Ethics and Information Technology
https://doi.org/10.1007/s10676-021-09606-x

ORIGINAL PAPER

Predictive privacy: towards an applied ethics of data analytics

Rainer Mühlhoff[1]

Accepted: 28 June 2021

Abstract
Data analytics and data-driven approaches in Machine Learning are now among the most hailed computing technologies in many industrial domains. One major application is predictive analytics, which is used to predict sensitive attributes, future behavior, or cost, risk and utility functions associated with target groups or individuals based on large sets of behavioral and usage data. This paper stresses the severe ethical and data protection implications of predictive analytics if it is used to predict sensitive information about single individuals or treat individuals differently based on the data many unrelated individuals provided. To tackle these concerns in an applied ethics, first, the paper introduces the concept of "predictive privacy" to formulate an ethical principle protecting individuals and groups against differential treatment based on Machine Learning and Big Data analytics. Secondly, it analyses the typical data processing cycle of predictive systems to provide a step-by-step discussion of ethical implications, locating occurrences of predictive privacy violations. Thirdly, the paper sheds light on what is qualitatively new in the way predictive analytics challenges ethical principles such as human dignity and the (liberal) notion of individual privacy. These new challenges arise when predictive systems transform statistical inferences, which provide knowledge about the cohort of training data donors, into individual predictions, thereby crossing what I call the "prediction gap". Finally, the paper summarizes that data protection in the age of predictive analytics is a collective matter as we face situations where an individual's (or group's) privacy is violated using data *other* individuals provide about themselves, possibly even anonymously.

Keywords Predictive analytics · Ethics of Big Data · Automated decision making · Bias · Privacy · Group privacy

Introduction

Data analytics and data-driven approaches in Machine Learning (ML) are now among the most hailed computing technologies in many industrial domains. One major application is the algorithmic prediction of human behavior, or human "fate" , if you will: Predictive Analytics (PA) leverages large behavioral data sets to classify individuals according to future risks, economic developments or expected costs and utility, as predicted from data correlations (O'Neil, 2016; Wachter & Mittelstadt, 2018; Mühlhoff, 2020a). For instance, online targeted advertising uses PA to decide which version of a message is shown to a user (Reilly, 2017; Duhigg, 2012). Differential insurance pricing uses PA to determine individual insurance risks and individual insurance premiums (Varner & Sankin, 2020). Hiring algorithms use PA to select, or short-list, job applicants (Bogen, 2019). Criminal recidivism scoring, such as the "COMPAS" system in the USA, predicts the likelihood of a criminal re-offending (Angwin et al., 2016; Fry, 2018).

PA is regularly used in contexts where decisions have a potentially life-changing impact on the affected individuals and social groups. This raises ethical concerns, some of which have been vividly discussed in fields such as algorithm ethics (Mittelstadt et al., 2016) and ethics of AI (cf. Coeckelbergh, 2020a), as well as by governmental bodies such as the European Commission's High-Level Expert Group on AI (2019) and industry organizations such as the IEEE (Chatila & Havens, 2019). Among the most debated ethical concerns are unfair bias and discrimination: predictive decisions might perpetuate existing or create new patterns of unfair discrimination against minority groups or vulnerable demographics (Barocas & Selbst, 2016). Other

✉ Rainer Mühlhoff
muehlhoff@tu-berlin.de; muehlhoff@ethikderki.de
https://RainerMuehlhoff.de

[1] Excellence Cluster Science of Intelligence, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany

Published online: 31 July 2021

concerns relate to transparency and explainability of PA: in the common depiction as "black boxes" , many ML and PA algorithms do not provide the necessary output to "justify" their decisions (or predictions) to a satisfying extent, making it ethically and legally questionable whether life-changing consequences can and should be based on such output (Mittelstadt et al., 2016). From a more structural perspective, critical commentators have added that the deployment of PA might stabilize socioeconomic inequality, exploitation and oppression on a macro-societal scale (O'Neil, 2016; Eubanks, 2017; Noble, 2018). Discussions in data protection, moreover, have been highlighting that PA challenges contemporary privacy frameworks such as the EU's General Data Protection Regulation (GDPR; Wachter, 2019; Zarsky, 2016; Mühlhoff, 2020b).

This paper will take a more detailed look at the data processing cycle of predictive systems in order to shape an applied ethics of PA that includes but also goes beyond the hot-spot problems of bias and explainability. The paper's contribution is both a normative and an analytic one: Normatively, I will propose the notion of "predictive privacy" as a central ethical principle that is being threatened by PA. The principle of *predictive privacy* seeks to protect individuals and groups against unfair treatment and infringements on their autonomy, dignity and well-being resulting from the use of information that is being merely *predicted* by leveraging statistical correlations with others' behavior. An individual's (or group's) predictive privacy is *violated* if sensitive information about that individual is statistically estimated against their will or without their knowledge on the basis of data of many other individuals, provided that these predictions lead to differential treatment or decisions that affect anyone's social, economic, psychological, physical, ... well-being or freedom. Analytically, the paper will dissect the various steps from data input to data output of predictive systems in order to discuss a spectrum of ethical concerns connected to each point in the data processing cycle. This step-by-step procedure is meant both as an academic contribution to a more nuanced understanding of the ethical challenges of PA and as a guide towards the operationalization of ethical thought in responsible implementations and political regulations of PA. Finally, the paper's overarching goal is to point out that tackling ethical issues of PA demands a collectivist instead of an individualist approach to privacy.

In the next section ("Predictive systems and predictive privacy"), I will briefly formalize my understanding of predictive systems and give a definition of *predictive privacy*. In the subsequent section ("Ethical evaluation of predictive systems"), I will dissect the typical data processing cycle of predictive systems, distinguishing six different steps grouped into two phases ("building" vs. "training the model"). In the section on "Collective ethical concerns", I will extend the data processing model by two more steps that are geared towards preventing mispredictions (including *some* types of bias), while I will be arguing that addressing these ethical issues requires a collectivist approach. In the Conclusion, I will relate the concept of predictive privacy to the fundamental ethical principle of human dignity, also discussing that there might be good reasons to abandon the use of PA completely, as technological means of making it ethically viable are rather limited.

Predictive systems and predictive privacy

By the term "predictive analytics" (PA) I refer to algorithms and techniques in the fields of data analytics (cf. McCue, 2007; Grindrod, 2014), computational statistics and Machine Learning (Goodfellow, Bengio, & Courville, 2016) that can be used to predict sensitive attributes (e.g., personal, health, social, financial, etc. information), future behavior, or cost, risk and utility functions of target individuals based on large sets of data about many other individuals. By the related term "predictive system", I refer to real-world socio-technological assemblages, e.g. in the context of government or business applications, in which PA is used to aid or automate decision making that affects human users, customers, patients, applicants, defendants, etc. Formally, a predictive system is based on a predictive model or function,

$$P_W : D_i \mapsto A_i,$$

that depends on a stock W of empirical knowledge ("training data") and, based on this, returns a prediction A_i for input data, D_i. Here, D_i is typically the information ("proxy data") available about an individual user or case i, while A_i contains a prediction of certain characteristics unknown about i.

A predictive model P_W critically depends on the knowledge W about many other empirical cases: PA is based on inductive statistical reasoning. That is, it applies statistical inferences, which are derived from a large cohort of known cases, to new cases. The use of PA in real-world predictive systems, therefore, represents an "emerging new empiricism" (Rieder & Simon, 2017), at the heart of which is the idea to estimate individual cases on the basis of lateral comparisons to a large number of other cases contained in the knowledge base W rather than deducing the result A_i from the individual properties D_i.

There is a subtle but crucial difference between "inference" and "prediction" that informs the choice of terminology in this paper. The term "inference" commonly refers to "deduc[ing] or conclud[ing] (something) from evidence and reasoning rather than from explicit statements."[1] As the

[1] *The Oxford Dictionary of English*, https://www.lexico.com/definition/infer.

reasoning that is implemented by data analytics and ML is a form of statistical reasoning, we more specifically mean *statistical inference* when referring to "inference" in this context. "Statistical inference", in turn, is understood in statistics as "[t]he process of drawing conclusions about a population on the basis of measurements or observations made on a sample of units from the population" (Everitt & Skrondal, 2010, p. 2017).[2] In this way, the model P_W is a product of statistical inference from the training data W; what it describes is some statistical knowledge that *generalizes* from the statistical sample captured in the training data W to a larger population. Crucially, this step works with the assumption that the sample is representative of the population.

