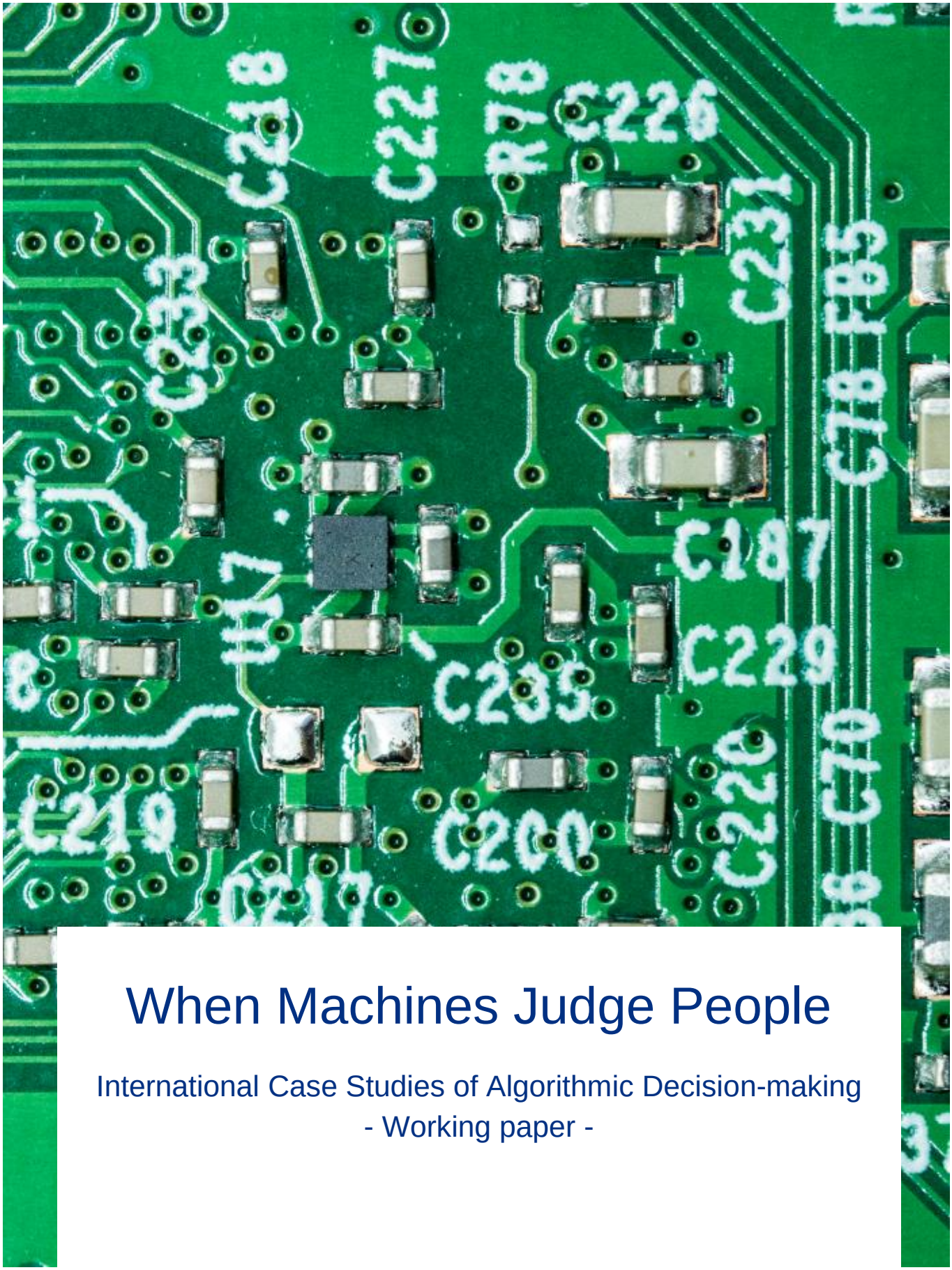

When Machines Judge People

International Case Studies of Algorithmic Decision-making
- Working paper -

BertelsmannStiftung

When Machines Judge People

International Case Studies of Algorithmic Decision-making
- Working paper -

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh, Germany
www.bertelsmann-stiftung.org

Responsible for content
Konrad Lischka
Ralph Müller-Eiselt

Authors
Konrad Lischka
Anita Klingel

DOI 10.11586/2017031

Content

1 Preface

Up to 70 percent of job applicants in the UK and in the United States are first evaluated by automated algorithmic procedures before an actual person sees their documents (Weber and Dwoskin 2014). Courts in nine US states use software that calculates risk predictions for the defendants in criminal proceedings (Angwin et al. 2016: 2). Automated predictions of creditworthiness are also used in the United States to determine the cost of insurance policies. The FBI automatically compares images of criminals with 411 million images from driving license, passport and visa data to identify possible suspects.

These four examples show that humans today are already assessed today using what is known as “algorithmic decision-making” (ADM) in many areas of life (Zweig 2016). ADM processes have been in use for many years, and categorize people without much debate over the fairness, clarity, verifiability or correctability of the methods. This may be due to the fact that the systems do not have very much in common with the artificial intelligence (AI) from science fiction. People often associate AI with fictional characters like HAL 9000 or Wintermute: intentionality and consciousness. Strong AI such as this, however, only currently exists in the world of literature and film. Neither of these characteristics describe the systems presented in this collection of case studies. And yet such systems already have considerable influence in court, on the allocation of loans and study places, on the deployment of police forces, on the calculation of insurance rates, and on the level of attention that callers receive in a customer service call. They are all programs developed to deal with specific problems which affect the lives of many people. This is not about science fiction, but about the present (Lischka 2015).

The case studies prepared in this working paper outline the opportunities and risks of such processes. Opportunities such as pattern recognition can help to predict the risk of lead poisoning in children depending on their place of residence (see chapter 2.3 Predicting lead poisoning) or to identify hotspots for specific offenses (for example break-ins, see chapter 2.8 Location-specific predictions of criminal behavior). An artificial neural network calculates the regional distribution of poverty in developing countries on the basis of satellite images almost as well as significantly more expensive local surveys. These results could be used to combat poverty wherever this is most needed and where the impact of aid measures will be greatest (see chapter 2.4 Predicting poverty distribution).

To take advantage of these opportunities for increased participation, algorithmic decision-making processes have to be clearly aimed at this goal in their planning, design and implementation. If this does not happen, the use of these tools can also quickly lead to greater social inequality. The risks and aberrations apparent in the chosen case studies illustrate the sources of error that can occur with many ADM processes. It is often the case that several of these shortcomings will be observed in an individual application scenario. In this working paper, however, each case study showcases one typical need for action which should be taken into account in order to design ADM processes for increased participation.

Table 1: Need for action in algorithmic decision-making processes (source: own representation)

Response	Description	Example
Ensure falsifiability	ADM processes can learn asymmetrically from mistakes. "Asymmetric" means that the system, by virtue of the design of the overall process, can only recognize in retrospect certain types of its own incorrect predictions. When algorithms learn asymmetrically, self-reinforcing feed-back loops may occur.	Recidivism predictions used in the legal system
Ensure proper use	Institutional logic can lead to ADM processes being used for completely different purposes than originally envisioned by their developers. Such inappropriate uses must be avoided.	Predicting individual criminal behavior
Identify appropriate logic model for social impact	Algorithm-driven efficiency gains in individual process steps can obscure the question of whether the means used to solve a social problem are generally appropriate.	Predicting lead poisoning
Make concepts properly measurable	Social phenomena or issues such as poverty and social inequality are often hard to operationalize. Robust benchmarks developed through public discussion are therefore helpful.	Predicting poverty distribution
Ensure comprehensive evaluation	The normative power of what is technically feasible all too easily eclipses the discussion of what makes sense from a social point of view. For example, the scalability of machine-based decisions can quickly lead to situations in which the societal appropriateness and consequences of using ADM processes have neither been debated nor verified.	Automatic face-recognition systems
Ensure diversity of ADM processes	Once developed, the decision-making logic behind an ADM process can be applied in a great number of instances without any substantial increase in cost. One result is that a limited number of ADM processes can predominate in certain areas of application. The more extensive the reach, the more difficult it is for individuals to escape the process or its consequences.	Preselection of candidates using online personality tests
Facilitate verifiability	Frequently, no effort is made to determine if an ADM process uses an appropriate concept of fairness. Doing so is even impossible if the logic and nature of an algorithm are kept secret. Without verification by independent third parties, no informed debate on the opportunities and risks of a specific ADM process can take place.	University admissions in France
Consider social interdependencies	Even when their use is very limited, the interdependences between ADM processes and their environment are highly complex. Only an analysis of the entire socio-informatic process can reveal the relationship between opportunities and risks.	Location-specific predictions of criminal behavior
Prevent misuse	Easily accessible predictions such as scoring results can be used for inappropriate purposes. Such misuse must be prevented at all costs.	Credit scoring in the USA

The case studies described here serve as a means of presenting in detail and summarizing the opportunities and risks of specific, partially evaluated ADM processes. This document reflects the first preliminary state of our discussion of the issue. We publish it as a working paper in order to contribute to a rapidly-evolving field upon which others can build. We therefore publish this working paper under a free license (CC BY-SA 3.0 DE) so that it can also be used, for example, as a basis for discussion in workshops or other examinations of the subject matter.

Algorithmic decision-making will only serve to benefit society if it is discussed, criticized and corrected. It is time for this discourse in Germany. We now have the opportunity to learn from international examples and experience and to shape development. Advances have already been made in this regard, particularly in the United States, where the White House under President Barack Obama presented a report on the challenges posed by machine decisions (Executive Office of the President, National Science and Technology Council und Committee on Technology 2016). In Germany, ADM processes are not yet so prevalent. German courts do not use ADM risk prediction. Only 60 of the 1000 largest companies in the country used computer-controlled candidate selection procedures in 2016 (Eckhardt et al. 2016: 8). What is more, automated face-recognition is integrated into the EasyPASS border control system at only seven airports in Germany (Bundespolizei 2015).

This means we are still able to determine how we want to use algorithms as a society. As the Algorithmwatch initiative puts it: “We have to decide how much of our freedom we allow ADM to preempt” (Algorithmwatch 2016). At the same time, we should discuss not only the “How”, but also the “Whether”: where society opts for solidarity and the communitarization of risks, for example, ADM processes must not individualize these risks. It is not the “technically possible”, but the “socially meaningful” that must be the guiding principle – so that machine decisions can serve to benefit mankind.

Ralph Müller-Eiselt
Senior Expert
Digitalization Taskforce
Bertelsmann Stiftung

Konrad Lischka
Project Manager
Digitalization Taskforce
Bertelsmann Stiftung

2 Case studies

Below we will present nine examples of the use of ADM processes. Our journey begins in the United States, with applications that are specific to this country, such as predicting the risk of recidivism amongst defendants in court or the risk of lead poisoning in Chicago. After a transnational example (evaluation of satellite images for mapping poverty), we come to Europe and a case study from France (university entrance). Following this, a number of US procedures which are also used in Germany (e.g. location-based “predictive policing”) show that the use of ADM procedures is a worldwide phenomenon which has also taken root in Germany. Each case description highlights a typical source of error and thus a need for action in terms of the future design of ADM processes for increased participation. These have been presented as clearly as possible, but in the full awareness that, unfortunately, the identified shortcomings frequently occur together in practice.

We have structured the presentation of the individual case studies in a uniform way to provide a basis for discussions with a quick overview of facts and assessments. A concise summary is provided following the description of the respective output of the systems as well as the underlying data and decision logic. We then present the consequences and available evaluation results of the processes, in order to consider not only the technology, but also the socio-informatic process as a whole. The section entitled “For discussion” briefly summarizes the opportunities and risks discussed in the thematic discussions. Where parallels to the case studies presented exist in Germany, we have outlined these briefly in the final section “Situation and relevance in Germany”.

2.1 Ensure falsifiability: recidivism predictions used in the legal system

Software calculates and predicts the probability of recidivism among offenders. Such algorithmic decision-making processes are now used at least once during the course of criminal proceedings in almost every US state (Barry-Jester, Casselman and Goldstein 2015). More than 60 predictive tools are available on the market, many of which are supplied by companies, including the widely-used COMPAS system from Northpointe.

2.1.1 Output: risk predictions and need for action on a scale of 1 to 10

The COMPAS system assesses individuals in different categories, which are measured against a total of 43 scale values (Northpointe 2015: 2). These include risk projections on the one hand: for example, the general risk of recidivism or the risk of recidivism for specific violent acts. Other categories are intended to identify and quantify the needs of the assessed individuals in order to systematize the planning of interventions. Categories such as poverty, addiction, and criminal associations are used for this purpose. All values are given on a scale from 1 to 10. According to the user manual, the probability of recidivism is “low” in the case of a risk prediction between 1 and 4, “medium” for a risk prediction between 5 and 7, and “high” for a prediction between 8 and 10 (ibid.: 11).

If judges have access to these predictions, they are able to decide whether and to what extent they will include them in their judgment.

2.1.2 Data and decision logic: comparison with test results for a norm group

The COMPAS system determines scoring values based on answers to 137 questions. The information comes from police files and from questionnaires completed by the individual being assessed. These include questions such as: Has one of your parents ever been arrested? How many of your friends/acquaintances consume illegal drugs? The respondents are also asked to evaluate statements such as: “A hungry person has the right to steal” (Angwin et al. 2016: 3).

The weighting of the answers in the calculation has not been made public. The outline of factors in the company user manual is very vague. With regard to the calculation for the prediction of the general risk of recidivism, the manual states the following: “The primary factors making up this scale involve prior criminal history, criminal associates, drug involvement, and early indicators of juvenile delinquency problems” (Northpointe 2015: 30).

The COMPAS output, scored on a scale of 1 to 10 and visible to judges and law enforcement officers, shows a comparison of the distribution of the values in a norm group. This norm group consists of 7,381 criminals who were assessed in US prisons in 2004 and 2005 using the COMPAS system (ibid.: 14). The scoring values determined in the 21 categories were broken down into deciles. For example, in the norm group, one-tenth of those assessed scored 23 or less in the “criminal personality” category, while the second decile scored 24 to 25, and so on. The COMPAS values thus indicate which decile of the norm group best resembles the assessed person based on their scoring value. The individual scales are different from one another, and the indication of decile makes the assessment manageable.

2.1.3 Consequences: the higher the risk prediction, the more likely the offender is to be imprisoned

Courts in many US states resort to such procedures for decisions relating to bail or early release. Pennsylvania is developing a procedure to also include risk prediction in the conviction process (Pennsylvania Commission on Sentencing 2016). In nine US states, such predictions are made available to judges during criminal proceedings (Angwin et al. 2016: 2). In some cases, judgments have been based on COMPAS predictions. In February 2013, Eric Loomis was arrested in Wisconsin for driving a car that had previously been used in a shooting. Loomis pleaded guilty to resisting arrest. He was sentenced to eight years’ imprisonment. The judge stated that the accused had been identified as a major risk to the community by the COMPAS assessment system (ibid.: 10).

