, dilemmatic cases, moral intuitions can appropriately guide ethical reasoning. This becomes evident when considering the tacit consensus about ethical target values or "attraction poles" that are embedded into sociological as well as psychological theories. Normative assumptions about the value of prosocial dispositions, rationality, moral development, openness, impulse control, positive emotionality, and the like are seldomly contested. Hence, these intuitive normative assumptions can in many cases guide data quality evaluations. This shall not downplay the potential complexities and conflicts that can be connected to such an evaluation or to AI technology assessments per se (Raji et al. 2020; Suresh & Guttag, 2020). But as a first step, moral intuitions might be a good "compass" to guide decisions on which features in data sets should be weighted stronger than others in order to promote beneficial model training.

5.2 Paternalism

Another potential objection to the ideas presented here is that they depict a form of ethical paternalism. The decision about what defines quality data contexts is made by machine learning practitioners and other members off the respective coalition of stakeholders that become involved in deciding which values and normative assumptions are to be implemented into technology. The decisions made in this coalition affect other collectives without their democratic consent. The idea of paternalistic thinking is that some individuals are more competent, rational, or versed than

others and that the former can decide for the latter for their advantage. Paternalistic approaches are mainly criticized because they not only limit the freedom of affected individuals, but this is also done without their consent (Sartorius, 1983). In this context, two main counterarguments can be raised. First, the idea of developing beneficial supervised machine learning applications follows, as the term suggests, the notion of being beneficial or advantageous for as many individuals as possible, hence promoting the common good. Second, machine learning applications do in nearly all cases affect individuals without their explicit consent or knowledge. Values are part of technologies itself (Brey, 2010), and the process of embedding certain values or choosing architectural designs is in many regards not a democratic one. Rather, a small group of technology developers possess the power to make far reaching decisions for a group of end users that can comprise millions or billions of individuals (Lessig, 2006). The decisive question is whether values are embedded in software in an arbitrary and nonreflective way, often perpetuating prejudice, biases, or misunderstandings, or whether these values and ethical norms are chosen carefully and consciously. Here, I opt for the latter. In addition to that, technology ethics indeed typically operates with paternalistic, top-down norms and standards. Asimov's Three Laws for Robots, for instance, is presumably the most well-known approach (Asimov, 2004). However, in the context of this paper, the paternalistic top-down approach is not solely embraced but combined with one that is bottom-up. Bottom-up approaches mean that a technical agent, in this case machine learning models, explores courses of action that represent morally praiseworthy examples (Wallach & Allen, 2009). Hence, the agent achieves "moral capabilities" by surveying its environment, similar to childhood development. Here, both top-down and bottom-up approaches are integrated. A top-down analysis is done with regard to the way training data is filtered or selected according to ethical criteria, while on the other hand a developmental or bottom-up approach is chosen in order to allow the machine to learn a certain behavior from data sets.

5.3 Transparency

Another objection, that is akin to the one above, is to stress that the proposed approach for beneficial machine learning applications is non-transparent, leaving affected individuals unwitting about the technical measures that are conditioning their user experience and filtering mechanisms for quality data. The counterargument to this objection is that platforms or software developers could and should have no problems whatsoever making transparency statements, thereby informing potential reviewers about the value and design choices, ultimately making them subject to public scrutiny. This would show that definitions about ethical quality dimensions of training data are in line with cultural consensuses, ethical theories, and moral intuitions. Revisiting the above-mentioned examples from chapter 4.1, the majority of people would consent to the claim that autonomous cars should be safe, that natural language generation should avoid simplicity and perform eloquently, that e-commerce platforms should promote sustainable shopping routines, that recommendation systems on social media platforms as well as search engines should not

foster political extremism and information that is "toxic" to the public discourse but promote quality content, expertise, and truth. In general, it should be a necessary prerequisite that through transparency statements the criteria by which data quality is evaluated are described and justified when putting technologies for digital behavior tracking as well as machine learning models that are trained with behavioral data in place.

5.4 Privacy

In order to select for quality data or data that represents ethical behavior, various tracking or surveillance techniques must be in place. This leads to a further objection, namely that the proposed approach for beneficial machine learning application goes hand in hand with privacy violations. This criticism requires a more detailed examination. First of all, although this might be a weak argument, the proposed approach does not necessarily opt for an extension of methods for recording behavioral data that are already entrenched (D'Mello & Kory, 2015; Zeng et al., 2009). Here, it is important that these methods are used for legitimate ends, not that they are abolished (Belliger & Krieger, 2018; Hagendorff, 2019a). As already discussed in a previous chapter, the idea of classifying people or people's behavior raises weighty ethical questions and has its ailments, especially with regard to the feasibility of data-driven assessments of sensitive traits like mental illnesses, intelligence, personality, and the like. However, data protecting measures that would prohibit these assessments like the one mentioned are primarily aiming at preventing unjust discrimination and at securing personal autonomy (Roßnagel, 2007). In the end, though, is it important to remember that from the pure existence of these technologies alone it does not necessarily follow that they are misused. On the contrary, when binding legal norms as well as strong ethical tenets are entrenched, tracking and profiling can be used for the common good, as it is described here.