With this in mind, I maintain that the statistical model P_W, which is a product of statistical inference and thus represents knowledge about a population, turns into a *predictive* model the moment it is applied to a *single* individual i specified by a data signature D_i. As I will explain in the next section ("Ethical evaluation of predictive systems"), this step turns statistical inference into prediction, as it projects a population-level statistical knowledge to a single individual, leading to an ethical and epistemological obstacle that I call the *prediction gap*. The term "prediction" commonly refers to the act of "[s]ay[ing] or estimat[ing] that (a specified thing) will happen in the future".[3] The component "to say *that*" is a central part of "prediction" but stands in stark contrast with the principle of statistical inference which produces statistical knowledge that weighs different possible options by different probabilities and thus is of the type "to say what *might be*".

When an inferential model P_W is applied to an individual case D_i, the result $P_W(D_i)$ is in fact still a probability distribution. Using this probability distribution to make a prediction reduces the information to a single alternative as the outcome: turning "might be X with probability p_X and might be Y with probability p_Y" into "saying that it is X" and treating the individual accordingly. Turning inferences into predictions thus involves *betting* on one specific value of $P_W(D_i)$ next to all other possible alternatives. This crucial difference between inference and prediction is the starting point of the ethical considerations in this paper and the reason why I use the term "predictive systems" instead of "inferential systems" (cf. Mühlhoff, 2020a). In this use of the term prediction I will not overly rely on the temporal aspect that is etymologically implied in "prediction". Prediction may refer to future behavior or events, but equally to facts that have already actualized but are unknown to the entity making the prediction. The decisive aspect about prediction in the present scope is not the future temporality of the information content, but rather that *bets are made* about single individuals in a situation of incomplete information.

Examples

An example from the area of "targeted advertising" illustrates how predictive models are trained in real business contexts. In 2012 it became known that some retailers in the US try to identify pregnant supermarket customers by their purchasing behavior in order to provide them with tailored advertising (Duhigg, 2012). The relevant training data was collected by tracking the purchasing behavior of customers over an extended period, for example, by utilizing customer loyalty programs or credit card data. Expectant parents could then be identified *retrospectively* in this data stock as soon as they bought relevant baby products, and predictive models could be trained to identify early "markers" in their behavior as compared with the purchasing behavior of non-pregnant customers. The resulting predictive models could recognize pregnant customers even before they, themselves, or their social environment, knew they were pregnant (ibid.).

Pregnancy prediction is an example of a *classifier*, that is, a PA that is used to categorize customers into predictive groups (pregnant vs. non-pregnant). Other important kinds of PAs predict *continuous values*, such as risk scores, revenue, costs or utility functions. For example, in credit scoring, PA is applied to recommend individual credit conditions, tailored to each applicant based on behavioral correlations in a pool of credit customers (Hurley & Adebayo, 2017; O'Neil, 2016). "Payday lending" providers such as the US company ZestFinance or the German company Kreditch specialize in leveraging a wide range of customer data to serve high-risk market segments such as "the world's unbanked", who are considered "uncreditworthy" on the traditional financial market (O'Dwyer, 2018). ZestFinance even identifies the use of capital vs. lower-case letters in online application forms and measures the time it takes the applicant to read through the terms and conditions. Operators claim they "don't know why" certain parameters, such as using capital letters, are relevant markers of credit risk (Lippert, 2014). Typically for PA, the criteria are self-learned by the system and not hard-coded by human programmers, which often means that they are not easily explainable to or traceable by human operators or users.

How prediction is a challenge to privacy and data protection

Besides the prediction gap problem, the observation that PA challenges personal intimacy and privacy in new ways is the second factor driving the ethical approach of this paper.

[2] See also Hacking (2016) for the full history of the "Logic of statistic inference ", and Efron and Hastie (2018) for applications to the age of computer based inference.

[3] *The Oxford Dictionary of English*, https://www.lexico.com/definition/predict.

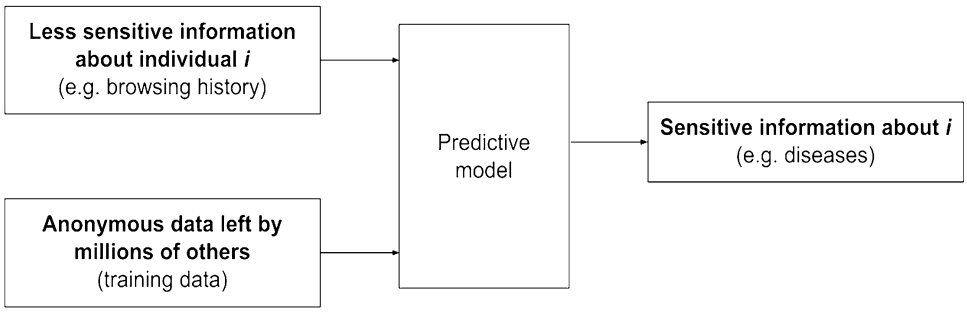

Fig. 1 Prediction of sensitive from less sensitive information

While a violation of intimacy was already evident in the case of pregnancy prediction, there is a host of further examples. Researchers at the University of Pennsylvania have shown that diseases such as depression, psychosis, diabetes or high blood pressure can be predicted from a user's postings on Facebook (Merchant et al., 2019). Kosinski, Stillwell and Graepel (2013) showed that sexual identity could be predicted from the same data. Facebook itself has announced the use of artificial intelligence to identify suicidal users based on their postings and automatically inform the authorities in acute cases (Goggin, 2019).

This suggests that PA raises a specific privacy concern: *Through PA, sensitive information about individuals or groups is predicted, potentially without the data subjects' knowledge, from less sensitive or more readily available information (proxy data) by leveraging the data left by millions of other users* (Fig. 1). As I argue, this presents a new challenge to data ethics and privacy regulation; here, privacy is compromised not by information the disclosed by the subject, but by the information revealed by many others (i.e., by the users of networked services as involuntary donors of training data). It makes for a qualitatively new data protection concern if information disclosed by many unrelated, potentially even anonymous, users helps estimate sensitive information about users who may lack representation in the training data (cf. Mantelero, 2016; Taylor, Floridi, & van der Sloot, 2016; Mittelstadt, 2017; Wachter & Mittelstadt, 2018; Mühlhoff, 2020b).

When I refer to the predicted information A_i as "sensitive information", I mean to address two distinct but related kinds of information at the same time: A_i could be some critical attribute of the data subject i, such as information on health, political views, socioeconomic status or any other attribute the data subject might not want to disclose to others in a given context. A_i can, however, also refer to data such as cost, risk or utility functions, expected turnout or profit, that are not intrinsic to the target individual but make sense only in relation to the business model of the entity running the prediction. That is, the PA may either classify target individuals according to their allegedly *inherent attributes* or estimated outcome of a certain *interaction* with the target individual, e.g. in a business context. Both, however, will be treated equivalently in terms of ethical concerns in this paper as the distinction is fluid. Models that predict critical attributes can easily be transformed into models predicting utility parameters. Conversely, utility parameters are "sensitive" as well if their estimation has a considerable impact on the data subject's access to resources, information and well-being. What makes the predicted information "sensitive" is not simply the fact that the individual might want to control how visible this information is to others. Rather, the term "sensitive" must be understood in a praxeological sense, referring to the potential effects and implications that the prediction has on the data subject as part of the sociotechnical system in which the predictive model operates.

Predictive privacy

The ethical approach of this paper addresses the twofold challenge regarding privacy and data protection due to the possibility of predicting sensitive information about individuals from proxy data through lateral comparisons to the data of many "data donors".[4] The first challenge relates to the prediction gap, the second to the estimation of sensitive information from the data of many others. Combined, I call this privacy concern "predictive privacy". A violation of predictive privacy is not committed by stealing or leaking information from someone's private sphere, but by deriving a prediction about individuals or profiling groups from data that were collected from *many other users* of networked digital services. Most users involuntarily contribute to the pool of training data by providing their data with consent, but may be unaware of how the data will be used *on others*. Also, proxy data about the target individual are mostly collected lawfully. Yet, all the involved individuals and groups might be unaware that the masses of feature-rich information provided by millions of users in contemporary internet

[4] I use the term "data donor" to refer to all of us in our role as everyday users of networked services. In this role, we involuntarily produce data. This includes personal and non-personal, identified and de-identified data, as well as data that we explicitly provide and data that is covertly collected about us, such as usage data, behavioral data, and metadata.

economy and media culture can be leveraged for predictions of information they would not want to disclose. To pinpoint this ethical concern, I will give a formal definition of the negative of predictive privacy—its violation:

Definition 1 An individual's or group's **predictive privacy is violated** if sensitive information about that person or group is predicted against their will or without their knowledge on the basis of data of many other individuals, provided that these predictions lead to decisions that affect anyone's social, economic, psychological, physical,... well-being or freedom.