2.1.4 Evaluation: different prediction errors for black and white individuals

The use of COMPAS in County Broward, Florida was looked into in 2016 by the US investigative journalism organization Propublica, which is funded by foundations. The reporters evaluated predictions of the risk of recidivism for 7,000 individuals arrested in the years 2013 and 2014. They checked whether these individuals were prosecuted for new crimes within the next two years. Key findings of the Propublica research:

- 20 percent of people with a prediction of recidivism for violent crime were prosecuted for such a crime within two years of the prediction (Angwin et al. 2016: 2).
- 61 percent of individuals (slightly better than chance) with a prediction of general recidivism were involved with the police again in the two years which followed – including for administrative offenses (ibid.).
- The type of prediction error differs between black and white individuals: the proportion of black people with a high recidivism prediction but with no recidivism within two years is twice as high as for white people (ibid.).

Some authors defend these results as fair, as the overall rate of recidivism among black defendants is significantly higher than among white defendants: “Racial differences in failure rates across race describe the behavior of defendants and the criminal justice system, not assessment bias” (Flores, Bechtel and Lowenkamp 2016: 13)

There are almost no independent evaluations for other recidivism prediction systems used in the United States. In most cases, validity was examined in only “one or two studies, and these studies were often conducted by the same people who developed the system” (Desmarais and Singh 2013: 53). In addition, almost all studies investigated only whether the systems predicted a higher risk of recidivism among known repeat offenders, but not whether individuals with a high recidivism prediction actually commit repeat offences (ibid.: 55).

2.1.5 For discussion: a lack of falsifiability can result in bias

The case example illustrates a core problem that can occur in many risk predictions: it is possible to skew your own decision-making basis in a feedback loop. Let us consider, for example, predictions of recidivism used in court. It is accepted that judges tend to prefer imprisonment to parole in cases with higher risk predictions. A longer period of imprisonment can increase the likelihood of recidivism, as individuals will be integrated into new criminal social contexts, for example. This can actually produce an increased rate of recidivism following imprisonment for people with high risk predictions. The prediction system thus proves itself to be correct (the predicted risk came to pass) and, in the long term, might even reinforce these biases if the system is programmed with new data based on a skewed ADM sample (O’Neil 2016a: 28).

Such biases would be favored if the system systematically impedes the falsifiability of a certain group of predictions. This is the case in the present example, where judges are more likely to opt for imprisonment than parole in cases with high risk predictions. People who are incorrectly labelled as “too high a risk” then have no way to prove that they would not have repeat offended if given parole. Potential systematic biases such as this must be sought out, checked, discussed and rectified before a process is used. The example from the judiciary system shows that this is not just about the design of an algorithm or a software package. If, for example, the consequence of a high scoring value for the accused was not imprisonment, but parole with intensive supervision, the process might leave more room for falsifiability. Thus both the possible and actual consequences of an ADM prediction should also be included in the analysis. Depending on how an ADM process is embedded in society, the predictions can restrict participation. If, for example, imprisonment were the only consequence of all risk predictions by design, emphasis would be placed on risk minimization over other functions of the judicial system, such as re-socialization for example (Christin, Rosenblat and Boyd 2015: 9).

Even the definition of fairness underlying the COMPAS system should have been subject to a broad social debate before being used. When designing ADM processes, developers must decide to implement fairness. In some cases, this operationalization inevitably goes hand in hand with a normative decision regarding which definition of fairness is deemed just. Such decisions should be preceded by a social debate as they touch on fundamental social issues. For example, when it comes to predicting the probability of recidivism among offenders, is it fair that every black

person will likely receive a higher risk prediction because the rate of recidivism is higher among black people? Or is it fair if both black and white people who are not repeat offenders are assigned the same risk category? Both fairness definitions are mutually exclusive. The debate over which of these is just began in the USA only after ADM processes had been used in court for several years, when an independent evaluation of the decisions revealed controversial biases. Deciding on the fairness principle that an ADM process should follow must involve a negotiation at the societal level in such cases. If such a debate does not take place, as in the case of COMPAS scoring, the process objectivizes the normative definitions of its small number of designers.

Further opportunities	Further risks
Unlike a judge, an algorithm-based prediction is not bound by daily constraints (e.g. time of day and breaks). An investigation of 1,112 judgments on the deferment of sentences to parole in Israel found that the likelihood of a positive decision for the defendant is greater at the beginning of the day and after food breaks than at other times (Danziger, Levav and Avnaim-Pesso 2011: 6890).	Justice is individualized. Judgments must be based on the individual case and the actions of the individual – not on similarity to norm groups. Are risk predictions not based on such comparisons? What is important here is that judges really evaluate the individual case. There is some evidence showing that where people deviate from risk predictions, they generally rule to the detriment of the accused and order imprisonment in spite of favorable predictions (Steinhart 2006: 70). In addition, the scope of application of the software used is potentially many times larger than that of a judge: once created by the team of developers, the decision logic will impact many more cases than the decision logic of a single judge.
The incarceration rate may fall, as judges are more likely to consider alternatives to imprisonment in the case of low risk predictions. The incarceration rate in Virginia has risen considerably more slowly since the introduction of risk predictions in 2002. In 2014, judges in Virginia sentenced almost half of all defendants charged with non-violent crimes to prison alternatives (such as rehabilitation programs). The number of prisoners in Virginia has risen by 5 percent since 2005, compared with a 31 percent rise in the previous decade (Angwin et al. 2016).	People perceive software-based predictions to be more reliable, more objective, and more meaningful than other information relating to a case, including their own impression (Hannah-Moffat, Maurutto and Turnbull 2009). This can result in predictions not being questioned in individual cases. The prediction, however, can still be influenced by human errors of judgment. The data basis for risk predictions can contain biases which seem to objectify the scoring value. For example, when the question “When did you first have contact with the police?” is included in the LSI-R prediction process, the risk projections are biased against people from neighborhoods with high levels of poverty and crime and a high police presence (O’Neil 2016a: 27).

2.2 Ensure proper use: predicting individual criminal behavior

20 of the 50 largest municipal law enforcement agencies in the US use Predictive Policing (Robinson und Koepke 2016: 20). The police in Chicago have been recording citizens who have a criminal record on a "Strategic Subject List" (SSL) using an ADM process since 2013. The system is developed by a team at the Illinois Institute of Technology, and funded by the US Department of Justice.

2.2.1 Output: software predicts victims and perpetrators of violence

Around 1,400 previously convicted people in Chicago are registered on the SSL list. Each person is assigned a scoring value between 1 and 500. The higher the value, the higher the risk of being involved in a future shooting or murder as either the perpetrator or the victim (Johnson 2016: 1). This is according to the SSL service statement. However even Police Superintendent Eddie Johnson publicly presents the predictions as a tool for the identification of high-risk perpetrators, claiming that the 1,400 people on the list are "responsible for the majority of the violence in the city" (Davey 2016). Essentially, investigators decide how to use the SSL predictions in their work.

2.2.2 Data and decision logic: no transparency

Ten variables from the police database are used to evaluate the ADM process. Which ones these are and how they are evaluated is kept confidential by the police under the auspices of "proprietary technology". A police representative gave the following examples of relevant information used in the process: Has a person been a victim of a shooting? Is the person's criminality trend line on the rise or on the decline? Were there any arrests due to gun crimes? (ibid.).

2.2.3 Consequences: home visits for high-risk persons and arrests

Police officers visited the homes of around 1,300 people with high scoring values, often together with social workers, in order to offer assistance (ibid.). Officials can also use the scoring values for investigations, and all police officers have access to the database (Johnson 2016: 1). A high value is not considered probable cause for a house search, for example. Nevertheless, the probability of an arrest is correlated with inclusion in the list, as shown by an evaluation by the RAND Corporation: "One potential reason why being placed on the list resulted in an increased chance of being arrested for a shooting is that some officers may have used the list as leads for closing shooting cases" (Saunders, Hunt, Hollywood, Criminol and Org 2016: 1).

2.2.4 Evaluation: virtually no prevention, more arrests, no influence on violent crimes

The RAND Corporation is evaluating the project. Conclusion on the first year of use: no effective prevention can be determined for those on the list, the system failed to predict 99 percent of murder victims between March 2013 and March 2014 (Saunders 2016). In May 2016, the police announced that over the course of the year more than 70 percent of people who had been shot and more than 80 percent of those who had been arrested in connection with shootings were on the SSL list (Davey 2016). There is no further information on the 2016 data, and no independent studies are available. Arrests do not tell us whether the individuals in question were actually the perpetrators or not. This could also be an effect of the list itself: investigative pressure is first exerted on those who are already known to the police.

> *"However, the Chicago Police failed to provide any services or programming. Instead they increased surveillance and arrests — moves that did not result in any perceptible change in gun violence during the first year of the program. (…) The names of only three of the 405 homicide victims murdered between March 2013 and March 2014 were on the Chicago police's list, while 99 percent of the homicide victims were not" (Saunders 2016: 1).*

Even if the quality of predictions and effectiveness of preventive measures were to rise rapidly, RAND ultimately expects only marginal benefits. Enormous progress would have to be made in order to reduce the city's murder rate by five percentage points. The quality of the predictions would have to be ten times that of the first year, and the effectiveness of the interventions for potential victims and perpetrators would need to be increased fivefold. RAND therefore advocates that other approaches are not ignored: "And after all that improvement — here's how

many lives would be saved: 21. In a city that reported 468 murders last year, that would be tremendous progress, but hardly the definitive solution" (ibid.).

2.2.5 For discussion: appropriate embedding is also a factor

According to current sources, the Strategic Subject List was developed in Chicago as a tool for prevention. However, in actual usage, the tool was rarely used as planned. The evaluating researchers were unable to establish a meaningful implementation of the predictions for preventive interventions as part of police work. Their conclusion:

> *"Overall, the observations and interview respondents indicate there was no practical direction about what to do with individuals on the SSL, little executive or administrative attention paid to the pilot, and little to no follow-up with district commanders. These findings led the research team to question whether this should be considered a prevention strategy" (Saunders et al. 2016: 10)*

This example demonstrates that the operative application and implementation of consequences also determines the impact of an ADM process. Not only is deliberate misuse a risk (see chapter 2.9. Prevent misuse, but so too is improper implementation, as in this example from Chicago shows.

Chicago lacked the necessary human resources for the planned preventive work on the basis of the SSL predictions. The available staff obviously used the predictions as an investigative tool instead, following the existing institutional logic. In this way, the software may narrow the investigator's focus in a search for suspects to only the people on the risk list. Such mechanisms threaten the presumption of innocence and risk jeopardizing the effectiveness of police work. According to its portrayal in the public eye, the SSL was not developed as a tool for seeking out criminality. How well the system is suited to this application would have to be evaluated on an independent basis. The example shows that the quality of ADM processes should also be measured in terms of their operative embedding in institutions and, above all, in terms of their true proper use.

Further opportunities	Further risks
Police resources could be used more effectively and with greater efficiency on the basis of the predictions.	The approach of the ADM process reduces successful police work to a single aspect: identifying suspects. Errors in system predictions are rarely considered, which can produce false incentives for the system's application. How many people on the list were wrongly suspected to be a threat, and even arrested? Such factors are not evaluated. Alternative indicators of system impact (e.g. trust in the police in individual neighborhoods, excessive violence in police operations) are not recorded. These factors, however, influence the willingness of residents to cooperate with the police, and can therefore improve clearance rates (The Leadership Conference on Civil and Human Rights et al. 2016: 2).
Preventive work could become more effective and efficient and, in the best case, crime rates may fall as a result.	The lack of transparency of the decision logic makes a comprehensive public debate impossible (ibid.: 1).

.

2.3 Identify appropriate logic model: predicting lead poisoning

Nearly 90 percent of the housing stock in Chicago was built before 1978 – the year in which lead-containing paint was banned in the US (Potash et al. 2015: 2039). As a result, lead poisoning in children is still a major problem in the city: in 2013, ten percent of children six yeasr and younger in Chicago had lead concentrations higher than the thresholds set by the US Center for Disease Control and Prevention, which is four times the US average (Hawthorne 2015). The existing measures are not put in motion by the city until a child is diagnosed with lead poisoning. Only then can the renovation of housing be arranged (Potash et al. 2015: 2040). There is a lack of political majorities for preventive building renovation with public funds (Hawthorne 2015).

The city is working with the University of Chicago to develop software to predict which buildings and which children are at the highest risk of lead poisoning, in order to provide early, targeted and therefore favourable interventions.

2.3.1 Output: ranking of particularly vulnerable children and high-risk buildings

The Chicago Department of Public Health (CDPH) wants to use the software to prioritize buildings and children for further measures by means of a risk assessment. The higher the risk of lead poisoning, the higher the ranking of the affected buildings and children concerned (Potash et al. 2015: 2042).

2.3.2 Data and decision logic: blood tests and inspections

The following data was available to researchers: 2.5 million lead poisoning blood test results for around one million children in Chicago between 1993 and 2013, giving the date, identity of the test subject, age and place of residence. In addition, they also had access to the results of 120,000 house inspections from the same period, giving the date and location. The researchers divided up the data sets and programmed several classification methods using a portion of the data, in order to then make predictions about the period to which the other data portion relates (ibid.: 2041) The first results, published by the researchers as part of a peer review process, show that information in the training data relating to age, inspection results and the condition of buildings at address level was particularly important in improving the quality of the prediction (ibid.: 2044).

2.3.3 Possible consequences: inspections of high-risk buildings

The CDPH cites prioritized inspections of high-risk buildings by inspectors and their subsequent renovation when limit values have been exceeded as a possible consequence of the risk ranking. As conceivable measures, the researchers also propose targeted advertising for blood testing in high-risk streets, the publication of address-related risk predictions for tenants, and targeting landlords on the basis of predictions (ibid.: 2046).

2.3.4 Evaluation: under way

The Chicago Department of Public Health is currently validating the model (Chicago Department of Public Health 2016: 3), and its application seems to focus on address-related risk predictions.

2.3.5 For discussion: efficiency gains do not always represent an appropriate logic model

If the predictions prove to be accurate, the city of Chicago could target parents of children at the highest risk of lead poisoning and prioritize the renovation of high-risk buildings. This would be an improvement compared to the status quo, where meagre resources are spread too widely. This is made possible by a fundamental advantage of machine decisions: algorithmic methods can evaluate far more factors and data than humans.