Moreover, techniques for tracking user behavior and assessing behavioral data quality can work by only using anonymized and aggregated data, avoiding opportunities to identify certain individuals. Although current trends in informational privacy research are extending the idea of privacy violations from a merely individualistic data protection perspective (Westin, 1967) towards notions of interdependent, group, collective, or predictive privacy (Biczók & Chia, 2013; Mühlhoff, 2021; Yu & Grossklags, 2016), the main "type" of privacy violation still remains in the form of a direct access or probabilistic inference to sensitive information that are tied to a particular person. But this "classical" form of privacy violation can be avoided by anonymization techniques, mainly differential privacy or k-anonymization (Dwork et al. 2006; Dwork 2008; Dwork & Roth, 2013; Samarati & Sweeney, 1998). These techniques imply to alter datasets by adding noise or by manipulating them in a way that individuals theoretically cannot be identified. That does not necessarily provide full privacy or full protection against re-identification, though. Moreover, making datasets more privacy preserving comes at a price, namely accuracy. On the flipside, implementing differential privacy can lead to access to new data sources that hitherto could not be collected due to privacy concerns (Kearns & Roth, 2020, p. 56).

In this way, privacy enhancing measurements can actually lead to an increase in the amount of sensitive behavioral data, which can then be used to assess training data quality from an ethical perspective. Eventually, the proposed approach, which is indeed very data-intensive, can go hand in hand with individual privacy preserving measurements.

5.5 Accuracy

Another objection is that techniques to digitally assess and rate human behavior may be inaccurate and create false positives and negatives. These techniques as well as data traces per se construct rather than represent an individual's true actions, traits, and states. Accordingly, behavioral data as well as the computational processing thereof cannot be condensed in information that represents "reality". Rather, different "realities" can be constructed from data and algorithms (Lewis, 2015; Matzner, 2016). They do not work impassive, but shape how we understand the world in a performative manner (Kitchin, 2017), while allowing probabilistic inferences on in situ behavior. In individual cases, this can lead to detrimental false positives or negatives. But the methods proposed in this paper all operate with aggregated and, in the best case, anonymized data, which means that individuals face unjust technological consequences only when tracking and profiling techniques fail significantly. Only in the unlikely event they come up with misclassifications in an overwhelming number of cases, the selection of quality training data and therefore the trained models would become skewed.

5.6 Discrimination

Akin to the aforementioned objection is the argument against algorithmic discrimination. The ideas presented in this paper could fall prey to such arguments since specific biases in data sets are intendedly promoted, resulting in "skewed" algorithmic decision making. Previous discourses on algorithmic discrimination rightly criticize that machine learning techniques perpetuate existing biases that are entrenched in data sets and therefore foster unfair discrimination. Under the umbrella term "fairness, accountability, and transparency in machine learning" (FAT ML), machine learning practitioners collect methods for reducing algorithmic discrimination primarily by dealing with protected attributes like gender, age, ethnicity, etc. (Dwork et al. 2011; Kleinberg et al., 2016; Veale & Binns, 2017). However, here I argue that one should reintroduce or promote "algorithmic discrimination", but, needless to say, not in the traditional way. Data sets may contain features that are critical in a way that they should be weighted stronger than others. This can be the case even if these features perpetuate biases, since biases can actually be necessary for fairness. Dutta et al. (2020) give the example of a hiring for fire fighters, where candidates should be able to lift heavy weights, which leads to a preference for men rather than women. In short, biases are acceptable if they are critical for the legitimate solution of a given task. Here, I propose to promote biases in data sets used to train machine learning models that lead to a preferability of features that are desirable from an

ethical point of view. At the same time, however, this means to put individuals at a "disadvantage" who produce behavioral data that originates in deeds, personal traits, or mental states that are socially less esteemed like risky behavior, detrimental norm violations, bad language, low education, political or religious extremism, flawed logical reasoning, impulsiveness, and the like. This way, machine behavior that results from recognizing statistical patterns in behavioral data is not "socialized" by general populations ($n = all$) but by specific subgroups ($n = X$, *where* $X \subsetneq all$) that comprise individuals who are most competent, eligible, or morally versed for a particular task. While the typical notion of algorithmic discrimination is pointing towards unfair computational outputs, the kind of algorithmic discrimination that is proposed here aims at introducing stricter filters that thwart particular data traces to become training data for machine learning. This way computational outputs manifest values that correspond to ethical virtues and that are socially accepted, appreciated, and sought-after like friendliness, literacy, truthfulness, positive emotionality, prosocial orientations, etc. Selecting for these values is a task that comes with a heavy responsibility. That is why it should not only be put on the shoulders of machine learning practitioners. Such a task can best be addressed when building a coalition of stakeholders, when combining technical research with the social sciences, psychology, domain-specific applied ethics, standard setting bodies, etc. As a rule of thumb, the extend and inclusiveness of this coalition correlates with the level of criticality of a particular application of algorithmic decision making (Wendehorst et al., 2019). By following such an integrated approach, unintended cognitive biases in the selection off intended training data biases can be tracked, but it can also be ensured that the selection of values and traits stands in accordance with democratic consensuses (Floridi et al., 2018; Brundage et al. 2020; Daly et al. 2019b; Danaher et al. 2017; Spindler et al., 2020).