Remark First, it is important to highlight that a prediction does not need to be accurate in order for predictive privacy to be violated. Adverse effects can equally result from inaccurate predictions. Secondly, I added the condition that the predictions lead to adverse decisions to provide a pragmatic definition of predictive privacy and a practice-based focus for this paper. I intend to focus my argument on data processing systems that have tangible *effects* in the world (e.g., in terms of discrimination), instead of stipulating an idealistic value of predictive privacy. For this purpose, I omit the more theoretical and abstract question of whether simply predicting something about somebody without *any* action following from it would also constitute a violation of predictive privacy.[5] Thirdly, in many cases, violations of predictive privacy do not materialize as a set of information that *someone* learns about someone, but as an immediate *effect of treating individuals or groups differently* in (automatic) interactions. Thus, the perpetrators of predictive privacy violations do not have to be (human) subjects; in general, violations of predictive privacy are attributed to the predictive system, as a whole. Even if no information is learned by anybody, it is the negative effects on individuals due to differential treatment that count for a violation of predictive privacy.[6]

Relation to other privacy conceptions

While fully anchoring the concept of predictive privacy in a theory of privacy is beyond the scope of this applied ethics paper, I will briefly relate the notion of predictive privacy to existing concepts in the recent literature.

Loi and Christen (2020) propose the notion of "inferential privacy" in their discussion of "group privacy":

> The inferential privacy of an entity (individual or group) X, is a measure of the logically valid inferences, about the sensitive features of X, that cannot be made about X, based on the available data about X. (Loi & Christen, 2020, p. 218)

With this definition, the authors address the privacy implications of "generalizable knowledge" such as statistical inferences, including the concern "that the privacy of an individual [can be] infringed through inferences made by virtue of data about other individuals" (Loi & Christen, 2020, p. 209). Although the concept of "inferential privacy" thus addresses a problem similar to that of the present paper, it falls one step short of the concerns I raise: predictive privacy does not restrict the cause of the privacy violation to "logically valid inferences" , but includes the "non-logical" *predictions* that result from applying statistical inferences to single cases, thus leaving the grounds of "[statistical] logic" in favor of *betting* on the possible outcome. In other words, predictive privacy is concerned with the material effects of predictions that *are made* in real-life systems, including predictions that are not logical, not justified, and possibly not even correct.

The concern about unjustified leaps from statistical inference to individual predictions might, however, be compatible with Wachter's and Mittelstadt's (2018) theory of a "right to reasonable inferences ". In their project, they also focus on privacy violation from data that are "created through deduction or reasoning rather than mere observation or collection from the data subject" (Wachter & Mittelstadt, 2018, p. 22). According to my analysis, the normative requirement that the creation of such data must always proceed "reasonably" implies that making predictions must be avoided; thus, respecting predictive privacy implies maintaining a right to reasonable inferences. The contrary, however, is *not* true, because protecting predictive privacy implies that even the use of "reasonably inferred" information might pose an ethical concern. In fact, at the heart of the notion of predictive privacy lies the concern that it is generally ethically wrongful to treat people on the basis of predictions, even if those predictions *are* reasonable. Consequently, demanding to protect predictive privacy is stronger than the right for reasonable inferences.

A common reference point to these debates is the concept of "group privacy" most famously proposed by Luciano Floridi and others (Floridi, 2014; Taylor et al., 2016). Mittelstadt (2017) applied this concept to predictive analytics. The starting point of group privacy here is that the differential treatment of individuals by predictive analytics often proceeds by "linking individuals into groups or classes of interest to the platform" (Mittelstadt, 2017, p. 475). The members of such profiling groups could be treated in a way that one would want to combat as a case of discrimination, compared to members of other profiling groups. Existing

[5] Note that my Definition 1 does not preclude conceptualizing predictive privacy in such a way that predictive privacy is also violated by predictive information that is not acted upon.

[6] Negative effects, of course, *could* include damage to informational self-determination if the predicted information is learned by an actual human or otherwise processed and circulated as information.

legal instruments, however, do not protect members of profiling groups from any resulting discrimination or privacy violation as those "algorithmically assembled group[s] ... [do] not necessarily align with [named] classes or attributes already protected by privacy and anti-discrimination law" (Mittelstadt, 2017, p. 475). The solution proposed by group privacy is then to attribute an own privacy right to the ephemeral, invisible, algorithmically determined "ad hoc groups" built by PA.[7] This approach, which is an important contribution to extend the scope of anti-discrimination legal instruments, differs from the ethical approach I am bringing forward in this paper. Predictive privacy does not tie the ethical concern to the precondition that mistreatment of an individual occurs at group scale. Predictive privacy is not looking at discriminated groups as "rights-holders" (Mittelstadt, 2017, p. 484), but at the differential treatment of individuals that is facilitated through leveraging millions of lateral comparisons with other individuals in the group of data donors. A group aspect is involved in predictive privacy, but in a different way: The collective of data donors goes from being rights-holders to *duty-bearers*. Predictive privacy makes it a duty for all of us, both in our roles as users and citizens, to ensure that no detrimental treatment of *others* is facilitated through the data (including de-identified data and usage data) that *we* submit to platforms and digital services.

Ethical evaluation of predictive systems: a step-by-step guide

In order to assess the ethical implications of predictive systems and precisely locate the violation of predictive privacy in these complex assemblages, I propose the following minimal model of the data processing cycle of predictive systems.

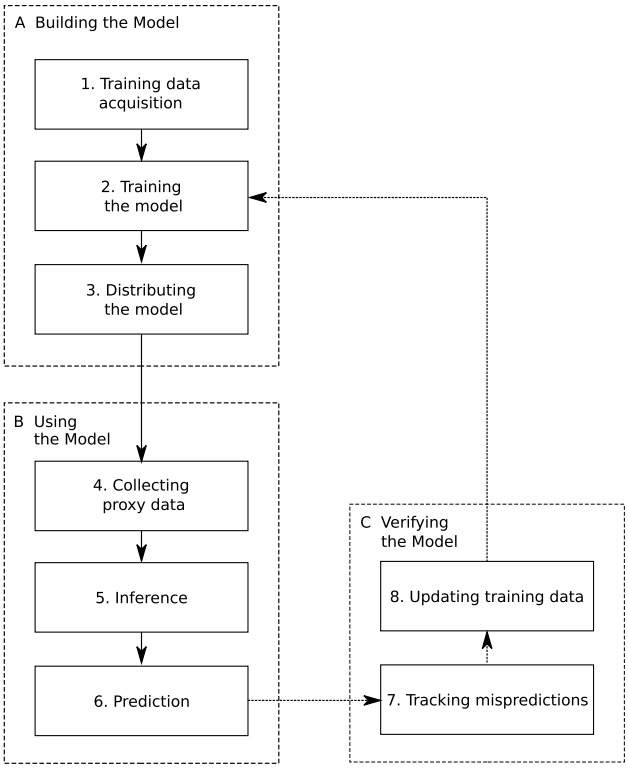

Fig. 2 Minimal model of typical data processing cycle for PA. Dashed lines: additional steps (7 + 8) and feedback loop to reduce Type A unfair bias (cf. the section "Collective ethical concerns")

The 6-step minimal model

Imagine the goal is to train a predictive model to predict or classify individuals i with respect to a certain target attribute or cost/risk/utility function, A. The data processing cycle that is necessary to build and use such a model based on data analytics and supervised ML algorithms can be broken down into the following minimal chain of steps (see Fig. 2):

(A) Building the model (training phase)

(1) *Data acquisition* a data set W covering a large number of users, individuals or cases is collected. The target variable A needs to be known for the data subjects in W to implement supervised learning. Typically, entries $(D_i, A_i) \in W$ for individuals i cover a long list of supplementary data fields D_i such as behavioral and usage data that could potentially be relevant as proxy variables to predict the target, A_i. As one does not prescribe to the learning algorithms which fea-

[7] Group privacy follows a similar argument to Anton Vedder's in his proposal of a concept of "categorical privacy" (Vedder, 1999). In many ways, categorical privacy is the closest precursor to what I propose here under the name of predictive privacy: Vedder has in mind discrimination and differential treatment based on aggregate information derived from de-identified personal information of many data subjects; and he points out the limitations of data protection regulation which ceases to apply "as soon as the data has ceased to be personal data in the strict sense", e.g., after de-identification. But his paper is focused on "data mining" technology before machine learning, which is reflected in his notion of "categorical privacy" as a form of privacy violation based on relating the target person to a "reference group" of data subjects who share most attributes. Predictive analytics, I argue, does not necessarily sort people into groups of *similar* others, but works with similarities *and* dissimilarities to *all* the others, or degrees of similiarity/dissimilarity. Thus, there is in general no "reference group" in predictive analytics, which makes it necessary to focus on the prediction (rather than grouping) aspect in the analysis of privacy threats.

tures (data fields) are relevant in predicting A_i, one endeavors to include as much data in W as are possibly available.