However, this advantage alone does not guarantee that more opportunities for participation will be created for all. A fundamental risk that can persist even in the case of accurate predictions is that the targeted use of resources by means of algorithms can override the question as to whether the nature and extent of the measures is based on an appropriate logic model. To develop the logic model, the overarching objectives and the options for action have to be made transparent.

In the present case study, it is notable that the use of the ADM process is being discussed regardless of the resources available for the investigation and renovation of buildings. What happens following a prediction of an

increased risk of lead poisoning? In 2015, the City of Chicago employed eleven inspectors to inspect houses for lead contamination, as well as three nurses. This represents just a quarter of the staff number available for these issues in 2010 (Hawthorne 2015). Any efficiency gains achieved through the use of ADM processes thus might only serve to balance out the shortcomings created by savings – if they are of any benefit at all. Perhaps more resources overall should be devoted to protecting children from lead poisoning. Perhaps even the existing mechanism of action is not sufficient or is not appropriate for achieving the desired goal from a societal perspective – perhaps the focus should be on preventing lead poisoning, instead of “merely” encouraging parents to seek blood tests to obtain a diagnosis of lead poisoning in their children. Such issues cannot be resolved by the design of an ADM process. They can, however, be used to determine the framework and the objectives of the design of an ADM process.

2.4 Make concepts properly measurable: predicting poverty distribution

In order to use development aid in a targeted manner and to evaluate the impact of measures, it is necessary to have up-to-date information on the local distribution of poverty. To develop a new data basis with a greater degree of variance, researchers have trained an artificial neural network to identify landscape features that are associated with extreme poverty in satellite images taken during daylight hours. The results were published in Science in August 2016 (Jean et al. 2016), however, the procedure is not yet used in practice.

2.4.1 Output: expenditure and wealth at village level

The software predicts daily per capita expenditure as per the World Bank definition for geographical clusters at village level in Nigeria, Tanzania, Uganda, Rwanda and Malawi as well as household wealth as per the definition of the Demographic and Health Survey Program (2014), which is used in US development aid.

2.4.2 Data and decision logic: comprehensive satellite images and surveys

Data on poverty distribution could feasibly be provided by surveys on purchasing power and wealth. In rural regions of Africa, however, such surveys are complex, expensive and therefore rare: between 2000 and 2010, 39 of the 59 countries in Africa conducted fewer than two such surveys (Patel 2016). As a result, researchers are on the lookout for other data sources that provide information about the distribution of poverty at village level. Data on mobile network usage is of some relevance, but is not publicly available. Satellite imagery taken at night is publicly available, but the informative value of such images is lower in regions where many people live in extreme poverty (as defined by the World Bank, 2015): where extreme poverty prevails, it is almost completely dark at night and any gradation is very slight (Jean et al. 2016: 790).

For this reason, the research team at the Stanford University Sustainability and Artificial Intelligence Lab is using day and night satellite imagery as well as current survey results on per capita expenditure and household wealth to train artificial neural networks in a series of steps. In the first step, a neural network maps out the characteristics of images taken during daylight hours and relates these to the light differences in the nighttime images. Some of these characteristics are also visible with the naked eye, such as roads, urban settlements, farmland, etc. (Horton 2016). In the second step, the researchers trained an artificial neural network to determine which of the mapped out characteristics in the daylight images are associated with the poverty distribution for the region, as determined by survey results. The software has, for example, worked out that the material composition (metal, straw, soil, grass) of roofing is related to per capita expenditure (Jean et al. 2016: 791).

2.4.3 Evaluation: better results than with mobile network data

The method provides better results than methods based on mobile network usage. The comparison shows that the quality of predictions relating to household wealth at village level in Rwanda is better (correlation coefficient for the mobile network method 0.62 vs. 0.75 for pattern recognition in daylight satellite imagery) (ibid.: 792). The study also shows that the models trained using survey data from one country can be applied in other countries:

> *"Pooled models trained on all four consumption surveys or all five asset surveys very nearly approach the predictive power of in-country models in almost all countries for both outcomes. These results indicate that, at least for our sample of countries, common determinants of livelihoods are revealed in imagery, and these commonalities can be leveraged to estimate consumption and asset outcomes with reasonable accuracy in countries where survey outcomes are unobserved" (a.a.O.: 794).*

2.4.4 Possible consequences: targeted deployment of relief measures

At present, it is not known which measures aid organizations would implement on the basis of these predictions. It is conceivable that there might be a positive impact on the people affected by the predictions (through the expansion of aid measures).

2.4.5 For discussion: public discourse on operationalization is needed

If algorithms based on available satellite imagery reliably predict the distribution of poverty, this could result in considerably more low-cost, more up-to-date and, above all, more needs-based aid measures. The reason for this

is a common feature of all ADM procedures: once a decision logic has been developed, it can be applied in any number of cases at comparatively low cost. In some cases it can even be applied under other framework conditions, such as to other countries, as in this case study.

These advances are possible because the ADM process can be optimized for the prediction of clear target variables, such as per capita expenditure in the present case. This key performance indicator has long been used in the field of development aid, and institutions such as the World Bank rely on its informative value. What is new is the prediction approach, not the measured value. It is advantageous in the use of ADM processes if their purpose and operationalization have previously been subject to (professional) public discourse.

Since the tools are only in the early stages of development, we can only speculate about the consequences of their application. The quality of the predictions should be checked by comparing them with random samples from survey results. This must be done in practice. If an ADM process is used for the local distribution of relief measures in practice, negative consequences are also conceivable. For example, aid measures could be reduced in areas which are better off by national comparison, even if poverty is problematic there by international standards.

2.5 Ensure comprehensive evaluation: automatic face-recognition systems

Since 2008, the United States FBI has been operating a system that uses facial recognition to analyze images of unknown persons and seeks matches in various databases which contain around 400 million images of US and foreign citizens (e.g. from visa applications). The US Government Accountability Office has criticized the fact that the reliability and error rate of the overall system have never been tested (United States Government Accountability Office 2016).

2.5.1 Output: up to 50 possible matches for a wanted person

Investigative authorities which possess images of individuals suspected of committing a specific criminal offense may ask the FBI to check these images against their existing database. The aim of such a request is either to confirm an existing suspicion or to ask for a list of possible suspects whose biometric criteria match those of the suspect. As a rule, requests can only be made via the FBI. Since 2011, however, investigative authorities from seven individual federal states have been able to directly access the database as part of a pilot project. Between 2011 and 2015, these authorities conducted more than 20,000 searches (interestingly, the number of inquiries per federal state ranges from 20 to 14,000). When a search request is made, the algorithm compares the image of the wanted person with the biometric information stored and generates a list of between 2 and 50 possible matches for their identity. This information is checked manually in the case of indirect access by what are known as biometric analysts, and reduced to one or two candidates which can then be passed on by the FBI to the requesting organization (e.g. local police department). A total of 29 such analysts worked for the FBI in 2015. In the case of direct access by a federal investigative authority, this intermediate step is not guaranteed. The process is used not only for investigations into violent crimes, but also for cases of theft or insurance fraud (ibid.).Investigators decide, on the basis of the submitted list of suspects, whether and against which of the persons indicated they will take further action.

2.5.2 Data and decision logic: “Criminal identities” vs. “civil identities”

The database is fed with voluntary submissions from the various US authorities, and currently comprises more than 30 million images. A distinction is made between “criminal identities” and “civil identities” on the basis of criminal records: “criminal” images are those taken in the context of arrests, convictions or imprisonment. More than 80% of the images belong to this category. “Civil identities” come from employee records, military service records, voluntary service records or immigration papers. Each image is associated with a complete set of fingerprints so that duplicates are automatically linked (even between criminal and civil identities). Once entered, images can only be removed from the database by the submitting organization or by court order. Through a special department (FACE), the FBI can also access other government databases (for example pictures from driving licenses and visa applications). Taking these external databases into account, the number of available images amounts to more than 411 million – and concerns some 64 million Americans (Garvie, Bedoya and Frankle 2016).

The decision logic consists of two steps: incoming images are analyzed for biometric criteria and then stored in the database. Civilian records can only be searched by the FBI, while requests relating to the records of people with a criminal record are open to all law enforcement agencies that submit images. However, if civilians are linked to criminal records, both are displayed. A list of the most promising matches will then be forwarded to Human Analysis in the case of an indirect request (see “Output”), or to the investigating authority using the system in the case of a direct request (United States Government Accountability Office 2016). Nothing is known about the criteria used in the algorithm.

2.5.3 Consequences: 64 million people under continuous scrutiny

In cases where an incorrect match is made between a person from the database and a search image, the suggested person can be wrongly suspected. Depending on the state, a positive match can even be used as evidence in court. Thus the algorithm plays a role in decisions about the freedom or imprisonment of citizens. Simply appearing on

the list of results has various consequences for those concerned: the information obtained can be used as a basis for house searches, data requests from internet providers and banks as well as arrests.

Ethnic discrimination must also be considered: the criminal record database contains more pictures of black people than white people. As a result, the algorithm is more likely to find a match for a black person than for a white person.

Last but not least, the use of facial recognition algorithms represents a legal limitation to the presumption of innocence on two levels: firstly, the algorithm provides sufficient initial suspicion to initiate an investigation against the person concerned. Secondly, a positive match can be regarded as proof of guilt in some federal states. The falsification rate of the algorithm used describes the probability of suspecting or even arresting innocent citizens on the basis of a possibly incorrectly matched database entry. Therefore, social discourse should be held about the acceptable limit of this rate prior to the implementation of such an algorithm. This relates to a fundamental question of fairness, which must be operationalized early on due to the peculiarities of ADM processes: Is society willing to accept the risk of wrongfully accusing citizens? How many? And how do we minimize the suffering caused in this way?

2.5.4 Evaluation: technically inadequate evaluation, lack of legal framework

Various control systems were developed prior to the deployment of the Next Generation Identification-Interstate Photo System (NGI-IPS) to ensure that the technology used is ethically permissible. The Department of Justice (DOJ), for example, demands a “Privacy Impact Assessment” (PIA) of any technology that compiles citizens’ data. In the case of the NGI-IPS, however, this was only conducted for the originally introduced and functionally much less extensive system in 2011. The updates and extensions which have been introduced since then were only reviewed in September 2015.

In addition, the FBI evaluated the success of the new algorithm internally. As part of this evaluation, both the identification rate (probability that the wanted person is among the 50 proposed matches) and the misidentification rate (probability of someone being wrongly suggested as a match) were checked. In the first instance, a hit rate of 85 percent was set as acceptable (only taking searches prompted externally into consideration, this corresponds to 3000 unsuccessful searches). So if the person you are looking for is included in the database, they should be present on a list of 50 possible matches in 85 percent of the cases. This target was achieved within the context of the evaluation with an actual result of 86 percent of cases. Only lists containing 50 possible matches were tested, however, although lists of between 2 and 50 possible matches are possible, with a default of 20 potential matches. No evaluation is available for these smaller lists.

The misidentification rate was not tested at all. The FBI argued on this matter that the lists contained only potential matches and were therefore not “positive”. However, both the United States Government Accountability Office (2016) and Garvie, Bedoya, and Frankle (2016) note that merely holding a suspicion represents a departure from the presumption of innocence – Garvie, Bedoya and Frankle describe the process as a “perpetual line-up”.

Irrespective of this concrete example, scientific opinion is also divided with regard to the effectiveness of algorithmic facial recognition in general (Revell 2016) – there are too many factors influencing the actual success rate. The importance of lighting is given as an example: in an experiment conducted in a subway station in Mainz, Germany, the real-time hit rate of the facial recognition algorithm varied between 60 percent during the day and 10 to 20 percent at night (Garvie, Bedoya and Frankle 2016). Other complicating factors range from the angle at which the image is taken, the resolution of the camera itself, and the quality of the reference picture, to plastic surgery, make-up and aging processes.

There is also the question of securing such a system against unlawfulness or unauthorized access. The possible scenarios range from hacker attacks to authorized users carrying out unlawful searches (e.g. for relatives) (Garvie, Bedoya and Frankle 2016). Here too, the evaluators criticize the inadequate structures currently in place to protect the system from such abuse. For example, only five federal states regulate police use of facial recognition algorithms at all. Not a single federal state has a legal requirement for the minimum identification rate at which use of

the systems is permissible. What is more, the severity of the offense required to permit the use of the system also varies from state to state (ibid.).

The development of the algorithm at the control points that are actually provided also raises fundamental questions. How do we design a control system that can keep up with the pace of a constantly changing algorithm and not only reflect the technical and legal complexity of its application and of the corresponding social debate with hindsight, but also ensure preventive control?

2.5.5 Discussion: comprehensive evaluation continues and analyzes indirect consequences

The case study shows that the scalability of machine decisions can quickly lead to deployment scenarios which have not been fully discussed in terms of social appropriateness and the resulting consequences. ADM processes allow searches to be carried out with a scope and frequency that was not possible with analog means. Many police stations have access to the FBI database. The database, in turn, links a variety of sources. This interlinking and the low effort of algorithmic facial recognition could result in:

- the process being used for petty offences.
- an ultimate increase in the number of errors due to an increase in search requests.
- an increased risk of misidentification for certain people because their images are contained in the database as a result of systematic bias. In high-poverty neighborhoods, for example, there is a higher likelihood of police checks, chance discoveries and, consequently, mug shots.

The consequences of this new type of facial recognition have not been sufficiently evaluated in the case study. The appropriateness of continuous automatic facial recognition has never been examined or discussed, neither before its introduction nor since. A comprehensive evaluation also includes the assessment of indirect consequences and the ongoing analysis of the system's actual application.

Further opportunities	Further risks
Increase in the likelihood of catching criminals: there are examples of criminals who were able to escape the long arm of the law for decades prior to the introduction of standardized recognition, but who have now been successfully identified.	It is impossible for US Americans to evade facial recognition by ADM processes. According to Garvie, Bedoya and Frankle, every second citizen in the US has been subjected to algorithmic facial recognition without their knowledge.
Increase in resource efficiency in the detection of criminality: without the use of an algorithm, searching a central database containing more than 400 million images for matches would be justified only in selected cases. The use of an algorithm permits access even in cases of slight suspicion.	Misidentification can lead to the false suspicion of innocent people. In addition to the inspection and monitoring of private data streams, the consequences of this false suspicion include possible investigation or even a conviction.