5.7 Systemic Imperatives

A further objection is to remark that beneficial machine learning applications, as supposed in this paper, stand in contradiction to systemic imperatives and goals of the economy. While making autonomous cars safer should result in having a competitive advantage, rendering recommendation systems on e-commerce platforms towards promoting sustainable, but more expensive products or on social media platforms towards less engaging content means that the platforms acquire fewer purchasers or users who can be influenced by online advertisements. However, the point of making something beneficial, in this case machine learning applications, is to overwrite systemic imperatives in case they have detrimental effects for particular individuals or society at large. Being successful according to the logic of a certain social system (Luhmann, 1995) does not necessarily mean that this success is morally justified (Habermas, 1987). This holds especially true with regard to the economy (Brand & Wissen, 2017). On a related note, beneficial machine learning applications, which are traditionally used in areas like health or crisis response, satellite image interpretation, climate action, poverty reduction, wildlife preservation, and the like (Chui et al., 2018) do in many cases not follow systemic but moral

imperatives. What is special in the case of the ideas presented here is that I do not propose to invent new machine learning application in hitherto undiscovered fields of society. Rather, I opt for reshaping applications that are already entrenched in areas that are structured by systemic imperatives and eventually aim at being profitable. In contexts that are purely dominated by monetary considerations, it is of course difficult to put the ideas into practice. Nevertheless, companies should at no point see the pursuit of profits as their only target. And as soon as they include social responsibility into their repertory of values, they can embrace the presented approach for beneficial machine learning.

6 Conclusion

In their classic book "Moral Machines", Wallach and Allen state: "The vision of learning systems developing naturally toward an ethical sensibility that values humans and human ethical concerns is an optimistic vision […]." (Wallach & Allen, 2009, p. 110) This paper is a tangible proposal how this vision could be put into practice. It stresses the importance of "feeding" machine learning applications not with all relevant behavioral data that is available, but with a particular selection of it, namely with quality data. Following the typical big data approach and using all available data to train models can have detrimental effects. This can not only be shown by, for instance, pointing at various cases of algorithmic discrimination. It only recently got obvious when COVID-19 caused dramatic changes in online shopping and other digitally recorded behavior, so that its inclusion in training sets caused machine learning applications to malfunction, making manual interventions necessary (Heaven, 2020). Thus, more rigorous mechanisms to filter training data sets have to be put in place, ensuring that, among others, only "good data" become training stimuli, meaning that digitally recorded behavior is classified and assessed along ethical criteria. Moral machine behavior is dependent on moral human behavior. Hence, both have to be linked. Incidentally, this idea is in accordance with the proposal to found "computational psychiatry" where computational problems like malfunctions or machine biases are seen in analogy to mental disorders (Schulz & Dayan, 2020). Taking up this proposal, this paper can be seen as an initial suggestion on how to give "therapy" to machine learning models.

Many machine learning applications acquire their "intelligence" by "capturing" (Mühlhoff, 2019, 1873) aggregated human cognitive and behavioral abilities. Hitherto, these aggregations of recordings of human behavior are hardly presorted before becoming training stimuli for machine behavior. This paper is a plea to do so and thereby to achieve truly beneficial machine learning. Its arguments start at the assumption that a person's social background, dispositions, educational level, etc. shapes his or her way of using digital devices. In turn, those devices are able to track, infer, and measure a user's personal states or traits. Depending on these attributes, an ethical assessment of data traces the respective user produces, namely the data traces that provide the basis to the measurements itself as well as data traces that are situated in the same context as the measurement, can take place. Subsequently, this assessment enables to single out quality data that are representing

behavior of individuals who are deemed to be especially competent, eligible, or morally versed for a particular task. This method for sampling out particular training stimuli stands in contrast to the *n = all* ideology. It serves to tailor training data in a way that machine behavior can correspond to values of the common good and become truly beneficial.

Acknowledgements This research was supported by the Cluster of Excellence "Machine Learning – New Perspectives for Science" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—Reference Number EXC 2064/1—Project ID 390727645. I would like to thank Sarah Fabi, Zeynep Akata, Ulrike von Luxburg, and Sebastian Bordt for very helpful comments on the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abdoli, N., Farnia, V., Delavar, A., Esmaeili, A., Dortaj, F., Farrokhi, N., et al. (2015). Poor mental health status and aggression are associated with poor driving behavior among male traffic offenders. *In Neuropsychiatric Disease and Treatment, 11*, 2071–2078.

Acar, Gunes; Eubank, Christian; Englehardt, Steven; Juarez, Marc; Narayanan, Arvind; Diaz, Claudia (2014): The Web Never Forgets. Persistent Tracking Mechanisms in the Wild. In Gail-Joon Ahn, Moti Yung, Ninghui Li (Eds.): Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14. ACM SIGSAC Conference. Scottsdale, Arizona. New York: ACM Press, pp. 674–689.

Amodei, Dario; Olah, Chris; Steinhardt, Jacob; Christiano, Paul; Schulman, John; Mané, Dan (2017): Concrete Problems in AI Safety. In arXiv, pp. 1–29.

Amoore, Louise (2011): Data Derivatives. On the Emergence of a Security Risk Calculus for Our Times. In Theory, Culture & Society 28 (6), pp. 24–43.

Anstey, K. J., Wood, J., Lord, S., & Walker, J. G. (2005). Cognitive, sensory and physical factors enabling driving safety in older adults. *In Clinical Psychology Review, 25*(1), 45–65.

Asimov, I. (2004). *I, Robot*. Random House LLC.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *In California Law Review, 104*, 671–732.

Belliger, Andréa; Krieger, David J. (2018): Network Public Governance. On Privacy and the Informational Self. Bielefeld: Transcript.

Bender, Emily M.; Gebru, Timnit; McMillan-Major, Angelina; Mitchell, Margaret (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?: ACM, pp. 1–14.