(2) *Training* W needs post-processing, normalization and clearing to ensure data quality (McCue, 2007). Subsequently, a predictive model P_W is trained by learning to predict A_i for every data pair $(D_i, A_i) \in W$. This step requires complex manual and often intuitive skills or even trial-and-error on the part of the model operators, for example, to avoid under- or over-fitting of the model. Depending on the learning algorithm, some part of W might be set aside to verify the predictive accuracy of the model in the course of training.[8]

(3) *Distribution* once the model P_W is trained to sufficient accuracy, the training data set W could be discarded. P_W can now be distributed (sold, published, deployed, pushed to client devices, etc.) for the prediction phase.

(B) Using the model (Prediction phase)

(1) *Collection of proxy data* to predict A_i for a concrete data subject i (e.g., a new customer, applicant, case, ...), proxy data D_i are actively collected about i from available sources. In some cases, D_i covers all the data fields that were available in the training data W (minus A_i, which is assumed to be unknown), in others, D_i might only cover a subset of the most significant predictive features.

(2) *Inference* P_W is applied to the proxy data D_i in order to obtain a statistical inference $A_i = P_W(D_i)$ to predict the target parameter for the data subject i. A_i is a *probability distribution* on a space of possible values of A.

(3) *Prediction* the inference A_i, which is a probability distribution in nature, is post-processed to obtain an actionable prediction $\bar{A}_i$. This, in turn, leads to a decision or action that might treat the data subject i differentially (e.g., an insurance premium is adjusted, a credit is denied, a recidivism risk is deemed too high, etc.).

[8] Generally, "predictive accuracy" is a complex matter as there are multiple metrics from sensitivity, specificity, positive and negative predictive value (classification algorithms) to logarithmic loss, F1 score, mean square error and many more. Adequate choice depends crucially on the context. Cf. Goodfellow, Bengio and Courville (2016).

Ethical discussion step by step

The different steps in this generic data processing cycle are linked to different ethical issues. Some of them are pragmatic, technical and operational, others are fundamental and lead to open philosophical and political debates. I will go through them one by one:

Step 1: data acquisition

- When collecting training data from users, the communication design in asking for consent is an ethical concern. Complicated terms and conditions might conceal the purpose and extent of data processing; UX-tricks in the design of confirmation dialogues might be used to increase the likelihood of uninformed and unintentional consent (cf. Nissenbaum, 2011; Mühlhoff, 2018).
- Promises about anonymisation of training data will make many users feel that the data processing is harmless, while they do not understand that they are contributing their data to a model that allows the prediction of sensitive information about *other* users, including those who would not consent to processing of their information.
- Further ethical questions arise if the predictive model is trained from data that is being re-purposed from other contexts. Also, parts of the training data (e.g., some additional features in the vector D_i) might be bought from (illegal or grey zone) data brokers without the users' knowledge.
- Other issues outside the scope of this paper concern secure data storage and the risk of re-identification in de-identified data.

Step 2: training the model

- Data post-processing is critical to reduce unfair bias. Bias can arise from misrepresentations of demographic groups in the data, from the mechanisms, places and social channels of data acquisition, from biases in society reflected in the data, from a lack of accuracy and completeness of the data, as well as from unconscious biases of the developers and operators of the data processing system (Friedman & Nissenbaum, 1996; Coeckelbergh, 2020a). An apt choice of post-processing methods can control for *some* of these sources of bias.
- The choice of metrics for the predictive accuracy of the model and, in fact, the choice of a training goal itself, can also be biased. For example, the reduction of false positives and false negatives can have a cancellation effect; optimization for predictive accuracy requires different measures in majority and minority segments. This is both a matter of priorities in the development of ML mod-

els and of awareness about biases by the designers and operators (including their own biases).
- There will be no technical solution that eradicates all bias. Trade-offs have to be made that are inherently political (Coeckelbergh, 2020a). Making these trade-offs and the underlying choices transparent is an ethical concern.

The ethical questions mentioned so far mostly coincide with ongoing discussions on predictive models and algorithmic decision making (see Mittelstadt et al., 2016; EU High-Level Expert Group on AI, 2019). In many of these cases, we already have clear ethical principles, if not legislation, which prescribe the common ethical judgments on these issues: It is understood that bias is one of the most salient problems in PA and ML (cf. Mittelstadt et al., 2016; Coeckelbergh, 2020a), that data re-purposing is forbidden (EU GDPR), that consent is to be based on information and data are to be protected against theft and misuse etc. (cf. Coeckelbergh, 2020a). I will therefore swiftly proceed to the ethical issues connected to steps 3–6. They are, for the most part, less debated, less regulated by law, and even philosophically open questions.

Step 3: distributing the model. Should trained models be free for sale?

The model consists of data (e.g. the matrix of weights of the trained neural network) that are derived from training data. If we assume that it is not possible to reconstruct training data from the model parameters,[9] does this automatically mean that distributing the model is less ethically sensitive than processing the training data?

Addressing this question, an individualistic perspective does not suffice. While information about individual data donors might not be revealed by distributing the model, the model comprises what Loi and Christen (2020) call "generalizable knowledge": it contains information on what the relevant features are to predict the target variable for *any* individual. This information was learned from the training data and reveals potentially sensitive information about the underlying *population*. The question is whether "we" as a group of everyday, involuntary data donors, or as society as a whole, want these insights from our aggregate data to be sold or made publicly available to others? This question depends highly on the domain of application and no general judgment is possible without a case-based and often political debate. Medical results, for instance, such as "smoking correlates with risk of cancer" , are an example of a beneficial generalizable knowledge gained from a statistical model using many case histories as underlying data. On the other hand, if a model can predict (or claims to predict) a rare, severe and incurable genetic disease from easily available proxy data, is it right that by publishing the model everyone with access to the proxy data about a concrete individual i will be able to predict this condition about i? Not only could this be abused by insurance companies or employers, but the affected individuals might not even know yet, or ever want to know, about their condition.

The trained model reveals information (statistical inferences) about the underlying cohort of training data donors that can, moreover, be used to predict potentially sensitive information A_i about any third individual i who did *not* contribute to the training data, who might not want A_i to be (anonymously) known about them and who might not be aware that by the (proxy) data they provide in a certain context, A_i could be predicted. Protecting predictive privacy is therefore a collective task, as it relates to the utilization of training data collected from a large cohort. This cannot easily be reduced to a mere sum of individual rights and responsibilities. Using and/or distributing predictive models thus breaches what Mantelero (2016) calls a "collective non-aggregative interest" in preserving our predictive privacy. However, as the examples show, protecting this non-aggregative interest can hardly be formulated as an absolute right as it has to be balanced against many beneficial applications of PA. Ethical evaluation depends on the concrete application. In general, it seems more interesting to debate *where* such a model may be circulated and used, and *who* exactly would benefit from it, rather than *if* it can be done (cf. Coeckelbergh, 2020a).

Step 4: how to collect proxy data ethically?

The "standard" objections listed in step 1 also apply to step 4, when proxy data about the target individual is collected. Additionally, however, there are more specific concerns:

Typically, the target variable A_i is a sensitive attribute or an unknown critical parameter (e.g. utility/risk/cost function), while proxy data D_i are either less sensitive or easier to obtain from the target individual i at the current stage in the process. D_i are often usage and behavioral data that, if collected pseudonymously, are not required to be classified as personally identifying information under data protection legislation. Nevertheless, critical information A_i can be predicted from D_i and even if the target individual consents to the collection of D_i, he/she has not consented to the collection of A_i. Users' awareness of the processing of their data must therefore be expanded so that they understand what information will be *derived* from the data they provide. In particular, when sensitive attributes are derived from other information, this should be named.