2.5.6 Situation and relevance in Germany

Automated facial recognition is used as part of the EasyPASS border control system at seven airports in Germany (Bundespolizei 2015). The German Federal Ministry of the Interior, together with Deutsche Bahn, developed a concept for monitoring railway stations using facial recognition in 2016 which has already been piloted in 20 stations (Plass-Fleßenkämper 2016). In 2016, Federal Minister of the Interior Thomas de Maizière called for the introduction of facial recognition systems at all German stations and airports ("*Terrorbekämpfung*" [Combating Terror] 2016). In February 2017, Deutsche Bahn announced that it would test intelligent video surveillance with facial recognition software at Berlin Südkreuz railway station: "This camera is a small miracle: it is designed to screen for people who are stored on a list of suspects using facial recognition. It should also register objects, such as luggage or packages, which have not been moved for a long period of time. It should even be able to recognize the typical behavior of pickpockets" (Kurpjuweit 2017).

2.6 Ensure diversity of ADM processes: pre-selection of candidates using online personality tests

In the UK and in the United States, 60 to 70 percent of applicants are subjected to automated competitions and tests. Personality tests play an important role in such testing. Conducting tests online is much cheaper than doing it in person. For this reason, agencies and employers are using online personality testing more frequently and at earlier stages of the selection procedure – this is also increasingly true of jobs paid at or below the national average in the service sector.(Weber and Dwoskin 2014).

2.6.1 Output: many applications are never seen by human eyes

Automated procedures are used to pre-select applications. A portion of these are immediately rejected on the basis of the online tests, even before a person has seen the applications. The employer can determine what percentage of the applications received are immediately refused. One test provider puts the percentage of automated rejections at around 30 percent (ibid.). Computer scientist Cathy O'Neil (2016a: 105) claims that 72 percent of applications in the US are assessed by machine only.

According to the limited information available, the pre-selection process appears to operate without human decision-makers.

2.6.2 Data and decision logic: online questionnaires on personality

There is little public knowledge about the decision logic of procedures. In the US, providers such as Kronos are legally opposed to requests for information from US federal agencies.

The online tests contain scale questions such as "I can experience many mood changes during the course of a day" and "When something very bad happens, I need some time to feel happy again" (Weber and Dwoskin 2014). These questions clearly seek to evaluate applicants according to the five-factor personality model. "People the system classified as 'creative types' tended to stay longer at the job, while those who scored high on 'inquisitiveness' were more likely to set their questioning minds toward other opportunities" (O'Neil 2016a: 109).

Some tests ask how long the candidate estimates their commute to their new place of work is. This information was used by a Xerox Services provider (US call center operator which hires 30,000 applicants per year) for the automated selection of candidates: those who had too long a commute were rejected because employees with long commutes were statistically more likely to quit than others. Xerox Services eliminated this criterion because it could systematically discriminate against people from poorer neighborhoods with a predominantly black population who cannot afford housing near the company. It is possible that the courts might consider this practice to constitute discrimination on the grounds of skin color, if someone were to sue (Weber and Dwoskin 2014).

2.6.3 Consequences: it is not just about *one* job, but about access to the labor market as a whole

Many employers in the US use the software offered by a select few service providers for the automated selection of applicants. This widespread use means that for jobs in the service sector paid below the national average the software acts as a gatekeeper for not just one, but for the majority of potential jobs. The consequences of this can be seen in the case of Kyle Behm. Following successful psychiatric treatment for bipolar disorder, the engineering student was back at university and looking for a part-time job. He had previously worked in supermarkets. He was then rejected by seven potential employers using similar online tests for minimum-wage positions in fast food shops, hardware stores and supermarkets (O'Neil 2016c: 1). Behm's father contacted the companies, and the majority offered Kyle a suitable job after closer examination, in exchange for Behm agreeing not to take legal action. He filed a complaint with the US Federal Equal Employment Opportunity Commission (EEOC). The investigation into the use of personality tests is ongoing.

For certain groups, automated procedures such as these can make access to the labor market more difficult on the whole. An earlier form of discrimination (e.g. based on foreign/ethnic-sounding names) could be replaced by another. Since the technology is used mainly in the low-wage sector, companies are unlikely to invest in the improvement and testing of the systems on the basis of individual cases. The systems do not have to find the best of the best, they just have to be more efficient than the previous selection system. Investments in the calibration of systems and in the continuous testing and updating of the decision-making logic and data stock will never be as important in this sector as in sectors with a low supply of labor, high demand, and correspondingly high salaries. In addition, it is virtually impossible to falsify the predictions of this ADM process in such a way that the system learns from the error. Even if Kyle Behm is rejected seven times but then gets a job elsewhere at a McDonald's and works his way up to manager within two years, none of the seven other companies will check the personality test used to find out why the prediction was wrong (O'Neil 2016c: 4).

2.6.4 Evaluation: there are no independent analyses

The effectiveness of the automated selection procedures has not yet been independently tested. Some companies, such as Xerox Services for example, report successes: "The attrition rate has dropped by 20 percent in some locations, and some people had been recruited who would not have been given a chance based solely on their CVs", according to the Personnel Manager at Xerox Services in 2014. A systematic evaluation does not appear to underlie this judgment, however: "'I don't know why this works', admits Ms Morse, 'I just know it works'" (Smedley 2014: 1). These statements should be treated with caution for two reasons: firstly, it is unlikely that companies would advertise any failures of automated procedures. Secondly, companies' key figures do not cover the impact of the procedure on rejected applicants, such as the question of whether systematic biases occur.

The US Society for Human Resource Management stated at a hearing of the Equal Employment Opportunity Commission that there is no empirical evidence relating to either the validity or the adverse effects of the testing procedures (Dunleavy 2016). A meta-analysis of 7,000 publications showed that personality tests are of very little relevance to future performance in the workplace (Morgeson et al. 2007).

2.6.5 For discussion: the more widespread the use of ADM, the greater the impact of even rare errors

The case of Kyle Behm illustrates a possible structural problem of ADM processes. In the selection of candidates, for example, people use guidelines in a decentralized way, thereby making their application diverse and different. Software, on the other hand, directly applies the defined process in the same way for each individual case. In a decentralized process, one of the decision-makers might have given Kyle Behm a chance. This centralization of decision-making logic becomes more of a burden the more institutions use the same ADM processes. According to the present sources, as a result of the broad use of certain ADM processes offered by a small number of service providers in the US, just a few procedures dominate certain sectors. Let's take the use of a pre-selection process in recruitment for low-paid service jobs, for example. This can make it virtually impossible for already marginalized groups to get past the first hurdle of the automated procedure.

At present, legal action appears to be a weal corrective instrument in such cases. This is because the lack of transparency of the procedures and decision-making logic makes it difficult for applicants to even begin finding a starting point for a complaint, let alone for legal action. On top of this, potential plaintiffs in the low-wage sector have limited financial resources at their disposal. Any rejected candidates must consider whether it is worth taking legal action rather than investing all their efforts in continuing their job hunt.

The effects of structurally biased falsification, which are also touched on in chapter 2.1, are also evident: if one of the ADM processes used for the selection of candidates systematically rejects suitable candidates, the present design does not allow the system to recognize this error or learn from it. The costs of an unjustified rejection for a low-wage job are low for the rejecting company, and there is therefore little incentive for the subsequent optimization of the procedure. If it is not possible to falsify the predictions, an important quality component for ADM processes is lost, as is the case with the widespread use of a select few procedures in the low-wage sector (O'Neil 2016c: 4). This can be remedied by increasing the range of different procedures used.

Further opportunities	Further risks
Formal qualifications can be of lesser importance in automated test procedures than in conventional ones. This opens up opportunities for previously disadvantaged groups, such as the long-term unemployed or low-skilled applicants. Skills are more important than credit score, and the needs of the labor market are more important than qualifications.	There is some evidence that some of the procedures disadvantage certain groups. People from low-income neighborhoods, for example, or people with mental illnesses, which may affect their personality test, but which will not necessarily impact on their ability to do the job for which they are being tested.
Rates of discrimination based on factors such as gender, foreign-sounding names, applicant photos, or openly communicated disabilities may fall. To illustrate the current situation in Germany: “To receive an invitation to interview, a candidate with a German name will have to submit an average of five applications, while a competing applicant with a Turkish name will have to submit seven” (Schneider, Yemane and Weinmann 2014: 4).	If the ADM processes are trained using data resulting from the current selection process, the existing discrimination will be allowed to continue (Trindel 2016).

2.6.6 Situation and relevance in Germany

According to a survey conducted in 2016, “6.0 percent of the 1,000 largest German companies currently use computer-controlled selection procedures” (Eckhardt et al. 2016: 8). However 47.5 percent of respondents believe that “the computer-controlled and automatic selection of applications will be used increasingly in the future” (ibid.). The same study came to a different conclusion in a survey of job seekers and those climbing the career ladder in Germany: four out of ten respondents stated that they had been faced with computer-controlled and automated selection tools at least once during their search for a position (ibid.). Aachen-based company Precire analyzes 15-minute speech samples of applicants using an ADM process to infer personality traits. Companies choose candidates for an interview on the basis of this selection process. The software is also intended to be used for the analysis of stress levels in the company (Morrison 2017)

2.7 Facilitate verifiability: university admissions in France

In 2009, the French Ministry of Education (Ministère de l'Education nationale, de l'Einseignement supérieur et de la Recherche) introduced a digital approval process for state universities – the Admission Post Bac (APB). An ADM process has been allocating students a study place, and thus a university, ever since. In 2016, the student representation organization *Droits des lycéens* [Rights for high school students] filed a lawsuit demanding the publication of the algorithm under French freedom of information law (Commission d'accès aux documents administratifs, CADA) (Thompson 2016).

2.7.1 Output: allocation of study places to high school graduates

In 2015, APB allocated approximately 740,000 pupils to more than 11,000 courses of study throughout France. As a result, 60 percent of the students were assigned to their first-choice university, 14 percent to their second choice and 8 percent to their third (Graveleau 2016). One percent of the study places were distributed by drawing lots.

The final decision on the admission or rejection of a candidate lies with the university.

2.7.2 Data and decision logic: place of residence and preference count

Each student can specify up to 24 preferred study courses (associated with a certain university in each case), which they then have to list in order of preference. The algorithm then uses this information to calculate which courses of studies have more candidates than places. For courses which are not oversubscribed, students are allocated their first choice. For oversubscribed courses, the algorithm gives priority to study applicants who completed their school-leaving qualifications in the school district (*académie*) in which their preferred school is located. In a second step, the algorithm evaluates the relative and absolute preference for a course of study: absolute preference refers to the actual position of the course of study in the student's list of preferences, while relative preference refers to its position in the list of preferences after all non-oversubscribed courses are removed. The process which sorts the candidates first prioritizes their location, then their relative, and finally their absolute preference for a course of study (Graveleau 2016) If there are still more candidates left than there are study places are available, a lottery decides the results. All places lower than the highest priority allocation on the student's list of preferences are automatically excluded and offered to other pupils ("Further Education in France" n.d.).

This decision logic was published only after considerable legal disputes between Droits des lycéens and the French Ministry of Education. Although the code was requested in January 2016, the Ministry intentionally avoiding publishing it until after the expiry of the 2016 deadlines, in order "to avoid causing concern" (Graveleau 2016). A schematic representation of the algorithm was published in June 2016, but this was in no way verifiable. The Ministry explained that this was for "reasons of security, particularly against hacker attacks", but also with the aim of "easier understanding", since the source code contained more than 250 pages of "incomprehensible code" (ibid.). After several complaints, the Ministry of Education finally published the code of the algorithm in October 2016. Unfortunately, however, this was also useless without any explanation of the variables used, and was also printed out and delivered by post (Berne 2016). Droits des lycéens subsequently published the documents on the GitHub collaborative platform and asked volunteers to help analyze the program. A draft law, which was intended to legally underpin the lottery procedure and the criteria used, was unexpectedly withdrawn by the Ministry in January 2017 (Stromboni 2017).

2.7.3 Consequences: implicit social selection based on place of residence and strategic selection of preferences

Droits des lycéens criticized two points: first, the initial selection by place of residence meant that students from Paris had a higher chance of getting in to one of the prestigious *Grandes Écoles*. Students from rural areas, on the other hand, were considerably disadvantaged (Frouillou 2016; de Coustin 2016). Second, these criteria, as well as the elimination of all options after the first allocation, results in students choosing their preferences strategically: (potentially) oversubscribed courses of study are given as high a priority as possible in order not to be eliminated by allocation to a course which is not oversubscribed. This, in turn, suggests to the system that there is a huge

uptake of oversubscribed courses and reduces the chances of obtaining one of the limited places in the lottery due to the high number of false preferences (Graveleau 2016).

2.7.4 Evaluation: consider the reproduction of social inequalities and workaround strategies

The results of APB allocation have previously been examined by Leila Frouillou (Sorbonne) (Lefauconnier 2016), who concluded that the allocation of study places is equally accessible to all, but that the system reproduces social inequalities. As a result of the regional allocation criterion, students from the *académies* in Créteil and Versailles have a much higher chance of admission to the elite universities – which are mainly based in Paris – than pupils from Marseilles or Lyon, regardless of their qualifications. However, as housing in the Paris region is generally much more expensive than in rural areas, allocations are implicitly made according to socio-economic background. In addition, being aware of the criteria makes it possible for well-off students to use workaround strategies: parents will sometimes move to the catchment area of the *académie* which is most likely to get their child into their desired university shortly before they complete their school-leaving qualifications. Others use preparatory courses (*prépas*) to avoid APB allocation. Since these courses take one to two years, during which time the student will not have any income, this route too is only open to well-off students (Frouillou 2016).

2.7.5 For discussion: a process that is not independently verifiable runs the risk of systematic bias

The case study illustrates the value of the verifiability of processes by independent third parties. As in the case of the university admissions in France, criteria which appear useful at first glance can lead to systematic discrimination in a selection process. If parents' place of residence has an influence on the allocation of their child's study place, social inequalities can reproduce where place of residence and social status are correlated. This effect can also be observed in other ADM processes. For example, a long commute to work is correlated with the employee quitting early on. Companies could therefore reject or defer applications from people who live too far away. If, however, place of residence is correlated with social status, such a system results in systematic discrimination (Trindel 2016)..

It is unclear whether the link between residence and social status in the design of the admissions procedure used in France has been overlooked or consciously accepted as a trade-off in favor of other, more important issues. In both cases, the original refusal of the Ministry of Education to publish the code is problematic. In the case of an error, this would have made it more difficult for checks to be made by outsiders. If the decision was a conscious one, it would have to be addressed in a broad public debate prior to the introduction of the allocation procedure. Such normative decisions should not be hidden behind the apparent objectivity of machine decisions.

The transparency of ADM processes vis-à-vis the public can encourage workaround strategies. The recognition of the criteria in the present example resulted in a change in the behavior of those being assessed (e.g. the high uptake of potentially oversubscribed study courses without a genuine study interest). A case could therefore be made for procedures which allow for clarification and independent evaluation without the complete publication of all procedural details. The Ministry of Education's argument, on the other hand, that the publication of the source code would expose weak points and thus invite hacker attacks, only appears to be justified in the short-term. Any exploited vulnerabilities could be subsequently patched, thus improving the algorithm. The Federal Ministry of Finance has even made its algorithm for calculating VAT available on GitHub (Berne 2016).