Bi, Bin; Shokouhi, Milad; Kosinski, Michal; Graepel, Thore (2013): Inferring the demographics of search users. In Daniel Schwabe, Virgílio Almeida, Hartmut Glaser, Ricardo Baeza-Yates, Sue Moon (Eds.): Proceedings of the 22nd International Conference on World Wide Web - WWW '13. New York: ACM Press, pp. 131–140.

Biczók, G., & Chia, P. H. (2013). *Interdependent privacy: Let me share your data*. Springer.

Bogard, William (1996): The Simulation of Surveillance. Hypercontrol in Telematic Societies. Cambridge: Cambridge University Press.

Bourdieu, Pierre (1984): Distinction. A Social Critique of the Judgement of Taste. Cambridge, Massachusetts: Harvard University Press.

Bourdieu, P. (1989). Social space and symbolic power. *In Sociological Theory, 7*(1), 14–25.

boyd, danah, . (2012). White Flight in Networked Publics. How Race and Class Shaped American Teen Engagement with MySpace and Facebook. In L. Nakamura & P. A. Chow-White (Eds.), *Race After the Internet* (pp. 203–222). Routledge.

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh; Kalai, Adam (2016): Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In arXiv, pp. 1–25.

Brand, Ulrich; Wissen, Markus (2017): Imperiale Lebensweise. Zur Ausbeutung von Mensch und Natur im globalen Kapitalismus. München: oekom Verlag.

Brey, P. (2010). Values in technology and disclosive computer ethics. In L. Floridi (Ed.), *The Cambridge Handbook of Information and Computer Ethics* (pp. 41–58). Cambridge University Press.

Calvo, R. A., Peters, D., & Cave, S. (2020). Advancing impact assessment for intelligent systems. *In Nature Machine Intelligence, 1*, 1–3.

Chui, Michael; Harryson, Martin; Manyika, James; Roberts, Roger; Chung, Rita; van Heteren, Ashley; Nel, Pieter (2018): Notes from the AI Frontier. Applying AI for Social Good. McKinsey Global Institute: McKinsey&Company, pp. 1–52.

Clarke, R. (1988). Information technology and dataveillance. *In Communications of the ACM, 31*(5), 498–512.

Clarke, R. (1994). The digital persona and its application to data surveillance. *In the Information Society, 10*(2), 77–92.

Crawford, Kate (2021): Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press.

Crawford, Kate; Dobbe, Roel; Dryer, Theodora; Fried, Genevieve; Green, Ben; Kaziunas, Elizabeth et al. (2019): AI Now 2019 Report. AI Now. New York. Available online at https://ainowinstitute.org/AI_Now_2019_Report.pdf, checked on 12/18/2019.

Čubranić-Dobrodolac, M., Lipovac, K., Čičević, S., & Antić, B. (2017). A model for traffic accidents prediction based on driver personality traits assessment. *In PROMET, 29*(6), 631–642.

Daly, Angela; Hagendorff, Thilo; Hui, Li; Mann, Monique; Marda, Vidushi; Wagner, Ben et al. (2019b): Artificial Intelligence, Governance and Ethics: Global Perspectives. The Chinese University of Hong Kong Faculty of Law Research Paper No. 2019–15. In SSRN Journal, pp. 1–41.

Daly, A., Devitt, K. S., & Mann, M. (Eds.). (2019a). *Good Data*. Institute of Network Cultures.

Danaher, J. (2018). Toward an ethics of AI assistants. *An Initial Framework. in Philos. Technol., 31*(4), 629–653.

Danaher, John; Hogan, Michael J.; Noone, Chris; Kennedy, Rónán; Behan, Anthony; Paor, Aisling de et al. (2017): Algorithmic governance. Developing a research agenda through the power of collective intelligence. In Big Data & Society 4 (2), 205395171772655.

Davidow, William; Malone, Michael S. (2020): Don't Regulate Artificial Intelligence: Starve It (Scientific American). Available online at https://blogs.scientificamerican.com/observations/dont-regulate-artificial-intelligence-starve-it/, checked on 5/8/2020.

Deleuze, Gilles (1992): Postscript on the Societies of Control. In October 59, pp. 3–7.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *In Journal of Educational Psychology, 105*(4), 1082–1099.

D'Mello, S., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *In ACM Comput. Surv., 47*(3), 1–36.

Domingos, Pedro (2015): The Master Algorithm. How the Quest for the Ultimate Learning Machine Will Remake Our World. New York: Basic Books.

Dutta, Sanghamitra; Venkatesh, Praveen; Mardziel, Piotr; Datta, Anupam; Grover, Pulkit (2020): An Information-Theoretic Quantification of Discrimination with Exempt Features. AAAI Conference on Artificial Intelligence, pp. 1–28.

Dwork, Cynthia (2008): Differential Privacy: A Survey of Results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, Angsheng Li (Eds.): Theory and Applications of Models of Computation. Berlin: Springer Berlin Heidelberg, pp. 1–19.

Dwork, C., et al. (2006). Differential Privacy. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, & J. C. Mitchell (Eds.), *Automata, Languages and Programming* (pp. 1–12). Springer.

Dwork, C., & Roth, A. (2013). The algorithmic foundations of differential privacy. *In FNT in Theoretical Computer Science, 9*(3–4), 211–407.

Dwork, Cynthia; Hardt, Moritz; Pitassi, Toniann; Reingold, Omer; Zemel, Richard (2011): Fairness Through Awareness. In arXiv, pp. 1–24.