[9] This is non-trivial, but feasible, e.g. using state-of-the-art privacy techniques such as differential privacy. Cf. Dwork (2006) and Abadi et al. (2016).

This ethical device gets more complicated if the real-world meaning of the target variable A_i is obfuscated. For instance, it might not be termed "likelihood to develop cancer" but "insurance risk coefficient". This makes it seem that no sensitive attribute inherent to the data subject is being predicted, but rather some parameter internal to the business model. Although these parameters might be highly correlated or even be the same (just under different names), it is to be expected that end users perceive information processing as less critical in the latter case. An ethical requirement, therefore, is to disclose *faithfully* to the user the full possible impact of the information that is being derived from D_i.

It might even happen that the output A_i of the predictive model is not an intermediate piece of information, but directly a yes/no decision in a certain real-world process (conflation of steps 5–6). In this case, the sensitive information that is *implicitly* being predicted from D_i remains opaque. Ethical behavior in this data processing step is thus a matter of honest intentions, fair treatment and "answerability" (Coeckelbergh, 2020b) in the communication relationship with the user. Legal and policy instruments are of limited effect in this respect, as operators have considerable opportunities to sneak through the regulatory grid by designing the human-computer interaction in a way that re-frames and re-phrases what is actually being done to make it seem harmless (Mühlhoff, 2018). External ethical auditing instead of mere regulation is, therefore, necessary to enforce ethical compliance.

Step 5: inference

In this step, a statistical inference is derived from proxy data about a data subject but the knowledge is not yet acted upon. Separating this step from step 6 is interesting both analytically and pragmatically. Analytically, this is the place for a range of theoretical ethical questions: Is the mere generation of statistical inferences about an individual unethical? That is, provided that we already took care of all ethical concerns up to this step, does deriving information from proxy data comprise an *additional* violation of anyone's rights, qua its nature as mere derived information, even if nothing happens in consequence? Or is the crucial ethical threshold only when the derived knowledge leads to actions and decisions, or when it circulates as information about someone that might shape perceptions, reputations, discourses etc.?[10]

It is important to bear in mind that the knowledge A_i is probabilistic and empirical knowledge. There are specific ethical implications tied to both the "probabilistic" and the "empirical" in this qualification. Connected to the "empirical", there is the risk of inductive fallacy: any inductive knowledge, generalizing from n observed cases (here n is the number of cases covered in the training data W), might fail on a new, $n+1$th, case. In the case of PA, inductive fallacy as a source of error occurs because the model P_W constitutes a generalizable knowledge about the cohort covered in the training data W. This knowledge is transferred, in step 5, to a new case that is potentially not included in W. (On the ethical implication of the "probabilistic" nature of A_i, see step 6.)

In pragmatic terms, the ethical concern in step 5 is that the statistical inference A_i be communicated to the data subject before, or independently of, any action being taken on it (step 6). Specifically, A_i should be made visible *as a probability distribution*. Making the probabilistic nature of the derived information transparent is of significant explanatory value to data subjects as this is a crucial step in the reasoning of PA (see discussion of step 6). This would also open the possibility to challenge the predictive outcome before it is turned into a prediction and, subsequently, a decision.

Step 6: the prediction gap

A severe, arguably even the *main*, ethical challenge of predictive systems is connected to the *probabilistic* nature of statistical inferences: In general, the result $A_i = P_W(D_i)$ of step 5 is not a single value but a probability distribution on a range of possible outcomes. In the case of a classifier (P_W sorting i into categories), inferred class membership could come with probability weights; if continuous values are predicted (such as credit scores or the likelihood to develop a disease), such inferences come with confidence intervals. In the transition from step 5 to 6, when inferences are turned into predictions, this uncertainty often gets disambiguated as the probabilistic nature of A_i needs to be broken down to *one* actionable result that feeds into a decision or differential treatment: The credit score might be taken as a single scalar value regardless of its statistical confidence; the category with the highest probability might be taken as *the* category of i in the case of classifiers. This reduction of statistical uncertainty means betting on one specific outcome next to all the other possible outcomes. This step is a significant source not only of error but of a specific violation of human dignity committed by predictive systems as it disambiguates diversity and uncertainty to make the target individual 'fit into an actionable category'. I refer to this disambiguation step as "crossing the prediction gap".

Conceptually, crossing the prediction gap is the moment where statistical inference is turned into a prediction that is implemented in a decision. As described in the section "Predictive systems and predictive privacy", statistical inferences always refer to populations. The information stored in A_i reads something like: "Within the cohort W, a case with

[10] See for an in-depth ethical debate of this general question (Basu, 2019).

data signature D_i belongs in 60% of cases to category A, in 30% of cases to category B, in 15% of cases to category C, ..." (in case of a classifier). Thus, A_i *is rather an information about the aggregated data W than it is an information about the individual i*. In step 6, however, the information that really refers to the statistical composition of the cohort W is projected on an individual case and turned into non-probabilistic knowledge. (It is irrelevant to the argument whether the individual i is a member of the cohort W.) Notice how this problem is different from inductive fallacy (see step 5). It raises the fundamental question of whether it is ethical to disambiguate individualized statistical inferences $A_i = P_W(D_i)$ in decision making, which basically amounts to transforming statistical patterns within a certain group of training data subjects into a judgment about an unrelated individual.

Pragmatically, visualizing the probability distribution A_i independently of its disambiguation $\bar{A}_i$, as recommended in step 5, is a contribution towards more transparency facing the ethical problem of crossing the prediction gap. In fact, both A_i and $\bar{A}_i$ should be communicated to the user in the interaction design.

There are further ethical questions relating to this point:

- Should crossing the prediction gap be connected to some ethical requirements concerning the mathematical features of the probability distribution A_i? For instance, one could demand a certain threshold regarding the confidence interval of a scalar predictor (utility/risk/cost function) or that the probability weight of one predicted class stands out from the others by a certain, significant margin (classifier)?
- Connected to this, we could articulate the ethical requirement that decision routines ought include 'non-predictability' as a possible output that stops the automated consequences (instead of, for instance, acting on the most pessimistic end of the predicted spectrum). 'Non-predictability' would amount to halting the process because the prediction gap cannot be crossed "ethically".

In the Conclusion, I will come back to the prediction gap to argue that no technical solution can be expected to resolve the fundamental threat to human dignity and autonomy that arises when aggregate inferences are turned into individual predictions. In the end, this is a question of how "we" as members of liberal democratic societies want to be treated by one another (see also Basu, 2019).[11]

The discussion in this section shows a range of ethical questions and demands linked to the preservation of predictive privacy. Most notably,

- Preservation of predictive privacy is not currently guaranteed by existing legal frameworks of data protection and requires a collectivist conception of privacy because target individuals' privacy is violated using data collected from *other* individuals (cf. Wachter, 2019; Wachter & Mittelstadt, 2018; Zarsky, 2016);
- Making predictive privacy a universal right enshrined in data protection and human rights legislation would emphasize that potential violations of predictive privacy must be negotiated and weighed against other rights and values, such as safety and freedom, in any specific context. This balancing will often be a political decision; but we would never even have that political debate unless predictive privacy is recognized as part of what is understood to be a right to privacy in a particular legal system.
- Violations of predictive privacy proceed by transferring behavioral patterns from cohorts to target individuals. No *stored* data about the target individual are *leaked* in a violation of predictive privacy. This is why protecting predictive privacy confronts us with a completely new ethical problem of informational privacy, which relates to "crossing the prediction gap".
- Reaching for sufficient transparency depends to a high degree on the good faith of PA operators, creating a significant legal enforcement challenge.

For these reasons, the protection of predictive privacy indeed requires an ethical debate and is irreducibly a matter of collective moral behavior in a highly capitalized and interest-driven context. It cannot be fully delegated to better regulation, although regulation should be updated to the specific challenges originating from Big Data analytics (cf. Wachter, 2019; Zarsky, 2016).

[11] Basu (2019) points to the literary figure of Sherlock Holmes as exemplifying behavior guided by a "kind of morally objectionably statistical reasoning" (Basu, 2019, p. 6) that closely resembles what I describe as the reasoning of predictive analytics. Basu argues that predictive statistical reasoning is "a way of looking at another person not as a person, but as an object that is determined by causal law, as something whose behaviour is to be predicted" (Basu, 2019, p. 8). She makes the compelling and much overlooked ethical point that this kind of wrong treatment begins at the level of *thought*, or at the *epistemic* level of what we believe of one another: people "can also be wronged by *what is believed of them*" (Basu, 2019, p. 2, emphasis in original). For this reason, the ethics of predictive privacy is closely related to the epistemological problems of a knowledge culture that increasingly relies on predictive reasoning.