Further opportunities	Further risks
A good procedure could ensure clarity and transparency in the allocation of school leavers to the limited (state) study places available.	In this case, the influence of place of residence on the allocation leads to discrimination via a proxy. Parental wealth and income have an influence on the place of residence of the students, therefore social status is an indirect criterion in the selection process. The example illustrates that crucial evaluation errors can occur in the design of ADM processes. The decision as to which data is collected and evaluated is not taken objectively.

The centralized allocation procedure relieves the burden on universities.	The APB process cannot be circumvented, since state institutions of higher education carry out their study place allocation centrally. This illustrates a fundamental risk of ADM processes: the ability to withdraw from them diminishes. Since the decision logic of a system with comparatively low additional effort can be applied to any number of cases, such methods favor centralization. However, this can also result in the extensive scaling of errors.

2.7.6 Situation and relevance in Germany

Although a comparable portal was launched in Germany in 2013 with hochschulstart.de, thanks to the highly individualized, decentralized decision-making criteria of German universities with regard to which students they accept, the risk of a dominant, standardized process is low. The importance of proximity to a student's place of residence is also lower in Germany due to the lack of local concentration of top universities.

In 2015, Karlsruhe Institute of Technology (KIT) developed something known as the "Admission oracle" for medicine, which calculates the individual wait time for each student for a place to study medicine. To make this possible, researchers used publicly accessible data from hochschulstart.de, including school-leaving exam grades; year of graduation; number of deferred semesters; whether the graduate had passed the medical entrance examination, and if so, with what grade; and whether (relevant) vocational training was undertaken in the meantime. The research team compared some 250,000 student profiles as well as the admission criteria for 36 universities with the relevant admission decision. An automatic learning algorithm then calculated the relevant weightings for admission using this data (Jung et al. 2015). The researchers' aim was to minimize student uncertainty about their application for a place to study medicine, thus preventing unnecessary wait times in cases where students will be turned away in the end.

A comparable attempt to open up state ADM processes by undertaking legal proceedings has not yet been documented.

2.8 Consider social interdependencies: location-specific predictions of criminal behavior

Predictive policing based on ADM processes can direct the focus of investigators to individuals (see chapter **Fehler! Verweisquelle konnte nicht gefunden werden.**). Another approach focuses on the locations that ADM processes predict – for example hotspots for specific offenses such as domestic burglaries. Well-known commercially available analysis programs which use this approach include Precobs (Institut für musterbasierte Prognosetechnik, Germany) and Predpol (Predpol, USA).

Authorities and criminologists have been working with geographical patterns of criminal activity since the nineteenth century, when this practice was developed in London. Long before software was deployed in the first ten years of the new millennium, human analysts were using crime statistics to identify high-crime areas. “For example, half of the crime in Seattle over a fourteen-year period could be isolated to only 4.5% of city streets. Similarly, researchers in Minneapolis, Minnesota found that 3.3% of street addresses and intersections in Minneapolis generated 50.4% of all dispatched police calls for service” (Gluba 2014: 5). Such retrospective analyses form the basis of evidence-based police work.

2.8.1 Output: burglary predictions for areas measuring 250 x 250 meters

Precobs software visualizes daily burglary predictions in 250 x 250 meter target areas (Brühl and Fuchs 2014). The system illustrates the predictions by changing the color of the grid squares: the anticipated rate of successive crimes is highest in red grid squares, becoming gradually lower in yellow, green and blue squares. For Precobs, a “near repeat” is when “at least two similar offenses occur within 72 hours within a defined geographical area” (Institut für musterbasierte Prognosetechnik 2014).

Predpol software, which is widely used in the United States and in Great Britain, provides predictions using a similar format. The area is divided into 150 by 150 meter squares, for each of which Predpol calculates the risk of criminal activity for the next two twelve-hour shifts, before displaying just the 20 squares with the highest risk on the map (Mohler et al. 2015: 10).

Ultimately, the police decide whether to deploy resources and how these should be distributed for each shift, based on the hotspot predictions.

2.8.2 Data and decision logic: non-personal crime characteristics

The prediction of hotspots for specific offenses is based on “near repeat” theory. This criminological approach assumes that following criminal offenses such as car theft, and domestic and car burglaries, the likelihood of further offenses in the local area will increase. Empirical studies from Great Britain, the US, the Netherlands, New Zealand and Australia show statistically significant near-repeat patterns for domestic burglaries (Ferguson 2012: 19).

Precobs analyzes a few parameters, such as the time of the offense, the type of loot (e.g. cash, till), the type of building (e.g. office, business premises, residential building) and the break-in method (e.g. lever, breaking a window by hand, or by foot) in order to recognize the patterns of serial offenders (Brühl 2014). The data are not personal, and information on offenders or victims is not included in the analysis (Schindler und Wiedmann-Schmidt 2015). Despite the clearly delineated objective and input data, Ralf Middendorf, a Precobs developer, says: “We see connections that we cannot explain” (Brühl 2014).

2.8.3 Consequences: greater police presence in risk areas

Police departments use Precobs and Predpol to set priorities when deploying patrols. In Kent in the UK and in Los Angeles for example (Mohler et al. 2015: 10), or in the Precobs pilot project in Munich: “If Precobs calculates that burglaries are expected in a certain area one day, the police will strengthen their patrol presence in that area” (Brühl and Fuchs 2014).

2.8.4 Evaluation: mixed results with regard to effectiveness

The few studies carried out under peer review on the use of location-specific predictions of criminal behavior have produced very different results. A study conducted by Predpol employees into Predpol use in Kent (UK) and Los Angeles concluded that the software correctly predicted the location of up to twice as many burglaries as human analysts: "Our results show that ETAS models predict 1.4-2.2 times as much crime compared to a dedicated crime analyst using existing criminal intelligence and hotspot mapping practice" (Mohler et al. 2015:1402). What's more, the crime rates in the hotspots identified by the Predpol software subsequently dropped after additional forces were sent to patrol these areas.

These results are not clear evidence of the efficacy of Predpol, however. With only four analysts in the control group, the quality of the human decisions is unknown, and other possible secondary effects are conceivable:

> *"But with only four human analysts of unknown effectiveness included in the study, the comparison is not wholly convincing. (…) More patrol time on ETAS hot spots could indeed be reducing crime; then again, on days when there is little crime for whatever reason, officers could have more time to visit suspect areas" (Perkowitz 2016).*

In 2011, the research-based evaluation in another pilot project using different software found no advantages of the ADM solution: "This study found no statistical evidence that crime was reduced more in the experimental districts than in the control districts" (Hunt, Saunders and Hollywood 2014).

The police in Milan use an ADM process to predict theft hotspots in some neighborhoods. The gendarmerie, which is responsible for other neighborhoods, does not use such processes. A comparison of the clearance rate in these neighborhoods shows that the clearance rate in the police area has improved by eight percentage points since the introduction of the prediction software (Mastrobuoni 2014: 7) – the study is, however, unpublished.

It is unclear whether crime actually decreases or merely shifts location due to the increased patrolling of hotspots. "Questions of the simple shift of crime by concentrating on specific areas, such as in the case of predictive policing, have not been adequately answered due to the research situation" (Gluba 2014: 11).

2.8.5 For discussion: social interdependency makes impact assessment more difficult

The accurate prediction of burglary hotspots, as in the case study, can even prevent certain offenses if there is increased police presence in these areas. The risks are smaller than for personalized predictions (Selbst 2016: 21). In this example, however, the present analyses show how complex the interactions between an ADM process and the environment can be. Even when use is very limited, the interdependences between ADM processes and their environment are highly complex. Only an analysis of the entire socio-informatic process can reveal the relationship between opportunities and risks. Questions to be answered include:

- Does location-specific "predictive policing" drive crime to other areas?
- Do systematic biases result from the fact that the predictions are based on statistics from the cases already known to the police, and the procedure therefore focuses attention on reported rather than unreported crime?
- Do criminals adapt their methods for the procedures used?

Further opportunities	Further risks
Human analysts also focus on location-based assessments, statistics, and information provided by informants. The systematization and evaluation of such sources for use in ADM processes can lead to the public discussion of decisions that would not previously have been visible (Prabhu 2015: 11).	Only certain crimes are territorial. There is a risk that more resources will be invested in the prosecution of these predictable types of offenses, due to quicker police success (Selbst 2016: 17).

2.8.6 Situation and relevance in Germany

Location-based predictions of criminality are regularly used in Europe, for example in Zurich (Baumgartner, 2015), Kent (Mohler et al. 2015) and Milan (Mastrobuoni 2014). In Germany, this form of predictive policing is currently in use or under development in 14 pilot projects and tests (Pilpul 2016).

2.9 Prevent misuse: credit scoring in the US

In the United States, three major national credit bureaus make up the market for private consumer credit ratings: TransUnion, Experian and Equifax. According to the US Consumer Financial Protection Bureau (CFPB), each of these suppliers processes around 1.3 billion updates each month for the profiles of more than 200 million consumers (Hurley and Adebayo 2016: 154). These credit bureaus provide predictions based on traditional models such as what is known as the FICO score, which exclusively uses credit history and relevant court judgments, relating to bankruptcy or foreclosure for example, as a data basis. According to estimates by the National Credit Reporting Association, these prediction models disqualify some 70 million US citizens from borrowing who do not receive a score due to a lack of data, or receive a poor repayment prediction based solely on limited information (Robinson and Yu 2014: 6). According to estimates from Experian, "unscoreables" (people with insufficient data for traditional models) include immigrants and college graduates, for example (Hurley and Adebayo 2016: 155).

2.9.1 Output: personal information removed, obligation of transparency is lost

Traditional prediction models provided by FICO and Vantagescore issue a score. The higher the value, the higher the estimate of creditworthiness. The credit bureaus themselves interpret the values for their assessments of borrowers. However, the ADM provider Vantagescore has publicly stated that a borrower in the highest level ("prime") will receive a score of between 661 and 780 (Vantagescore 2013).

The output also influences whether predictions fall within the transparency requirements of the FCRA regulation on credit information. Since the law is aimed at individuals, it does not capture aggregated marketing scores, which provide information about rows of houses, for example.

> *"Aggregated marketing scores – which are computed on a household or block level, and arguably not tied to any one consumer's identity – have become a primary way for credit bureaus to sell, and for creditors and other actors to use, consumers' credit histories to market to them with greater precision. These products often come within a hair's breadth of identifying a person. (…) In other words, it provides detailed insight into the financial characteristics of the 'group' of people in a single household – and does so putatively without triggering any of the protections of the FCRA" (Robinson and Yu 2014: 17).*

Potential lenders may also include other factors in their decision. As FICO states in information for customers: "Your credit score is calculated from your credit report. However, lenders look at many things when making a credit decision, such as your income, how long you have worked at your present job and the kind of credit you are requesting" (Fair Isaac Corporation 2017). It is unclear how often such decisions are automated and when a human decision-maker is involved. For some applications, it is obvious that the decisions have been purely machine-based and automated: for example, with online credit applications or the evaluation of waiting callers on a hotline based on their credit scoring.

2.9.2 Data and decision logic: credit providers, power suppliers, social networks

Two laws define which predictions are permissible: the Fair Credit Reporting Act (FCRA) requires that the data sold about individuals be relevant and accurate and that it may only be used for certain permitted purposes. The Equal Credit Opportunity Act (ECOA) prohibits the inclusion of protected characteristics such as race or age in credit rating systems (Robinson and Yu 2014b: 6).

Two fundamental changes have shaped the situation in the USA:

- New prediction models use data sources such as social networks and consumer profiles to predict the creditworthiness of people who were not previously given a scoring (see: Data and decision logic:).
- Credit bureaus are developing new predictions which, perhaps counterintuitively, are not covered by FCRA regulation: "marketing scores", for example, can be based on information relating to creditworthiness, however, these are not used for lending, and are instead used for things like pricing, etc. (ibid.).

Robinson and Yu (2014b: 4) distinguish between three approaches to predicting creditworthiness in the US based on the data used in ADM processes:

- Traditional prediction models use only information relating to the repayment of loans, obtained for example from credit card companies or mortgage lenders.
- Mainstream alternative models also evaluate data obtained from member companies of credit bureaus, but also consider the regular settlement of outstanding payments, e.g. utilities (electricity, water, etc.) (ibid.: 10)
- Fringe alternative models also use data not related in any way to the settlement of outstanding payments to predict creditworthiness. Depending on the model, this might include social media profiles, location data from the applicant's smartphone, information about purchasing behavior, or evaluations of how quickly a user scrolls through the information on the credit provider's website (ibid.: 13 et seq.). These models are often referred to in the US as "alternative credit decisioning tools" (ACDT).

The decision logic of the systems is difficult to understand, even for traditional prediction models. The two largest providers of traditional prediction models, FICO and Vantagescore, operate many different versions of their ADM processes for different customers (Hurley und Adebayo 2016: 155). The information offered by these two providers on how these work does not differentiate between different procedures.

- FICO states that the scoring value is based on information relating to individual behavior across five categories: past payment history, extent of the individual's credit history, amount of existing loans, nature of existing loans, and amount of new loan requests. The report quantifies the weighting of the categories with percentages, but relativises them. The weighting is different for each scoring: "… it's impossible to measure the exact impact of a single factor in how your credit score is calculated without looking at your entire report. Even the levels of importance shown in the FICO Scores chart are for the general population, and will be different for different credit profiles" (Fair Isaac Corporation 2017).
- Vantagescore specifies the same basic requirements for the prediction of creditworthiness as FICO, but also considers exhausted credit lines. Borrowers are given behavioral tips for obtaining a good score. For example: "Maintain a mix of accounts (credit cards, auto, mortgage) over time to improve your score. Prime consumers have an average of 13 loans. Typically the oldest loan is more than 15 years old" (Vantagescore 2013) This is a possible reference to the decision logic: the ADM process may compare the similarity of the profile of a potential borrower with the profiles of borrowers who make reliable repayments.

An important factor in all established models is the age of the available data: "Credit files that have gone more than six months with no reported activity are considered 'stale' by the FICO algorithm, and will not produce a score" (Robinson und Yu 2014c: 17). This may offer some explanation as to why approximately 70 million US citizens are not given a credit prediction due to a lack of data.

Citron and Pasquale criticize the fact that this information makes the decision logic of the procedures unclear and is of no help to the individual concerned: "Looking forward, a consumer has no idea, for example, whether paying off a debt that is sixty days past due will raise her score. The industry remains highly opaque, with scored individuals unable to determine the exact consequences of their decisions" (2014: 18).