Eady, T. A. (2019): Why Tesla's Fleet Miles Matter for Autonomous Driving (Medium). Available online at https://towardsdatascience.com/why-teslas-fleet-miles-matter-for-autonomous-driving-8e48503a462f, checked on 5/11/2020.

Engelmann, Severin; Chen, Mo; Fischer, Felix; Kao, Ching-yu; Grossklags, Jens (2019): Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior. In Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19, pp. 69–78.

Erikson, E. H. (1980). *Identity and the life cycle*. W.W. Norton.

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *In the British Journal of Sociology, 30*(4), 415–441.

Eubanks, Virginia (2018): Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Marting's Press.

Eyal, Nir; Hoover, Ryan (2014): Hooked. How to build Habit-Forming Products. Princeton, NJ: Princeton University Press.

Fergusson, D., Swain-Campbell, N., & Horwood, J. (2003). Risky driving behaviour in young people: prevalence, personal characteristics and traffic accidents. *In Australian and New Zealand Journal of Public Health, 27*(3), 337–342.

Floridi, Luciano; Cowls, Josh; Beltrametti, Monica; Chatila, Raja; Chazerand, Patrice; Dignum, Virginia et al. (2018): AI4People - An Ethical Framework for a Good AI Society. Opportunities, Risks, Principles, and Recommendations. In Minds and Machines 28 (4), pp. 689–707.

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

Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; Daumeé, Hal, III; Crawford, Kate (2018): Datasheets for Datasets. In arXiv, pp. 1–17.

Ghani, Rayid; Schierholz, Malte (2017): Machine Learning. In Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, Julia Lane (Eds.): Big Data and Social Science. A Practical Guide to Methods and Tools. Boca Raton: CRC Press, pp. 147–186.

Gitelman, L. (Ed.). (2013). *"Raw Data" Is an Oxymoron*. The MIT Press.

Grusky, David B. (2019): Social stratification. Class, race, and gender in sociological perspective. London: Routledge.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. *In International Journal on Advances in Software, 10*, 1–20.

Guntuku, S., & Chandra; Yaden, David B., Kern, Margaret L., Ungar, Lyle H., Eichstaedt, Johannes C. . (2017). Detecting depression and mental illness on social media: an integrative review. *In Current Opinion in Behavioral Sciences, 18*, 43–49.

Habermas, Jürgen (1987): Theorie des kommunikativen Handelns. 2 volumes. Frankfurt a.M: Suhrkamp (1).

Hagendorff, Thilo (2019c): Maschinelles Lernen und Diskriminierung. Probleme und Lösungsansätze. In Österreichische Zeitschrift für Soziologie 44 (1), pp. 53–66.

Hagendorff, T. (2019a). From privacy to anti-discrimination in times of machine learning. *In Ethics and Information Technology, 33*(3), 331–343.

Hagendorff, T. (2019b). Jenseits der puren Datenökonomie - Social-Media-Plattformen besser designen. In C. Ochs, M. Friedewald, T. Hess, & J. Lamla (Eds.), *Die Zukunft der Datenökonomie* (pp. 327–342). Springer.

Haggerty, K. D., & Ericson, R. V. (2000). The surveillant assemblage. *In the British Journal of Sociology, 51*(4), 605–622.

Haidt, Jonathan (2001): The Emotional Dog and Its Rational Tail. A Social Intuitionist Approach to Moral Judgment. In Psychology Review 108 (4), pp. 814–834.

Harari, G. M., Müller, S. R., Aung, M. S. H., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *In Current Opinion in Behavioral Sciences, 18*, 83–90.

Hargittai, E. (2008). The Digital Reproduction of Inequality. In D. B. Grusky (Ed.), *Social Stratification* (pp. 936–944). Westview Press.

Hart, D., Hofmann, V., Edelstein, W., & Keller, M. (1997). The relation of childhood personality types to adolescent behavior and development: a longitudinal study of Icelandic children. *In Developmental Psychology, 33*(2), 195–205.

Heaven, Douglas Will (2020): Our weird behavior during the pandemic is messing with AI models (MIT Technology Review). Available online at https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/, checked on 6/15/2020.

Hu, Jian; Zeng, Hua-Jun; Li, Hua; Niu, Cheng; Chen, Zheng (2007): Demographic prediction based on user's browsing behavior. In Carey Williamson, Mary Ellen Zurko, Peter Patel-Schneider, Prashant Shenoy (Eds.): Proceedings of the 16th International Conference on World Wide Web - WWW '07. New York: ACM Press, pp. 151–160.

Hyman, M. M. (1968). Accident vulnerability and blood alcohol concentrations of drivers by demographic characteristics. *In q. J. Stud. Alcohol Suppl., 29*(S4), 34–57.

Irani, L. (2016). The hidden faces of automation. *In XRDS, 23*(2), 34–37.

Irvin, Jeremy; Rajpurkar, Pranav; Ko, Michael; Yu, Yifan; Ciurea-Ilcus, Silviana; Chute, Chris et al. (2019): CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In arXiv, pp. 1–9.

Jo, Eun Seo; Gebru, Timnit (2019): Lessons from Archives. Strategies for Collecting Sociocultural Data in Machine Learning. In arXiv, pp. 1–11.

John, Oliver P.; Naumann, Laura P.; Soto, Christopher J. (2008): Paradigm Shift to the Integrative Big Five Trait Taxonomy. History, Measurement, and Conceptual Issues. In Oliver P. John, Richard W. Robins, Lawrence A. Pervin (Eds.): Handbook of Personality. Theory and Research. New York: The Guilford Press, pp. 114–158.