Collective ethical concerns: from predictive privacy to unfair bias and discrimination

One of the main concerns related to automated decision making based on PA is the potential contribution of this technology to stabilizing or even increasing social and economic inequalities and power differentials within societies (Amoore, 2020; Mühlhoff, 2020a; O'Neil, 2016). Often, the underlying ethical problems of this societal effect are discussed using the terms "unfair bias" and discrimination (cf. Friedman & Nissenbaum, 1996; Barocas & Selbst, 2016; Mittelstadt et al. 2016). Extending on the data processing minimal model presented in the previous section, I will now discuss ethical requirements to PA operators that contribute to reducing unfair bias and discrimination.

There is a negative and a neutral meaning to "discrimiation" : the negative connotation implies the unfair treatment of certain (protected) societal or demographic groups, while the etymologically more original and neutral meaning refers to the act of distinguishing different cases and possibilities from one another. Notably, PA is *made to discriminate*—in the neutral sense of the term: It is PA's exact purpose to draw distinctions (based on data) and to be biased towards those features that correlate with a fixed target variable. This is why the attribute "unfair" is often prepended to "bias" and "discrimination" to enable a critical discussion of bias that does not question the general viability of PA technology. The ensuing discussion relies on the implicit assumption that discriminating between alternatives using PA is not objectionable per se; yet, within this general approval of PA, "unfair" forms of discrimination are to be avoided (cf. Amoore, 2020). I will follow this (questionable) assumption *only in the present section* in order to make a contribution towards operational measures against unfair bias. The general ethical question of whether we should discriminate using PA *at all* is discussed in the next section.

Two types of unfair bias

Speaking of *unfair* bias and discrimination does not yet solve the problem of defining what makes a discrimination unfair (Coeckelbergh, 2020a). To elucidate this, we may differentiate between two types of unfair bias (that may also be combined):

Type A unfair bias, or "misprediction bias", boils down to a systematic *misprediction* of a group of cases resulting from a malfunctioning of the predictive system. In this case of unfair bias, there is a group of individuals I with proxy data signatures D_i, $i \in I$, whose predictions $A_i = P_W(D_i)$ tend to miss the real values of A_i by a significant margin. The members of I are, therefore, going to be treated in a way that "does not fit them" and "does not do them justice"—as evaluated *within* the logic and criteria of the predictive system itself. Type A unfair bias comprises an adverse effect even from the perspective of system operators themselves: the system "misses certain opportunities" and should be improved. Sometimes, however, system operators do not prioritize eradicating all kinds of predictive errors equally, as costs and awareness of different kinds may vary. This is an important indirect cause of Type A unfair bias. For instance, in the case of hiring algorithms it is much easier in terms of system design to reduce false positives (wrong hires) than false negatives (missed hires), see below. In the case of gender misprediction in facial analytics brought forward by Joy Buolamwini, we have an misprediction bias that resulted from unawareness in the engineering teams and biased training data (cf. Buolamwini & Gebru, 2018).

Type B unfair bias, also called "biased prediction" , is not connected to misprediction but to effects of "accurate" predictions that are outright unethical or socially and politically undesirable from an overarching standpoint. PA tends to cluster individuals by dis-/similarities that sometimes correlate with structures such as gender, ethnic background, class, wealth, social status or level of education. This potentially results in a perpetuation of existing discriminatory patterns and social inequalities when predictive systems are deployed to treat these clusters differently in terms of access to resources, information, education, etc. It is important to notice that this kind of unfair bias has, in general, *no technological solution* (cf. Amoore, 2020; Coeckelbergh, 2020a). Recognizing, fighting and eliminating this kind of unfair discrimination is an ethical and political goal that might go against the interests of operators because it will require trade-offs between "efficiency' of the predictive system (internal goal) and avoiding this type of unfair discrimination (collective goal). Therefore, Type B unfair bias must be politicized in order to impose constraints on self-interest driven operators in the name of equality as a higher, collective goal.

In many real examples of unfair bias, Types A and B are combined. Still it is of use to tell these two dimensions apart from each other analytically, for instance, when tech discourses are quick to claim that they can "fix" a certain instance of unfair bias. Then it is important to know that they will only be able to fix the Type A and hardly the Type B component of that bias; hence the distinction shows the intrinsic limitations of technical approaches.

There is extensive philosophical discussion of Type B unfair bias for instance in the ethics of algorithms (cf. Mittelstadt et al., 2016). In this section I will exclusively address Type A bias, as I intend to discuss ethical imperatives to PA operators to reduce unfair bias. By that choice I do *not* suggest that technical solutions will generally suffice in tackling

unfair bias. They might allow to reduce Type A bias to some extent, but reducing Type B bias will always be a matter of political will and regulation. It cannot be left to technological solutionism.

Extending the minimal model

A key ethical demand to counter the problem of Type A bias (systematic misprediction) is the constant verification and re-training of predictive models in the prediction phase through the implementation of feedback loops (cf. O'Neil, 2016, pp. 148–155). This means: staying in touch with the target individuals to *detect* mispredictions and *feed* this information *back* as training data to calibrate the model. In order to achieve this, the data processing cycle proposed in the previous section must be extended in a circular way, as follows:

⋮ (appending the list of steps from the previous section ("Ethical evaluation of predictive systems"), cf. Fig. 2)

(C) Model Verification/Feedback Loop:

(1) Appropriate measures are to be used, and a persistent relationship with the target individual i is to be built, so that mispredictions can be detected. That is, retrospectively, for target individuals i, the real value $\hat{A}_i$ of the predicted parameter $\bar{A}_i$ is to be measured so that the difference (or distance) $|\hat{A}_i - \bar{A}_i|$ can be used as an indicator of the quality of the prediction.
(2) Knowledge about the quality of predictions is to be re-inserted in → step 2 of the Training Phase. That is, the data pair $(D_i, \hat{A}_i)$ is to be added to the training data W to re-train the model on the extended data set. This closes the feedback loop to constantly update the model.

There are a number of limitations and objections to this principle. First, as mentioned above, it is not guaranteed that this procedure will eradicate all unfair bias, not even that of Type A. For instance, if biases result from developer and operator blind spots or are passed on from society through training data, they may not be eradicated by this method. Neither is this procedure itself ethically unproblematic, as tracking users to detect misprediction from their future performance raises ethical and privacy concerns in and of itself. Secondly, detecting mispredictions can be impossible if decisions that result from the prediction prevent the target individual to act contrarily: if, in criminal recidivism prediction, a defendant is predicted to re-offend and therefore stays in prison, it is not detectable whether this prediction was wrong. In such cases, it is a strong ethical imperative to verify the predictive model with more sophisticated techniques. For instance, a kind of A/B testing approach could be adopted: in a representative and random subgroup of cases, decisions of the predictive system could be calculated but not acted upon in order to compare predictions to the real outcome of such cases.

Thirdly, the possibilities and appropriate mechanisms for implementing feedback in the data processing life cycle is highly domain specific. Let me briefly illustrate this using further examples: In the case of **hiring algorithms**, keeping track of rejected individuals is very hard and undesirable in terms of privacy and data protection. Using false negatives (wrongly rejected applicants) to re-train the model is therefore difficult, while data on false positives is easily available. This will result in a *bias towards eradication of false positives*. The fact that employers usually have more incentives to reduce false positives than false negatives further adds to this imbalance. Altogether this presents hiring algorithms as an extremely problematic kind of technology ethically, as it is hard to build such systems fairly (cf. Sanchez-Monedero, Dencik, & Edwards, 2020).