The decision logic of alternative prediction models are even more difficult to explain on the basis of publicly available information. ZestFinance, a provider of new prediction models, processes up to 10,000 data points per credit applicant, including mobile phone payments and even behavioral data, including "unusual observations, such as whether applicants use proper spelling and capitalization on their application form, how long it takes them to read it, and whether they bother to look at the terms and conditions" (O'Neil 2016a: 144). The founder and managing director of the provider Douglas Merrill explained in an interview that the decision logic used by his company is not always clearly comprehensible in each individual case: "Merrill acknowledges that in many cases, there's no explanation for why a particular data point helps or hurts a credit score. For instance, borrowers who write in all-caps are

riskier, the firm's credit scoring system discovered after underwriting thousands of loans. 'We don't know why. It just is', said Merrill" (Koren 2015).

2.9.3 Consequences: scoring affects insurance premiums and applicant selection

Many US citizens will not be granted loans – or will be granted only very expensive loans – because their profiles do not contain sufficient data for predictions to be made.

This situation may be made all the more serious by the fact that the procedures originally developed for predicting the default probability of loans are also used as indirect indicators for altogether separate questions. Some examples:

Insurance: predictions of creditworthiness influence the cost of car insurance in many US states. According to a price analysis by the consumer organization Consumer Reports, in some cases a below-average scoring value can increase the cost of premiums by up to 1301 dollars per year, regardless of driving behavior (Consumer Reports 2015). This practice is permitted in all US states, with the exception of California, Hawaii and Massachusetts. In some states, the price premiums for people with bad credit predictions may be even higher than for those with drink-driving convictions (O'Neil 2016a: 149).

Candidate selection: According to a survey conducted by the US Society for Human Resource Management (SHRM) in 2012, 47 percent of personnel departments used credit ratings for selecting candidates: 34 percent of respondents used these for some candidates, while 13 percent checked the credit rating of all candidates (Society for Human Resource Management 2012: 8). According to a US aid organization, jobs in the low-wage sector are particularly affected: "The people contacting her group, she said, are 'mostly lower-wage workers', especially those applying to big retail chains" (Rivlin 2013).

Those affected are faced with unemployment:

> *"Among survey respondents who are unemployed, 1 in 4 says that a potential employer has requested to check their credit report as part of a job application. 1 in 10 survey respondents who are unemployed have been informed that they would not be hired for a job because of the information in their credit report. Among job applicants with blemished credit histories, 1 in 7 has been advised that they were not being hired because of their credit" (Traub 2013: 9).*

2.9.4 Evaluation: virtually no independent studies, evidence of age discrimination

No current, independent, representative and systematic investigation of the quality of the different prediction models has been undertaken. The 2012 investigation published by the US Federal Reserve Bank on the possible disparate impact of the scoring values is based on data sets from the years 2003 and 2004 (Avery, Brevoort and Canner 2012: 3). The results from these 300,000 data sets along with demographic information indicate that there is no difference in treatment on the basis of ethnicity or gender, however, there is evidence that the processes discriminate against young people:

> *"Our results provide little or no evidence that the credit characteristics used in credit history scoring models operate as proxies for race or ethnicity. (…) We do, however, find some evidence that credit characteristics associated with the length of an individual's credit history (…) may have a disparate impact by age. In particular, we find that the predictiveness of this credit characteristic increases when the credit scoring model is estimated in an age neutral environment" (Avery, Brevoort und Canner 2012: 27).*

This result illustrates a weakness of traditional prediction models: those who are lacking in certain areas (age, data, payment history) are given more expensive loans – or no loans at all in cases of insufficient data.

The reliability of traditional prediction models was investigated by the US Federal Trade Commission (FTC) in 2012. In the context of this investigation, 1,001 participants assessed their credit reports (2,968 in total). Results: 26 percent of the participants discovered errors in their information; 21 percent achieved a correction by contesting

this information; only 13 percent received a different scoring value following this correction (Federal Trade Commission et al. 2012: 5).

The Policy and Economic Research Council (PERC) examined how the inclusion of payments to electricity and telecommunications providers affects credit ratings. The assessment of mainstream alternative models was positive. 25 percent of those examined in the study, who had previously had insufficient payment information for traditional prediction models ("thin-file population"), were classified into a better risk category following the inclusion of supplier data. Only six percent would have received a poorer classification to a lower risk level with the inclusion of this expanded information (Turner et al. 2012: 6).

Providers of fringe alternative prediction models promise to allow far more people access to credit. However, there have been no independent, representative and systematic evaluations on the models used, as explained by Robinson and Yu (2014):

> *"Less still is known about the financial startup scene, which relies on even more exotic data. For example, ZestFinance boasts that its "big data underwriting model provides a 40% improvement over the current best-in-class industry score." But it is unclear how accurate the "best-in-class industry score" actually is for Zest's target population of consumers, much less how ZestFinance measures up to that benchmark" (a.a.O.: 27).*

2.9.5 For discussion: misuse of scoring values transfers participation effects

The case study highlights the implications of a concept such as creditworthiness being used as a proxy value for the classification of people in many other areas of life. In some cases, the concept of creditworthiness is clearly misused, for example to offer expensive credit cards to people with bad scores or to place their calls to call centers at the bottom of the list (O'Neil 2016b: 132 ff.).

The automated processing of ADM procedures serves to facilitate such examples of misuse. The legal framework in the United States has led to the automated procedures of insurance companies translating personal credit ratings into so-called marketing scores. The current prevailing perception is that these are not subject to credit regulation in the US, however, they are ultimately sufficient for discriminatory treatment on an individual basis by companies (Robinson and Yu 2014b: 6).

The credit score offers a clear and simple answer – in the examples of misuse, however, this is not the answer to the pertinent question. Whether an individual's credit score actually tells us anything about their work performance or risk of accident is more than doubtful. Such an approach merely transfers disadvantages from one sphere of life to another. These examples show that the effects of ADM processes on individuals cannot be evaluated solely on the basis of data protection logic.

Further opportunities	Further risks
ADM-based alternative models for the prediction of creditworthiness can provide people with access to loans that would previously have been evaluated by traditional models as high risk due to a lack of information about their payment behavior (Turner et al. 2012: 23; Hurley and Adebayo 2016: 156).	These alternative models are rarely subject to independent research (Robinson and Yu 2014: 27). They may also systematically discriminate against certain groups of people, for example people with writing disabilities in cases where spelling errors in the loan application are considered a sign of increased risk of default (O'Neil 2016a: 144).

2.9.6 Situation and relevance in Germany

In Germany, credit bureaus provide information on companies and individuals. The Federal Data Protection Act permits scoring under certain conditions. For example, the data used to calculate the probability value must be demonstrably significant for the calculation of the probability of the behavior in question, on the basis of a scientifically recognized mathematical-statistical method; it is not permitted for predictions to be made on the basis of just address data (§ 28b Federal Data Protection Act, BDSG).

In 2014, the German Federal Supreme Court (BGH) ruled that people who have been assessed are not entitled to know precisely how the evaluation of their future behavior has been calculated. In the proceedings, a woman who had not been granted a loan after being assessed by the German private credit bureau Schufa filed a suit for information. In her view, the data provided by Schufa did not meet the legal requirements. The BGH ruled that the "abstract method of score value calculation need not be communicated" and that the following information, in addition to other data, is protected as a trade secret: "(...) the calculation parameters incorporated in the first step of the score formula, such as the statistical values used, the weighting of individual calculation elements in the determination of the probability value, and the formation of possible comparison groups as a basis for the scoring" (Bundesgerichtshof 2014: 1). A constitutional appeal against this judgment is currently pending ("Schufa-Klägerin zieht vor Verfassungsgericht" 2014).

3 Conclusion

The case studies show that ADM processes influence decisions about people, as well as how this occurs. When machines evaluate us and their predictions impact civil liberties – as in the case of use by the courts or by the police – or equal opportunities – as in the selection of candidates or assessments of creditworthiness – society must consider the fairness and participation impacts of these processes.

It is worth looking closely at the specific context, because not every ADM process poses the same risks. The social requirements on ADM processes can vary, depending on the impact of these processes on society and on basic individual rights. Automatic spell-checkers or navigation systems have different effects on a person's life than systems that assign credit risk or risk of criminal behavior.

The aggregation of the opportunities and risks from the case studies (see Table 2: Abstract overview of opportunities and risks from the case studies on the following page) points to some overarching participation-critical factors of ADM processes. These are based on different aspects of the socio-informatic process as a whole, and affect different levels. Three examples:

- *Designing ADM processes at the micro and meso level:* the selection of data and determination of criteria at the beginning of the development process may involve normative assumptions that affect fundamental social issues in certain cases.

- *Provider and operator structure at the macro level:* The diversity of different ADM processes and operators can strengthen participation (e.g. by means of creditworthiness predictions aimed at people who were previously excluded by the system), as well as expand options for withdrawal and increase falsifiability. Accordingly, monopolistic structures increase the risk of the individual being "excluded from the system".

- *Dealing with ADM predictions at the micro, meso and macro level:* The interaction of technology, society and individuals has a huge influence on the handling of, and thus the impact of, algorithms. Important questions therefore include: How do people (both ADM developers and users, as well as the general population) deal with the automated predictions? Do these processes include opportunities to refute ADM predictions?

Further systematic analyses of possible sources of error at different stages of ADM processes are required here – from the definition of objectives and the measurability of the concepts, to data collection, the selection of algorithms, and the embedding of the process in the social context (cf. Zweig 2016). Quality criteria for ADM processes are required which include all levels and steps. The needs for action highlighted in the introduction can serve as an initial basis in this regard (cf. Table 1: Need for action in algorithmic decision-making processes).

An important, almost fundamental quality characteristic should once again be highlighted at this point: an analysis of the opportunities, risks and societal consequences was only possible because independent third parties were able to examine the quality of the machine decisions. Institutions such as the investigative journalism organization Propublica (see chapter 2.1 Recidivism predictions used in the legal system), the US Government Accountability Office (see chapter 2.5 Automatic face recognition systems) or the student representation organization Droits des lycéens (see chapter 2.7 University admissions in France) have invested time and money in data collection, data analysis and legal disputes, in order to attempt to explain the algorithms used and to provide some transparency in this regard. The question of whether a social debate about the impact of specific ADM processes is even possible currently depends on institutions such as these. This needs to change since the verifiability and transparency of algorithmic decisions form an indispensable knowledge basis for solution-oriented social discourse, with the aim of ensuring that ADM processes are designed for increased participation and that machine decisions serve people.

Table 2: Abstract overview of opportunities and risks from the case studies (source: own representation)

Dimension	Opportunities	Risks
Normative assumptions	When an ADM process is designed, normative decisions (e.g. about fairness criteria) must be made before the process is used. This offers an opportunity to discuss ethics issues thoroughly and publically at the very start and to document decisions.	ADM processes can hide normative decisions behind their design. If discussion is only possible once the design phase is complete, any normative assumptions are more likely to be accepted as unalterable.
Data	Software can analyze a much greater volume of data than humans can, thereby identifying patterns and answering certain questions faster, more precisely and less expensively.	The data used for an ADM process can contain distortions that are seemingly objectified by the process itself. If the causalities behind the correlations are not verified, there is a significant danger that unintentional, systematic discrimination will become an accepted part of the process.
Consistency of application	Algorithm-based predictions apply the predetermined decision-making logic to each individual case. In contrast to human decision makers, software does not have good and bad days and does not in some cases arbitrarily use new, sometimes inappropriate criteria.	In exceptional cases, there is usually no possibility for assessing unexpected relevant events and reacting accordingly. ADM systems unfailingly make use of any incorrect training data and faulty decision-making logic in each and every case.
Scalability	Software can be applied to an area of application that is potentially many times larger than what a human decision maker can respond to, since the decision-making logic used in a system can be applied at very low cost to a virtually limitless number of cases.	ADM processes are easily scalable, which can lead to a decrease in the range of such processes that are or can be used, and to machine-based decisions being made much more often and in many more instances that might be desirable from a societal point of view.
Verifiability	Data-driven and digital systems can be structured in a way that makes them clear and comprehensible, allows them to be explained and independently verified, and provides the possibility of forensic data analysis.	Because of process design and operational model, independent evaluations are often only possible, comprehensible or institutionalized to a limited degree.
Adaptability	ADM processes can be adapted to new conditions by using either new training data or self-learning systems.	The symmetry of the adaptability in all directions depends on how the process is designed. One-sided adaptation is also possible.
Efficiency	Having machines evaluate large amounts of data is usually cheaper	Efficiency gains achieved through ADM processes can hide the fact

	than having human analysts evaluate the same amount.	that the absolute level of available resources is too low or inadequate.
Personalization	ADM processes can democratize access to personalized products and services that for cost-related reasons were previously only available to a limited number of people. For example, before the internet, numerous research assistants and librarians were required to provide the breadth and depth of information that results from a single search-engine query.	When ADM processes are the main tools used for the mass market, only a privileged few have the opportunity to be evaluated by human decision makers, something that can be advantageous in non-standard situations when candidates are being preselected or credit scores awarded.
Human perception of machine-based decisions	ADM processes can be very consistent in making statistical predictions. In some cases, such predictions are more reliable than those made by human experts. This means software can serve as a supplementary tool which frees up time for more important activities.	People can view software-generated predictions as more reliable, objective and meaningful than other information. In some cases this can prevent people from questioning recommendations and predictions or can result in their reacting to them only in the recommended manner.

4 Bibliography

Algorithmwatch (2016). “Das ADM-Manifest”. https://algorithmwatch.org/das-adm-manifest-the-adm-manifesto/ (Download 19.2.2017).

Angwin, Julia, Lauren Kirchner, Jeff Larson and Surya Mattu (2016). “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks”. 23.5.2015. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Download 11.12.2016).

Avery, Robert B., Kenneth P. Brevoort and Glenn Canner (2012). “Does Credit Scoring Produce a Disparate Impact?”. *Real Estate Economics* 40 s1. S65–S114.

Barry-Jester, Anna M., Ben Casselman and Donna Goldstein (2015). “The New Science of Sentencing”. *The Marshall Project.* 4.4.2015. https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.xXEp6R5rD (Download 24.4.2017).

Baumgartner, Fabian (2015). “Deutliche Zunahme im Jahr 2015: Wieder mehr Einbrüche in der Stadt Zürich“ (Marked increase in 2015. More break-ins again in Zurich). *Neue Zürcher Zeitung* 27.10.2015. http://www.nzz.ch/zuerich/stadt-zuerich/wieder-mehr-einbrueche-in-der-stadt-zuerich-1.18636278 (Download 24.4.2017).