Jordan, Michael I.; Mitchell, Tom M. (2015): Machine learning. Trends, perspectives, and prospects. In Science 349 (6245), pp. 255–260.

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin.

Kant, Immanuel (1977): Kants Werke, Akademie Textausgabe. Anmerkungen der Bände I-[IX]: Walter de Gruyter.

Kearns, Michael; Roth, Aaron (2020): The Ethical Algorithm. The Science of Socially Aware Algorithm Design. New York: Oxford University Press.

Khan, Iftikhar Ahmed; Brinkman, Willem-Paul; Fine, Nick; Hierons, Robert M. (2008): Measuring personality from keyboard and mouse use. In Joaquim Jorge (Ed.): Proceedings of the 15th European Conference on Cognitive Ergonomics the Ergonomics of Cool Interaction - ECCE '08. New York: ACM Press, pp. 1–8.

Kitchin, Rob; Dodge, Martin (2011): Code/Space. Software and Everyday Life. Cambridge, Massachusetts: The MIT Press.

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

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *In GeoJournal, 80*(4), 463–475.

Kleinberg, Jon M.; Ludwig, Jens; Mullainathan, Sendhill (2016): A Guide to Solving Social Problems with Machine Learning (Harvard Business Review). Available online at https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning, checked on 12/1/2017.

Köbis, Nils; Bonnefon, Jean-François; Rahwan, Iyad (2021): Bad machines corrupt good morals. In Nat Hum Behav.

Kohlberg, Lawrence; Levine, Charles; Hewer, Alexandra (1983): Moral stages. A current formulation and a response to critics. Basel: Karger.

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: an interdisciplinary challenge. *In IEEE Intell. Transport. Syst. Mag., 9*(1), 90–96.

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *In Machine Learning, 95*(3), 357–380.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *In Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802–5805.

Kuss, Daria J.; Griffiths, Mark D. (2017): Social Networking Sites and Addiction. Ten Lessons Learned. In International journal of environmental research and public health 14 (3).

Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *In SIGKDD Explor. Newsl., 12*(2), 74–82.

Lerner, R. M. (Ed.). (2015). *Handbook of child psychology and developmental science*. Hoboken, New Jersey: Wiley.
Lessig, Lawrence (2006): Code. Version 2.0. New York: Basic Books.
Lewis, K. (2015). Three fallacies of digital footprints. *In Big Data & Society, 2*(2), 1–4.
Lischka, Konrad; Stöcker, Christian (2017): Digitale Öffentlichkeit. Wie algorithmische Prozesse den gesellschaftlichen Diskurs beeinflussen. Arbeitspapier. Gütersloh: Bertelsmann Stiftung, pp. 1–88.
Loevinger, J. (1997). Stages of Personality Development. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of Personality Psychology* (pp. 199–208). Elsevier.
Los, Maria (2006): Looking into the future: surveillance, globalization and the totalitarian potential. In David Lyon (Ed.): Theorizing Surveillance. The panopticon and beyond. Cullompton: Willian Publishing, pp. 69–94.
Luhmann, N. (1995). *Social Systems*. Stanford University Press.
Lyon, David (2003): Surveillance as social sorting. Computer codes and mobile bodies. In David Lyon (Ed.): Surveillance as Social Sorting. Privacy, risk, and digital discrimination. London: Routledge, pp. 13–30.
Matz, S. C., & Netzer, O. (2017). Using Big Data as a window into consumers' psychology. *In Current Opinion in Behavioral Sciences, 18*, 7–12.
Matzner, Tobias (2016): Beyond data as representation. The performativity of Big Data in surveillance. In Surveillance & Society 14 (2), pp. 197–210.
Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A., & Brox, T. (2018). What makes good synthetic training data for learning disparity and optical flow estimation? *In Int J Comput vis, 126*(9), 942–960.
Mayer-Schönberger, Viktor; Cukier, Kenneth (2013): Big Data. A Revolution That Will Transform How We Live, Work, and Think. New York: Eamon Dolan.
McCloud, R. F., Okechukwu, C. A., Sorensen, G., & Viswanath, K. (2016). Entertainment or health? exploring the internet usage patterns of the urban poor: a secondary analysis of a randomized controlled trial. *In Journal of Medical Internet Research, 18*(3), 1–12.
McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *In Journal of Personality, 60*(2), 175–215.
Mittelstadt, Brent; Russell, Chris; Wachter, Sandra (2019): Explaining Explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19, pp. 1–10.
Mjolsness, Eric; DeCoste, Dennis (2001): Machine learning for science. State of the art and future prospects. In Science 293 (5537), pp. 2051–2055.
Mossberger, Karen; Tolbert, Caroline J.; Stansbury, Mary (2003): Virtual inequality. Beyond the digital divide. Washington, D.C.: Georgetown University Press.
Mühlhoff, Rainer (2019): Human-aided artificial intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning. In New Media & Society, 1–17.
Mühlhoff, Rainer (2021): Predictive Privacy: Towards an Applied Ethics of Data Analytics. In SSRN Journal, pp. 1–24.
Myers, Isabel Briggs; Myers, Peter B. (1995): Gifts Differing. Understanding Personality Type. Palo Alto: Davies-Black.
Ng, Andrew (2020): Facebook Likes Extreme Content (The Batch). Available online at https://blog.deeplearning.ai/blog/the-batch-facebooks-unruly-algorithm-ai-that-does-the-dishes-new-life-for-old-data-models-that-take-shortcuts-yolo-returns, checked on 6/19/2020.
Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: biases, methodological pitfalls, and ethical boundaries. *In Front. Big Data, 2*, 1–33.
O'Neil, Cathy (2016): Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers.
Pang, Bo., & Lee, L. (2008). Opinion mining and sentiment analysis. *In FNT in Information Retrieval, 2*(1–2), 1–135.
Patel, T. (2006). Comparing the usefulness of conventional and recent personality assessment tools. *In Global Business Review, 7*(2), 195–218.
Paul, B., & Baltes; David L. Featherman; Richard M. Lerner; Orville Gilbert Brim; Marion Perlmutter, . (1978). *Life Span Development and Behavior*. Academic Press Inc.
Perrons, R. K., & McAuley, D. (2015). The case for "n«all": why the big data revolution will probably happen differently in the mining sector. *In Resources Policy, 46*, 234–238.
Picard, R. W. (1997). *Affective computing*. MIT Press.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., et al. (2019). Machine behaviour. *In Nature, 568*(7753), 477–486.