In **criminal recidivism scoring**, keeping track of target individuals is generally not difficult. Detecting false negatives is easy (defendant will re-offend and thus be treated by the justice system again). False positives are hard to detect as imprisoned defendants will not have an opportunity to show that they would not re-offend. Given the strong life-changing consequences of recidivism prediction, an in-depth ethical debate on this technology is needed with strong demands towards control of mispredictions. This imbalance in detecting mispredictions between false positives and negatives makes alone this technology highly questionable on principal grounds in terms of fairness and human dignity (cf. Fry, 2018).[12]

Credit scoring is particularly tricky because there are significant differences in the views of society, target individuals, and operators about what counts as a misprediction. While the targeted Individuals are concerned about having their credit applications rejected or being offered a disproportionately high interest rate, the benefit to which the operators will optimize their decision-making procedures

[12] The ProPublica investigative analysis of biases in the US recidivism prediction system "COMPAS" (Angwin et al., 2016) shows how these differences in controlling for mispredictions may be interwoven with social structures such as race: as the investigation shows, the rate of false positive predictions (wrong prediction that an individual will re-offend) is disproportionately higher for black compared to white defendants; the rate for false negative predictions (wrong prediction that someone will *not* re-offend) is disproportionately higher for whites. This is illustrative of how predictive systems can be embedded in social reality in a way that the externalities of predictive errors are disproportionately carried by social groups that are already discriminated.

is a mix of profit maximization, acceptance rate maximization (competitive advantage over other credit providers), and credit risk control. Hence, offering disproportionately high-interest rates to certain demographic groups who are disadvantaged concerning access to credit and do not have the means to defend themselves against this form of exploitation will not be perceived as misprediction from the perspective of operators. In credit scoring, it is therefore an illusion that PA will only be used for the 'legitimate purpose' of risk control towards investors, while it is highly likely that PA models will learn from the data how best to exploit specific psychological, (sub-)cultural and socioeconomic vulnerabilities of specific demographic groups. Implementing feedback loops can be expected to even *further* this learning effect of predictive systems, particularly with regard to making the system learn how 'bold' an offer can be in each individual case before customers turn away. Therefore, in the case of credit scoring, there is a high ethical demand for external auditing and regulation. Feedback loops that control for unfair discrimination should be implemented and supervised by external instances which should additionally act as data trustee, for it is undesirable from a collective point of view to make data of turned-away customers available to the banks.

Conclusion and outlook

We have seen in the previous section ("Collective ethical concerns") that PA is *made* to discriminate—if discriminating means to mark the distinguishing or peculiar features of something. The discussion of unfair bias, which is popular in criticism of AI and data analytics today runs the risk of implicitly endorsing the fact *that* PA is already being used and will continue to be used, thereby acquiescing to focus only on how the technology should be made "fairer" (cf. Amoore, 2020). But since this is not just a matter of technological improvements, we should also ask the fundamental question of whether "we" as democratic societies want this technology to be used at all. Of course, given the already ubiquitous deployment of PA, this question might seem abstract and out of touch with reality. All the more we should ask ourselves this question to finally find more awareness of the potential harms of PA technology and effective regulation to prevent them.

In the section "Ethical evaluation of predictive systems", it turned out that a specific and qualitatively new ethical concern about PA is related to *crossing the prediction gap*: The leap from aggregate statistical inferences to individual predictions implies that individuals are being judged by behavioral comparisons ("pattern matching") to all other individuals within the pool of training data subjects. The result $A_i = P_W(D_i)$ of the PA for proxy data D_i is knowledge about the cohort (W) and not only about the individual i. i is thus judged on the basis of similarities and dissimilarities with the individuals covered in W. This step is not only the source of unfair bias, but fundamentally raises the question of whether we should ethically and politically allow individuals to be "locked" into a logic of behavioral comparison and "pin-pointed" to what appears to be their most likely future behavior or performance. Arguably, individuals are stripped of their autonomy and dignity when they are judged according to a "people like you"-scheme (O'Neil, 2016) that, like Sherlock Holmes (cf. Basu, 2019), suggests that "we know already what you are like".[13] Since the result A_i of the PA is a probability distribution, acting on that basis in automated decision-making often means disambiguating it by betting on the most likely among possible outcomes. In practice, this means that people are preemptively treated *as if* they already reveal a certain attribute, thereby depriving the target individual of the principle diversity and self-authorship of choices, opinions and behavior. This form of treatment, when implemented on a large scale in automated systems, leads to social stratifications and chasms, as *predictive knowledge has a "performative effect" on social reality* (cf. Matzner, 2016): Predictive systems produce and stabilize precisely the kinds of social differences and inequalities that they claim to merely detect in the world (Amoore, 2020; O'Neil, 2016; Mühlhoff, 2020a).

In future research, it would be fruitful to reach a better understanding of how PA technology is embedded in a technological culture of our time that might be called "digital behaviorism", characterized by a transition from a statistical to what might be called a *predictive epistemology*. Notably, this culture is not likely to respect and support a fundamental unpredictability, self-authorship and natural contingency of peer behavior. Instead, my analysis suggests that the implicit techno-cultural mindset (or should we call it an "ideology"?) today rests on the following tacit assumptions:

1. Behavior can be understood in terms of statistical laws using data from digital interfaces (Data positivism).

[13] In a similar argument, Vedder (1999) speaks of a "deindividualization of the person": "Persons are judged and treated more and more as *members of a group* (i.e., the reference group that makes up the data or information subject) rather than as individuals with their own characteristics and merits" (Vedder, 1999, p. 277, emphasis in original). However, my ethical argument does not reduce to people being treated as members of groups. Vedder's analysis of data mining in the era before machine learning always relates the individual to a pre-determined "reference group" of supposedly *similar* individuals. This is not necessarily what contemporary predictive analytics does, whose models *learn* relevant combinations of predictive features that generally do not map to meaningful social groups that could be used as empirical reference groups. Put another way: the models I have in mind do not put people into "buckets" , but they compute *individual* predictions from similarities *and dissimilarities* with *all* the individuals captured in the training data.

2. It is only a matter of having enough data points to obtain accurate knowledge of these laws that covers all relevant social phenomena and variations that might ever occur (Closure to emerging diversity and to the future).
3. Once this Big Data threshold is reached, statistical laws can be epistemologically short-circuited into individual predictions, i.e., the statistical nature of the laws (inherent reference to cohorts) is tacitly *replaced by a predictive interpretation of those laws* (Disappearance of the prediction gap).

Future ethical research, as well as political debate, needs to critically scrutinize how digital networked media routinely apply an epistemology of "herd patterns" , "swarm principles" and "the law of the big number" to the will and behavior of human beings. The apparent conflict between this epistemology and the idea of autonomous, self-determined and self-responsible human beings points to a fundamental challenge to human dignity that is not yet fully understood. As the European Commission's High-Level Expert Group on AI declared, human dignity demands that data subjects should be treated "with respect due to them as moral *subjects*, rather than merely as *objects* to be sifted, sorted, scored, herded, conditioned or manipulated" (EU High-Level Expert Group on AI, 2019). I maintain that protecting predictive privacy is an inevitable part of making sense of and reasserting human dignity in the age of Big Data and AI by extending the scope of privacy to include the harm we help inflict to *others* through the data traces we leave behind.

Along these lines, the concept of predictive privacy has both a descriptive and a normative intention: I propose the concept as a descriptive tool to raise awareness of a new form of attack on privacy and human dignity. Normatively, predictive privacy sits at the intersection of ethics, data protection and anti-discrimination, based on the ethical intuition that persons are wronged if they are judged by behavioral comparisons with others on facts they would not want to reveal about themselves (cf. Basu, 2019). Raising awareness of predictive privacy in our societies means challenging the *liberal* presupposition underlying the Western notion of "privacy", according to which each data subject should be able to decide for themselves what data they want to disclose about themselves. However, given the challenges of Big Data and AI, data protection is not a private choice. Rather, we need a *collectivist approach to data protection* (cf. Mantelero, 2016). Predictive privacy makes visible and debatable that everyone is potentially affected by the data *others* disclose, and everyone, simply by using everyday networked services (even anonymously), influences the predictive privacy of others as an involuntary data donor. Only by using data from millions of "normal people" who think they have "nothing to hide" can PA algorithms learn what "normal" (translate: "privileged") means so that predictive systems can discriminate against allegedly non-normal, dangerous, sick,... persons. This challenge usually goes under the radar of mainstream privacy discourses, which assume there is always a "hacker", stalker or perpetrator who wants to steal information about you. For a violation of predictive privacy, there does not need to be a person or entity that personally obtains information about the target; it is sufficient if the target to be is treated differently, e.g. by an automatic decision making system.