Berne, Xavier (2016). “Pour dévoiler l’algorithme d’Admission Post-Bac, l’Éducation nationale opte pour le papier”. 19.10.2015. https://www.nextinpact.com/news/101809-pour-devoiler-l-algorithme-d-admission-post-bac-l-education-nationale-opte-pour-papier.htm (Download 8.2.2017).

Brühl, Jannis (2014). “Ermitteln mit ‚Predictive Policing‘-Algorithmen: Polizei-Software soll Verbrechen voraussagen” (Investigating with predictive policing: police software to forecast crime). *Süddeutsche Zeitung* 12.9.2014. http://www.sueddeutsche.de/digital/ermitteln-mit-predictive-policing-algorithmen-polizei-software-soll-die-zukunft-voraussagen-1.2121942 (Download 24.4.2017).

Brühl, Jannis, and Florian Fuchs (2014). “Polizei-Software zur Vorhersage von Verbrechen: Gesucht: Einbrecher der Zukunft” (Police software to forecast crime: sought: burglars of the future). *Süddeutsche Zeitung* 12.9.2014. http://www.sueddeutsche.de/digital/polizei-software-zur-vorhersage-von-verbrechen-gesucht-einbrecher-der-zukunft-1.2115086 (Download 24.4.2017).

Bundesgerichtshof (2014). “Urteil des VI. Zivilsenats vom 28.1.2014 – VI ZR 156/13” (Ruling of 6th Civil Senate). 28.1.2014. http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&sid=6e60182ecc6717735edc41833c989561&nr=66910&pos=0&anz=1 (Download 24.4.2017).

Bundespolizei (2015). “EasyPASS”. http://www.easypass.de/EasyPass/DE/EasyPASS-RTP/rtp_node.html (Download 19.2.2017).

Chicago Department of Public Health (2016). “Fifty Years Fighting Lead in Chicago The Plan for a Lead Free Generation”. 5.7.2016. https://www.cityofchicago.org/content/dam/city/depts/cdph/food_env/general/Lead_Poison_Prevention_Program/CDPH_LeadBrochure_10172016.pdf (Download 24.4.2017).

Christin, Angele, Alex Rosenblat and Danah Boyd (2015). “Courts and Predictive Algorithms”. *Data & CivilRight: Criminal Justice and Civil Rights Primer*. 27.10.2015. http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf (Download 24.4.2017).

Citron, Danielle Keats, and Frank A. Pasquale (2014). “The Scored Society: Due Process for Automated Predictions”. *Washington Law Review* (89) 1. 1–33.

Consumer Reports (2015). “How a Credit Score Affects Your Car Insurance”. http://www.consumerreports.org/cro/car-insurance/credit-scores-affect-auto-insurance-rates/index.htm#creditmap (Download 24.4.2017).

Danziger, Shai, Jonathan Levav and Liora Avnaim-Pesso (2011). “Extraneous factors in judicial decisions”. *Proceedings of the National Academy of Sciences* (108) 17. 6889–6892. https://doi.org/10.1073/pnas.1018033108 (Download 24.4.2017).

Davey, Monica (2016). “Chicago Police Try to Predict Who May Shoot or Be Shot”. *The New York Times* 23.5.2016. http://www.nytimes.com/2016/05/24/us/armed-with-data-chicago-police-try-to-predict-who-may-shoot-or-be-shot.html (Download 24.4.2017).

de Coustin, Paul (2016). “APB: les explications du ministère ne lèvent pas tous les doutes”. *Le Figaro* 6.2.2016. http://etudiant.lefigaro.fr/les-news/actu/detail/article/l-algorithme-d-admission-post-bac-se-devoile-20621/ (Download 8.2.2017).

Desmarais, Sarah L., and Jay P. Singh (2013). “Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States”. Lexington, KY: Council of State Governments. https://csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf (Download 24.4.2017).

Dunleavy, Eric M. (2016). “Written Testimony of Eric M. Dunleavy, PhD, Director”. Washington D.C. https://www.eeoc.gov/eeoc/meetings/10-13-16/dunleavy.cfm (Download 24.4.2017).

Eckhardt, Andres, Tim Weitzel, Sven Laumer, Christian Maier, Caroline Oehlhorn, Jakob Wirth and Christoph Weinert (2016). “Techniksprung in der Rekrutierung” (Technological leap in recruitment). https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/isdl/Recruiting_Trends_2016_-_Techniksprung_in_der_Rekrutierung_v_WEB.PDF (Download 24.4.2017).

Fair Isaac Corporation (2017). “How FICO Credit Score is Calculated”. http://www.myfico.com/crediteducation/whatsinyourscore.aspx (Download 30.1.2017).

Federal Trade Commission et al. (2012). “Report to Congress Under Section 319 of the Fair and Accurate Credit Transactions Act of 2003”. December.

Felten, Ed, National Science and Technology Council und Committee on Technology (2016). “Preparing for the Future of Artificial Intelligence”. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (Download 24.4.2017).

Ferguson, Andrew Guthrie (2012). “Predictive Policing and Reasonable Suspicion”. *Emory Law Journal* 259. https://papers.ssrn.com/abstract=2050001 (Download 24.4.2017).

Flores, Anthony W., Kristin Bechtel and Christopher T. Lowenkamp (2016). “False positives, false negatives, and false analyses: A rejoinder to ‚machine bias: There’s software used across the country to predict future criminals and it’s biased against blacks”. *Unpublished manuscript*. https://www.researchgate.net/profile/Christopher_Lowenkamp/publication/306032039_False_Positives_False_Negatives_and_False_Analyses_A_Rejoinder_to_Machine_Bias_There’s_Software_Used_Across_the_Country_to_Predict_Future_Criminals_And_it’s_Biased_Against_Blacks/links/57ab619908ae42ba52aedbab.pdf (Download 24.4.2017).

Frouillou, Leila (2016). “Post-bac admission: an algorithmically constrained ‘free choice’”. http://www.jssj.org/wp-content/uploads/2016/07/JSSJ10_3_VA.pdf (Download 24.4.2017).

“Further Education in France”. *Angloinfo*. http://www.angloinfo.com/how-to/france/family/schooling-education/further-education (Download 8.2.2017).

Garvie, Clare, Alvaro M. Bedoya and Jonathan Frankle (2016). *The Perpetual Line-Up*. Washington D.C.: Center on Privacy and Technology at Georgetown Law. https://www.perpetuallineup.org/sites/default/files/2016-12/The%20Perpetual%20Line-Up%20-%20Center%20on%20Privacy%20and%20Technology%20at%20Georgetown%20Law%20-%20121616.pdf (Download 24.4.2017).

Gluba, Alexander (2014). *Predictive Policing – taking stock*. LKA Niedersachsen: Hannover. https://netzpolitik.org/wp-upload/LKA_NRW_Predictive_Policing.pdf (Download 24.4.2017).

Graveleau, Séverin (1 June 2016). "Admission post-bac, l'algorithme révélateur des failles de l'université". *Le Monde.fr* 1.6.2016. http://www.lemonde.fr/campus/article/2016/06/01/admission-post-bac-l-algorithme-revelateur-des-failles-de-l-universite_4929949_4401467.html (Download 24.4.2017).

Hannah-Moffat, Kelly Hannah, Paula Maurutto and Sarah Turnbull (2009). "Negotiated Risk: Actuarial Illusions and Discretion in Probation". *Canadian Journal of Law and Society* (24) 03. 391–409. https://doi.org/10.1017/S0829320100010097 (Download 24.4.2017).

Hawthorne, Michael (2015). "Could Chicago prevent childhood lead poisoning before it happens?". *Chicago Tribune* 16.7.2015. http://www.chicagotribune.com/news/ct-lead-poisoning-solutions-20150707-story.html (Download 3.1.2017).

Horton, Michelle (2016). "Stanford scientists combine satellite data, machine learning to map poverty". *Stanford News Service* 18.8.2016. http://news.stanford.edu/press-releases/2016/08/18/combining-satellg-to-map-poverty/ (Download 3.1.2017).

Hunt, Priscilla, Jessica M. Saunders and John S. Hollywood (2014). *Evaluation of the Shreveport predictive policing experiment*. Santa Monica CA: RAND Corporation.

Hurley, Mikella, and Julius Adebayo (2016). "Credit Scoring in the Era of Big Data". *Yale JL & Tech.* 18. 148–275.

Institut für musterbasierte Prognosetechnik (2014). "Near Repeat Prediction". http://www.ifmpt.de/prognostik/ (Download 9.1.2017).

Jean, Neil, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell and Stefano Ermon (2016). "Combining satellite imagery and machine learning to predict poverty". *Science* (353) 6301. 790–794.

Johnson, Eddie T. (July 14, 2016). "Special Order S09-11 Strategic Subject List (SSL) Dashboard". http://directives.chicagopolice.org/directives/data/a7a57b85-155e9f4b-50c15-5e9f-7742e3ac8b0ab2d3.html (Download 24.4.2017).

Jung, Dominik, Lorenz Kemper, Benedikt Kaempgen and Achim Rettinger (2015). "Predicting the Admission into Medical Studies in Germany: A Data Mining approach. Open Access at KIT". https://doi.org/10.5445/IR/1000045460 (Download 24.4.2017).

Koren, James Rufus (2015). "Some lenders are judging you on much more than finances". *Los Angeles Times* 19.12.2015. http://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html (Download 24.4.2017).

Kurjuweit, Klaus (2017). "Berliner Bahnhof: Bahn testet intelligente Videoüberwachung am Südkreuz" (German railway tests video surveillance at Berlin Südkreuz). *Der Tagesspiegel* 20.2.2017. http://www.tagesspiegel.de/berlin/berliner-bahnhof-bahn-testet-intelligente-videoueberwachung-am-suedkreuz/19413266.html (Download 21.2.2017).

Lefauconnier, Natacha (2016). “Leïla Frouillou: ‘APB promeut un libre choix d’études tout en étant socialement inégalitaire‘“. *Educpros* 16.6.2016. http://www.letudiant.fr/educpros/entretiens/leila-frouillou-apb-promeut-un-libre-choix-d-etudes-tout-en-etant-socialement-inegalitaire.html (Download 9.2.2017).

Lischka, Konrad (2015). “Wie die KI-Debatte falsch läuft und wo Software heute teilautonom entscheidet” (How the AI debate goes wrong and where software already makes partially autonomous decisions). 14.6.2015. http://www.konradlischka.info/2015/06/blog/wie-die-ki-debatte-falsch-laeuft-und-was-software-heute-schon-autonom-entscheidet/ (Download 24.4.2017).

Mastrobuoni, Giovanni (2015). “Crime is terribly revealing: Information technology and police productivity”. Unpublished Paper. http://cep.lse.ac.uk/conference_papers/01_10_2015/mastrobuoni.pdf (Download 24.4.2017).

Mohler, George O., Martin B. Short, Sean Malinowski, Mark Johnson, George E. Tita, Andrea L. Bertozzi P. Jeff Brantingham (2015). “Randomized controlled field trials of predictive policing”. *Journal of the American Statistical Association* (110) 512. 1399–1411. https://doi.org/http://dx.doi.org/10.1080/01621459.2015.1077710 (Download 24.4.2017).

Morgeson, Frederik P., Michael L. Campion, Robert L. Dipboye, John R. Hollenbeck, Kevin Murphy and Neal Schmitt (2007). “Reconsidering the use of personality tests in personnel selection contexts”. *Personnel psychology* (60) 3. 683–729.

Morrison, Lennox (2017). “Speech analysis could now land you a promotion”. http://www.bbc.com/capital/story/20170108-speech-analysis-could-now-land-you-a-promotion (Download 20.1.2017).

Northpointe (2015). “Practitioners Guide to COMPAS Core”. http://images.google.de/imgres (Download 24.4.2017).

O’Neil, Cathy (2016a). *Weapons of math destruction: how big data increases inequality and threatens democracy* (First edition). New York: Crown.

O’Neil, Cathy (2016b). *Weapons of math destruction: how big data increases inequality and threatens democracy.* ed. 1 New York NY: Crown.

O’Neil, Cathy (2016c). “How algorithms rule our working lives”. *The Guardian* 1.9.2016. https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives (Download 24.4.2017).

Patel, Prachi (2016). “Fighting Poverty With Satellite Images and Machine-Learning Wizardry”. 18.8.2016. http://spectrum.ieee.org/tech-talk/aerospace/satellites/fighting-poverty-with-satellite-data-and-machine-learning-wizardry (Download 2.1.2017).

Pennsylvania Commission on Sentencing (2016). “Risk Assessment Project”. http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment (Download 14.12.2016).

Perkowitz, Sidney (2016). “Should we trust predictive policing software to cut crime?”. 27.10.2016. https://aeon.co/essays/should-we-trust-predictive-policing-software-to-cut-crime (Download 12.1.2017).

Pilpul, Martin (2016). “Wo Predictive Policing eingesetzt wird” (Where predictive policing is used). December 2016. https://blog.pilpul.me/wo-predictive-policing-eingesetzt-wird/ (Download 24.4.2017).

Plass-Fleßenkämper, Benedikt (2016). “Automatische Gesichtserkennung gegen den Terror – kann das funktionieren?” (Automatic face-recognition against terror - can it work?). *wired.de* 25.8.2016. https://www.wired.de/collection/tech/automatische-gesichtserkennung-gegen-den-terror-kann-das-funktionieren (Download 12.2.2017).

Potash, Eric, Joe Brew, Alexander Loewi, Subhanrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Read Mansour and Rayid Ghani (2015). "Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning". *KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2039–2047. https://doi.org/10.1145/2783258.2788629 (Download 24.4.2017).

Prabhu, Robindra (2015). *Predictive policing – Can data analysis help the police to be in the right place at the right time?* Oslo: Teknologirådet. https://teknologiradet.no/wp-content/uploads/sites/19/2013/08/Predicitve-policing.pdf (Download 24.4.2017).

Revell, Timothy (2016). "Concerns as face recognition tech used to 'identify' criminals". *New Scientist* 12.1.2016. https://www.newscientist.com/article/2114900-concerns-as-face-recognition-tech-used-to-identify-criminals/ (Download 24.4.2017).

Rivlin, Gary (2013). "Employers Pull Applicants' Credit Reports". *The New York Times* 11.5.2013. http://www.nytimes.com/2013/05/12/business/employers-pull-applicants-credit-reports.html (Download 24.4.2017).

Robinson, David, and Logan Koepke (2016). "Stuck in a Pattern – Early evidence on "predictive policing" and civil rights". *Upturn* August 2016. https://www.teamupturn.com/reports/2016/stuck-in-a-pattern (Download 24.4.2017).

Robinson, David, and Harlan Yu (2014). "Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace". https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf (Download 24.4.2017).