Raji, Inioluwa Deborah; Smart, Andrew; White, Rebecca N.; Mitchell, Margaret; Gebru, Timnit; Hutchinson, Ben et al. (2020): Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In arXiv, pp. 1–12.

Rawls, J. (1999). *A Theory of Justice*. Oxford University Press.

Roßnagel, A. (2007). *Datenschutz in einem informatisierten Alltag*. Gutachten im Auftrag der Friedrich-Ebert-Stiftung.

Samarati, Pierangela; Sweeney, Latanya (1998): Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. In Technical Report SRI, pp. 1–19.

Samitsch, Christoph (2015): Data Quality and its Impacts on Decision-Making. How Managers can benefit from Good Data. Wiesbaden: Springer.

Sartorius, R. (Ed.). (1983). *Paternalism*. University of Minnesota Press.

Schneier, Bruce (2015): Data and Goliath. The Hidden Battles to Collect Your Data and Control Your World. New York: W. W. Norton & Company.

Schulz, Eric; Dayan, Peter (2020): Computational Psychiatry for Computers. In iScience 23 (12), p. 101772.

Schulze, Gerhard (1996): Die Erlebnis-Gesellschaft. Kultursoziologie der Gegenwart. Frankfurt am Main: Campus Verlag.

Schwartz, H., Andrew; Eichstaedt, Johannes C., Kern, Margaret L., Dziurzynski, Lukasz; Ramones, Stephanie M., Agrawal, Megha, , et al. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *In PloS One, 8*(9), 1–16.

Segalin, Cristina; Celli, Fabio; Polonio, Luca; Kosinski, Michal; Stillwell, David; Sebe, Nicu et al. (2017): What your Facebook Profile Picture Reveals about your Personality. In Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Phoebe Chen et al. (Eds.): Proceedings of the 2017 ACM on Multimedia Conference - MM '17. New York: ACM Press, pp. 460–468.

Sheng, Emily; Chang, Kai-Wei; Natarajan, Premkumar; Peng, Nanyun (2019): The Woman Worked as a Babysitter: On Biases in Language Generation. In Kentaro Inui, Jing Jiang, Vincent Ng, Xiaojun Wan (Eds.): Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: Association for Computational Linguistics, pp. 3405–3410.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *In J Big Data, 6*(1), 1–48.

Silvervarg, A., Raukola, K., Haake, M., & Gulz, A., et al. (2012). The Effect of Visual Gender on Abuse in Conversation with ECAs. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, & J. C. Mitchell (Eds.), *Intelligent Virtual Agents* (pp. 153–160). Springer.

Sloterdijk, Peter (2009): Du mußt dein Leben ändern. Über Anthropotechnik. Frankfurt a.M: Suhrkamp.

Solaiman, Irene; Clark, Jack; Brundage, Miles (2019): GPT-2: 1.5B Release. OpenAI. Available online at https://openai.com/blog/gpt-2-1-5b-release/, checked on 11/13/2019.

Spera, C. (2005). A review of the relationship among parenting practices, parenting styles, and adolescent school achievement. *Educational Psychology Review, 17*(2), 125–146.

Spindler, M., Booz, S., Gieseler, H., Runschke, S., Wydra, S., & Zinsmaier, J. (2020). How to achieve integration? In B. Gransche & A. Manzeschke (Eds.), *Das geteilte Ganze* (pp. 213–239). Springer Fachmedien Wiesbaden.

Stieglitz, Stefan; Dang-Xuan, Linh (2012): Impact And Diffusion Of Sentiment In Public Communication On Facebook. In ECIS 2012 Proceedings 98, pp. 1–12.

Stieglitz, Stefan; Dang-Xuan, Linh (2013): Emotions and Information Diffusion in Social Media. Sentiment of Microblogs and Sharing Behavior. In Journal of Management Information Systems 29 (4), pp. 217–248.

Sullivan, P. M., & Knutson, J. F. (2000). Maltreatment and disabilities: A population-based epidemiological study. *Child Abuse & Neglect, 24*(10), 1257–1273.

Suresh, Harini; Guttag, John V. (2020): A Framework for Understanding Unintended Consequences of Machine Learning. In arXiv, pp. 1–10.