In forthcoming contributions, I will show how the concept of predictive privacy can inform new approaches to regulation and legislation. Indeed, sharper regulation of Big Data technology is *needed*, as existing frameworks such as the EU GDPR are insufficient in preventing toxic effects of anonymized mass data (Wachter, 2019; Zarsky, 2016). The principle of predictive privacy inspires a regulatory approach that *treats predicted information similar to personal information*. That is, the processing of predicted information could be *generally forbidden*, similar to how the GDPR generally prohibits the processing of personal information unless there is a legal foundation. This would incidentally preserve the benefits of PA, e.g., in medical diagnostics (with patient consent). Pragmatically, it is important that regulation of predictive analytics must aim to limit the *application* of predictive models to human beings, not just the creation of such models. That is, data protection regulation must be explicitly extended to the stages "*after* data collection" (Wachter, 2019; cf. also Vedder, 1999), in which derived data, including trained AI models, are deduced from pools of collected data, and may reveal nothing about the training data donors, but can be used to violate the predictive privacy of *any* target individual—whether or not they are part of the training data.

However, and perhaps most importantly, an ethical awareness of predictive privacy as a fundamental value among a large majority of data subjects is urgently needed as precondition for any successful regulation. Any political answer to the social challenges of Big Data technologies requires a timely democratic debate about the ethical values at stake through our everyday use of networked technology. Introducing the concept of predictive privacy hopefully is as step towards such a debate.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

References

Abadi, M., Chu, A., Goodfellow, I., Brendan McMahan, H., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security—CCS'16* (pp. 308–318). https://doi.org/10.1145/2976749.2978318.

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.

Angwin, J., Kirchner, L., Larson, J., & Mattu, S. (2016, May). Machine bias. Retrieved August 18, 2020, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

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

Basu, R. (2019). What we epistemically owe to each other. *Philosophical Studies, 176*(4), 915–931. https://doi.org/10.1007/s11098-018-1219-z.

Bogen, M. (2019). All the ways hiring algorithms can introduce bias. *Harvard Business Review*. Retrieved April 3, 2020, from https://hbr.org/2019/05/allthe-ways-hiring-algorithms-can-introduce-bias.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency. PMLR* (pp. 77–91).

Chatila, R., & Havens, J. C. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In M. I. A. Ferreira, et al. (Eds.), *Robotics and well-being* (Vol. 95, pp. 11–16). Springer. https://doi.org/10.1007/978-3-030-12524-0_2.

Coeckelbergh, M. (2020a). *AI ethics. The MIT press essential knowledge series*. The MIT Press.

Coeckelbergh, M.(2020b). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics, 26*, 2051–2068. https://doi.org/10.1007/s11948-019-00146-8.

Duhigg, C. (2012, February). How companies learn your secrets. *The New York Times*. Retrieved February 28, 2020, from https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Dwork, C. (2006). Differential privacy. In M. Bugliesi, et al. (Eds.), *Automata, languages and programming: 33rd international colloquium, ICALP 2006*, Proceedings, Part II, Lecture Notes in Computer Science 4052, Venice, Italy, July 10–14, 2006, (Vol. 2, pp. 1–12).

Efron, B., & Hastie, T. J. (2018). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press. https://doi.org/10.1017/CBO9781316576533.

EU High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. Retrieved May 3, 2020, from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Eubanks, V. (2017). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Everitt, B., & Skrondal, A. (2010). *The Cambridge dictionary of statistics* (4th ed.). Cambridge University Press.

Floridi, L. (2014). Open data, data protection, and group privacy. *Philosophy and Technology, 27*(1), 1–3. https://doi.org/10.1007/s13347-014-0157-8.

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

Fry, H. (2018). *Hello world: Being human in the age of algorithms* (1st ed.). W.W. Norton & Company.

Goggin, B. (2019, January). *Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts*. Business Insider. Retrieved February 28, 2020, from https://www.businessinsider.com/facebook-isusing-ai-to-try-to-predict-if-youre-suicidal-2018-12.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning. Adaptive computation and machine learning*. The MIT Press.

Grindrod, P. (2014). *Mathematical underpinnings of analytics: Theory and applications*. Oxford University Press.

Hacking, I. (2016). *Logic of statistical inference*. Cambridge University Press.

Hurley, M., & Adebayo, J. (2017). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, *18*(1), 5.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of USA, 110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110.

Lippert, J. (2014, October). ZestFinance issues small, high-rate loans, uses big data to weed out deadbeats. *Washington Post*. Retrieved March 10, 2020, from https://www.washingtonpost.com/business/zestfinance-issues-small-high-rateloans-uses-big-data-to-weed-out-eadbeats/2014/10/10/e34986b6-4d71-11e4-aa5e-7153e466a02d_story.html.

Loi, M., & Christen, M. (2020). Two concepts of group privacy. *Philosophy and Technology, 33*, 207–224. http://doi.org/10.1007/s13347-019-00351-0.

Mantelero, A. (2016). Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection. *Computer Law and Security Review, 32*(2), 238–255.

Matzner, T. (2016). Beyond data as representation: The performativity of Big Data in surveillance. *Surveillance and Society, 14*(2), 197–210.

McCue, C. (2007). *Data mining and predictive analysis: Intelligence gathering and crime analysis*. Butterworth-Heinemann.

Merchant, R. M., Asch, D. A., Crutchley, P., Ungar, L. H., Guntuku, S. C., Eichstaedt, J. C., Hill, S., Padrez, K., Smith, R. J., & Andrew Schwartz, H. (2019). Evaluating the predictability of medical conditions from social media posts. *PLoS ONE, 14*(6). https://doi.org/10.1371/journal.pone.0215476.

Mittelstadt, B. (2017). From individual to group privacy in Big Data analytics. *Philosophy and Technology, 30*(4), 475–494. ISSN 2210-5433, 2210-5441. Retrieved December 20, 2019, from https://doi.org/10.1007/s13347-017-0253-7.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*. https://doi.org/10.1177/2053951716679679.

Mühlhoff, R. (2018). Digitale Entmündigung und User Experience Design: Wie digitale Geräte uns nudgen, tracken und zur Unwissenheit erziehen. *Leviathan Journal of Social Sciences*, *46*(4), 551–574. https://doi.org/10.5771/0340-0425-2018-4-551.

Mühlhoff, R. (2020a). Automatisierte Ungleichheit: Ethik der Künstlichen Intelligenz in der biopolitischen Wende des Digitalen Kapitalismus. *Deutsche Zeitschrift für Philosophie, 68*(6), 867–890. https://doi.org/10.1515/dzph-2020-0059.

Mühlhoff, R. (2020b). Prädiktive Privatheit: Warum wir alle etwas zu verbergen haben. In C. Markschies & I. Hermann (Eds.), *#VerantwortungKI – Künstliche Intelligenz und gesellschaftliche*

Folgen (Vol. 3/2020). Berlin-Brandenburgische Akademie der Wissenschaften.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32–48.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

O'Dwyer, R. (2018, May). Are you creditworthy? The algorithm will decide. *Undark Magazine*. Retrieved March 10, 2020, from https://undark.org/2018/05/07/algorithmiccredit-scoring-machine-learning/.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Reilly, M. (2017). Is Facebook targeting ads at sad teens? Retrieved August 6, 2020, from https://www.technologyreview.com/2017/05/01/105987/is-facebook-targeting-adsat-sad-teens/.

Rieder, G., & Simon, J. (2017). Big Data: A new empiricism and its epistemic and socio-political consequences. In W. Pietsch, J. Wernecke, & M. Ott (Eds.), *Berechenbarkeit der Welt?* (pp. 85–105) Springer. https://doi.org/10.1007/978-3-658-12153-2_4.

Sanchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to solve the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. Retrieved January 22, 2020, from http://arxiv.org/abs/1910.06144.

Taylor, L., Floridi, L., & van der Sloot, B. (2016). *Group privacy: New challenges of data technologies*. Springer.

Varner, M., & Sankin, A. (2020, February). Why you may be paying too much for your car insurance. Retrieved March 2, 2020, from https://www.consumerreports.org/car-insurance/why-you-may-be-paying-too-much-for-your-car-insurance/.

Vedder, A. (1999). KDD: The challenge to individualism. *Ethics and Information Technology*, *1*(4), 275–281.

Wachter, S. (2019). Data protection in the age of big data. *Nature Electronics,* 2(1), 6–7. https://doi.org/10.1038/s41928-018-0193-y.

Wachter, S., & Mittelstadt, B. (2018). A right to reasonable inferences: Re-thinking data protection law in the age of Big Data and AI. Preprint. LawArXiv. Retrieved December 20, 2019, from https://osf.io/mu2kf.

Zarsky, T. Z. (2016). Incompatible: The GDPR in the age of big data. *Seton Hall Law Review*, *47*, 995.