Tellegen, Auke; Waller, Niels G. (2008): Exploring Personality Through Test Construction: Development of the Multidimensional Personality Questionnaire. In Gregory J. Boyle, Gerald Matthews, Don

Saklofske (Eds.): The SAGE Handbook of Personality Theory and Assessment: Volume 2. Personality Measurement and Testing. London: SAGE Publications Ltd, pp. 261–292.
Brundage, Miles; Avin, Shahar; Wang, Jasmine; Belfield, Haydn; Krueger, Gretchen; Hadfield, Gillian et al. (2020): Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. In arXiv, pp. 1–80.
Tsutsui, Satoshi; Kerola, Tommi; Saito, Shunta; Crandall, David J. (2018): Minimizing Supervision for Free-space Segmentation. In arXiv, pp. 1–10.
Turow, Joseph (2012): The Daily You. How the New Advertising Industry Is Defining Your Identity and Your Worth. New Haven: Yale University Press.
Underwood, Mimi (2015): Updating Our Search Quality Rating Guidelines (Google Blog). Available online at https://webmasters.googleblog.com/2015/11/updating-our-search-quality-rating.html, checked on 5/29/2020.
Brundage, Miles; Avin, Shahar; Clark, Jack; Toner, Helen; Eckersley, Peter; Garfinkel, Ben et al. (2018): The Malicious Use of Artificial Intelligence. Forecasting, Prevention, and Mitigation. In arXiv, pp. 1–101.
van Dijck, J. (2014). Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. *In Surveillance & Society, 12*(2), 197–208.
van Ly, Minh; Martin, Sujitha; Trivedi, Mohan M. (2013): Driver classification and driving style recognition using inertial sensors. In : 2013 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE, pp. 1040–1045
Vaughn, M. G., Define, R. S., Delisi, M., Perron, B. E., Beaver, K. M., Fu, Q., & Howard, M. O. (2011). Sociodemographic, behavioral, and substance use correlates of reckless driving in the United States: findings from a national Sample. *In Journal of Psychiatric Research, 45*(3), 347–353.
Veale, Michael; Binns, Reuben (2017): Fairer machine learning in the real world. Mitigating discrimination without collecting sensitive data. In Big Data & Society 4 (2), 1–17.
Vester, Michael (2001): Soziale Milieus im gesellschaftlichen Strukturwandel. Zwischen Integration und Ausgrenzung. Frankfurt: Suhrkamp.
Wallach, Wendell; Allen, Colin (2009): Moral Machines. Teaching Robots Right from Wrong. New York: Oxford University Press.
Wang, Richard Y.; Strong, Diane M. (1996): Beyond Accuracy. What Data Quality Means to Data Consumers. In Journal of Management Information Systems 12 (4), pp. 5–33.
Wang, X., Huang, K., & Yang, Li. (2019). Effects of socio-demographic, personality and mental health factors on traffic violations in Chinese bus drivers. *In Psychology, Health & Medicine, 24*(7), 890–900.
Wayne, N. L., & Miller, G. A. (2018). Impact of gender, organized athletics, and video gaming on driving skills in novice drivers. *In PloS One, 13*(1), 1–12.
Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *In Journal of Marketing, 80*(6), 97–121.
Wendehorst, Christiane; Woopen, Christiane; Haberer, Johanna; Heckmann, Dirk; Hansen, Marit; Kelber, Ulrich et al. (2019): Opinion of the Date Ethics Commission. Berlin: Data Ethics Commission of the Federal Government, pp. 1–240. Available online at https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2, checked on 7/4/2021.
West, Mark; Kraut, Rebecca; Chew, Han Ei (2019): I'd blush if I could: closing gender divides in digital skills through education, pp. 1–146.
West, R., & Hall, J. (1997). The role of personality and attitudes in traffic accident risk. *In Applied Psychology, 46*(3), 253–264.
Westin, A. F. (1967). *Privacy and Freedom*. Atheneum.
Willke, Helmut (2005): Symbolische Systeme. Grundriss einer soziologischen Theorie. Weilerswist: Velbrück.
Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming. *In SIGCAS Comput. Soc., 47*(3), 54–64.
Woodall, P., Oberhofer, M., & Borek, A. (2014). A classification of data quality assessment and improvement methods. *In IJIQ, 3*(4), 298–321.
Xu, C., Ding, Z., Wang, C., & Li, Z. (2019). Statistical analysis of the patterns and characteristics of connected and autonomous vehicle involved crashes. *In Journal of Safety Research, 71*, 41–47.
Young, M. D. (1994). *The Rise of the Meritocracy*. Transaction.

Yu, Pu., & Grossklags, J. (2016). Towards a model on the factors influencing social app users' valuation of interdependent privacy. *In Proceedings on Privacy Enhancing Technologies, 2*, 61–81.

Zeng, Zhihong; Pantic, Maja; Roisman, Glenn I.; Huang, Thomas S. (2009): A survey of affect recognition methods: audio, visual, and spontaneous expressions. In IEEE Trans. Pattern Anal. Mach. Intell. 31 (1), pp. 39–58.

Zhuang, Fuzhen; Qi, Zhiyuan; Duan, Keyu; Xi, Dongbo; Zhu, Yongchun; Zhu, Hengshu et al. (2020): A Comprehensive Survey on Transfer Learning. In arXiv, pp. 1–31.

Zillien, N., & Hargittai, E. (2009). Digital distinction: status-specific types of internet usage. *In Social Science Quarterly, 90*(2), 274–291.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *In Journal of Information Technology, 30*, 75–89.