Saunders, Jessica (2016). "Pitfalls of Predictive Policing". *RAND*. http://www.rand.org/blog/2016/10/pitfalls-of-predictive-policing.html (Download 9.12.2016).

Saunders, Jessica, Priscilla Hunt and John S. Hollywood (2016). "Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot".*Journal of Experimental Criminology* (12) Sept. 347–371. https://doi.org/10.1007/s11292-016-9272-0 (Download 24.4.2017).

Schindler, Jessica, and Wolf Wiedmann-Schmidt (2015). "Kriminalität: Im roten Bereich" (Crime: in the red zone). *Der Spiegel* 10/2015. http://www.spiegel.de/spiegel/print/d-132040367.html (Download 24.4.2017).

Schneider, Jan, Ruta Yemane and Martin Weinmann (2014). "Diskriminierung am Ausbildungsmarkt: Ausmaß, Ursachen und Handlungsperspektiven" (Discrimination on the education market: extent, causes and plans for action). Berlin: Forschungsbereich beim Sachverständigenrat deutscher Stiftungen für Integration und Migration (SVR). http://www.svr-migration.de/wp-content/uploads/2014/11/SVR-FB_Diskriminierung-am-Ausbildungsmarkt.pdf (Download 24.4.2017).

"Schufa-Klägerin zieht vor Verfassungsgericht" (Constitutional appeal for Schufa plaintiff). *Spiegel online* 4.11.2014. http://www.spiegel.de/wirtschaft/soziales/schufa-klaegerin-reicht-vor-verfassungsgericht-beschwerde-ein-a-964030.html (Download 24.4.2017).

Scoring (2010). "§ 28b Bundesdatenschutzgesetz Dritter Abschnitt – Datenverarbeitung nicht-öffentlicher Stellen und öffentlich-rechtlicher Wettbewerbsunternehmen (§§ 27–38a), Erster Unterabschnitt – Rechtsgrundlagen der Datenverarbeitung (§§ 27–32)" (§ 28b Federal Data Protection Act, Section Three - Data processing by non-public bodies and undertakings governed by public law which compete on the market (§§ 27–38a), First Subsection – Legal basis to data processing (§§ 27–32))

Selbst, Andrew D. (2016). *Disparate Impact in Big Data Policing.* (SSRN Scholarly Paper No. ID 2819182). Rochester NY: Social Science Research Network. http://papers.ssrn.com/abstract=2819182 (Download 24.4.2017).

Smedley, Tim (2014). “Forget the CV, data decide careers”. *Financial Times* 9.7.2014. https://www.ft.com/content/e3561cd0-dd11-11e3-8546-00144feabdc0 (Download 24.4.2017).

Society for Human Resource Management (2012). “SHRM Survey Findings: Background Checking – The Use of Credit Background Checks in Hiring Decisions”. https://perma.cc/MMG9-QF4M (Download 24.4.2017).

Steinhart, David (2006). *Juvenile detention risk assessment: A practice guide to juvenile detention reform*. Band 1. Baltimore MD: Annie E Casey Foundation.

Stromboni, Camille (2017). “APB : le gouvernement recule sur le tirage au sort à l’entrée à l’université”. *Le Monde.fr*. 18.1.2017. http://www.lemonde.fr/campus/article/2017/01/18/apb-le-gouvernement-recule-sur-le-tirage-au-sort-a-l-entree-a-l-universite_5064779_4401467.html (Download 24.4.2017).

“Terrorbekämpfung: De Maizière will Gesichtserkennung und Rucksackverbote” (Fighting terror: De Maizière wants facial recognition and rucksack ban). *Die Zeit* 21.8.2016. http://www.zeit.de/politik/deutschland/2016-08/terrorbekaempfung-thomas-de-maiziere-gesichtserkennung-flughaefen (Download 24.4.2017).

The Demographic and Health Surveys Program (2014). “Wealth Index Construction”. http://www.dhsprogram.com/topics/wealth-index/Wealth-Index-Construction.cfm (Download 3.1.2017).

The Leadership Conference on Civil and Human Rights, American Civil Liberties Union, Brennan Center for Justice, Center for Democracy, Technology, Center for Media Justice, Color of Change, Data&Society, Demand Progress, Electronic Frontier Foundation, freepress, media mobilizing project, 18MR.org, National Hispanic Media Coalition (NHMC), OpenMIC, Open Technology Institute and Public Knowledge (2016). “Predictive Policing Today: A Shared Statement of Civil Rights Concern”. 31.8.2016. http://civilrightsdocs.info/pdf/FINAL_JointStatementPredictivePolicing.pdf (Download 24.4.2017).

Thompson, Madeleine (2016). “The French Educational Algorithm of Inefficiency”. *Brown Political Review* 11.8.2016. http://www.brownpoliticalreview.org/2016/11/french-educational-algorithm/ (Download 24.4.2017).

Traub, Amy (2013). “Discredited: How employment credit checks keep qualified workers out of a job”. *Demos* 7.

Trindel, Kelly (2016). “Written Testimony of Kelly Trindel”. Washington DC. https://www.eeoc.gov/eeoc/meetings/10-13-16/trindel.cfm#fn6. (Download 24.4.2017).

Turner, Michael A., Patrick D. Walker, Chaudhuri Sukanya and Robin Varghese (2012). *A New Pathway to Financial Inclusion: Alternative Data, Credit Building, and Responsible Lending in the Wake of the Great Recession*. Durham NC: Policy & Economic Research Council.

United States Government Accountability Office (2016). “FACE RECOGNITION TECHNOLOGY FBI Should Better Ensure Privacy and Accuracy”. (No. GAO-16-267). http://www.gao.gov/products/GAO-16-267 (Download 24.4.2017).

Vantagescore (2013). “What influences your VantageScore Credit Score?”. https://www.vantagescore.com/pdf/VantageScore%20Infographic%2005.pdf (Download 24.4.2017).

Weber, Lauren, und Elizabeth Dwoskin (2014). “Are Workplace Personality Tests Fair?”. *Wall Street Journal* 30.9.2014. http://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257 (Download 24.4.2017).

World Bank (2015). “FAQs: Global Poverty Line Update”. http://www.worldbank.org/en/topic/poverty/brief/global-poverty-line-faq (Download 3.1.2017).

Zweig, Katharina Anna (2016). “2. Arbeitspapier: Überprüfbarkeit von Algorithmen“. (2nd Working Paper: Verifiability of algorithms) 7.7.2016. http://algorithmwatch.org/zweites-arbeitspapier-ueberpruefbarkeit-algorithmen/ (Download 24.4.2017).

5 Executive Summary

Processes of algorithmic decision-making (ADM) now evaluate people in many areas of life. ADM processes have been used for years to categorize people, without any real discussion of whether those processes are fair or how they can be explained, verified or corrected. One potential reason for this is that the systems have little to do with artificial intelligence (AI) as it appears in science fiction. People often associate AI with qualities exhibited by fictional characters like HAL 9000 or Wintermute: intentionality and consciousness. Yet, until now, powerful AIs of this sort have only been found in literary works and films, and have nothing to do with the systems presented in this collection of case studies. The latter, however, already play a significant role in deciding legal matters, approving loans, admitting students to university, determining where and when police officers are on duty, calculating insurance rates and assisting customers who call service centers. All are programs which are specially designed to address specific problems and which impact the lives of many people. It's not about the future according to science fiction, it's about everyday reality today.

The nine case studies presented in this working paper demonstrate the opportunities and risks associated with such processes. The journey begins in the United States, with applications that we found only there, such as algorithms used in the criminal justice system to predict recidivism. Software-assisted pattern recognition, moreover, can help predict the risk of lead poisoning among children, depending on where they live. One transnational example illustrates some of the opportunities ADM offers: an artificial neuronal network that uses satellite photos to determine the regional distribution of poverty in developing countries almost as accurately as considerably more expensive on-site surveys. The results could be used to combat poverty by targeting those areas in the most distress and where, consequently, assistance can have the greatest impact. An example from France (university admissions) and several processes used in the US and subsequently adopted by Germany (e.g. location-specific predictive policing) show that the use of ADM processes is a global phenomenon, one that is also becoming more prevalent in Germany.

Every example highlights a typical problem and, with it, the need for a corrective response when designing future ADM processes meant to increase participation. To the greatest extent possible, the discussion here treats the problems as if they were discrete phenomena, knowing full well that the identified shortcomings often occur concurrently in practice.

To take advantage of the opportunities ADM offers in the area of participation, one overall goal must be set when ADM processes are planned, designed and implemented: ensuring that participation actually increases. If this is not the case, the use of these tools could in fact lead to greater social inequality. The risks and unwanted consequences seen in the chosen examples illustrate where corrective responses are required. Numerous potential problems can often be observed in the individual application scenarios. In each example given in this working paper, we highlight a typical response that should be considered when ADM processes designed to increase participation are developed.

Table 3: Corrective responses to ADM processes

Response	Description	Example
Ensure falsifiability	ADM processes can learn asymmetrically from mistakes. "Asymmetric" means that the system, by virtue of the design of the overall process, can only recognize in retrospect certain types of its own predictions which proved incorrect. When algorithms learn asymmetrically, the danger always exists that self-reinforcing feedback loops will occur.	Recidivism predictions used in the legal system

Ensure proper use	Institutional logic can lead to ADM processes being used for completely different purposes than originally envisioned by their developers. Such inappropriate uses must be avoided.	Predicting individual criminal behavior
Identify appropriate logic model for social impact	Algorithm-driven efficiency gains in individual process steps can obscure the question of whether the means used to solve a social problem are generally appropriate.	Predicting lead poisoning
Make concepts properly measurable	Social phenomena or issues such as poverty and social inequality are often hard to operationalize. Robust benchmarks developed through public discussion are therefore helpful.	Predicting patterns of poverty
Ensure comprehensive evaluation	The normative power of what is technically feasible all too easily eclipses the discussion of what makes sense from a social point of view. For example, the scalability of machine-based decisions can quickly lead to situations in which the appropriateness and consequences for society of using ADM processes have neither been debated nor verified.	Automatic face-recognition systems
Ensure diversity of ADM processes	Once developed, the decision-making logic behind an ADM process can be applied in a great number of instances without any substantial increase in cost. One result is that a limited number of ADM processes can predominate in certain areas of application. The more extensive the reach, the more difficult it is for individuals to escape the process or its consequences.	Preselection of candidates using online personality tests
Facilitate verifiability	Frequently, no effort is made to determine if an ADM process is sufficiently fair. Doing so is even impossible if the logic and nature of an algorithm is kept secret. Without verification by independent third parties, no informed debate on the opportunities and risks of a specific ADM process can take place.	University admissions in France
Consider social interdependencies	Even when use is very limited, the interdependences between ADM processes and their environment are highly complex. Only an analysis of the entire socio-informatic process can reveal the relationship between opportunities and risks.	Location-specific predictions of criminal behavior
Prevent misuse	Easily accessible predictions such as scoring results can be used for inappropriate purposes. Such misuse must be prevented at all costs.	Credit scoring in the US

This publication documents the preliminary results of our investigation of the topic. We are publishing it as a working paper to contribute to this rapidly developing field in a way that others can build on. We are therefore making it available as a working paper using a free license (CC BY-SA 3.0 DE), so that it might serve as the basis for discussion in workshops or during other considerations of the topic.

The case studies show how ADM processes influence decisions made about people. When machines evaluate us and when their predictions – as used in the legal system or by law enforcement officials – affect personal rights or – as is the case when candidates are being selected or credit assessed – issues of equality, then society must discuss the fairness of these processes and their impact on participation.

This is where a close look must be taken at the specific context, since not all ADM processes are equally risky. What society demands of ADM processes can vary depending on the consequences these processes have for

society as a whole or for individuals and their basic rights. Spelling-correction programs and navigation systems have a different impact on a person's life than processes which flag a person as being a credit risk or likely to commit a crime.

In sum, the opportunities and risks in the examples presented here point to a number of general factors related to ADM processes that can critically affect participation. These factors involve different aspects of the overall socio-informatic process and can be found on different levels. Here are three examples:

- **Shaping ADM processes on the micro and macro level:** Choosing data and setting criteria at the start of a development process can themselves reflect normative principles which sometimes touch on fundamental social issues.

- **Structure of suppliers and operators on the macro level**: Having a range of ADM processes and operators can increase participation (e.g. through credit assessments of people who have not been part of the system in the past), can make it easier to avoid the ADM process and can expand possibilities for falsification. Conversely, monopolistic structures increase the risk that individuals will "fall out of the system" and get left behind.

- **Use of ADM forecasts on the micro, meso and macro level**: The interplay of technology, society and individuals has a major impact on how and when algorithms are used and the influence they thus have. Key questions that must therefore be asked are: How do people (ADM developers and users, and the general public) deal with automated predictions? Do the processes include the possibility of challenging ADM results?

What are needed here are additional systematic analyses of the potential shortcomings of ADM processes on different levels – from the definition of the goals and the efforts to measure the issues at hand, to data collection, the selecting of algorithms and the embedding of processes in the relevant social context. Criteria are needed for determining the benefits of ADM processes on all levels and in all steps. The responses discussed here can provide initial impetus for addressing these issues (see Table 1: Need for action in algorithmic decision-making processes).

Table 4: Summary of opportunities and risks found in case studies

Dimension	Opportunities	Risks
Normative principles	When an ADM process is designed, normative decisions (e.g. about fairness criteria) must be made before the process is used. This offers an opportunity to discuss ethics issues thoroughly and publically at the very start and to document decisions.	ADM processes can contain hidden normative decisions. If discussion is only possible once the design phase is complete, any normative principles are more likely to be accepted as unalterable.
Data	Software can analyze a much greater volume of data than humans can, thereby identifying patterns and answering certain questions faster, more precisely and less expensively.	The data used for an ADM process can contain distortions that are seemingly objectified by the process itself. If the causalities behind the correlations are not verified, there is a significant danger that unintentional, systematic discrimination will become an accepted part of the process.

Consistency of application	Algorithm-based predictions apply the predetermined decision-making logic to each individual case. In contrast to human decision makers, software does not have good and bad days and does not in some cases arbitrarily use new, sometimes inappropriate criteria.	In exceptional cases, there is usually no possibility for assessing unexpected relevant events and reacting accordingly. ADM systems unfailingly make use of any incorrect training data and faulty decision-making logic.
Scalability	Software can be applied to an area of application that is potentially many times larger than what a human decision maker can respond to, since the decision-making logic used in a system can be applied at very low cost to a virtually limitless number of cases.	ADM processes are easily scalable, which can lead to a decrease in the range of such processes that are or can be used, and to machine-based decisions being made much more often and in many more instances that might be desirable from a societal point of view.
Verifiability	Data-driven and digital systems can be structured in a way that makes them clear and comprehensible, allows them to be explained and independently verified, and provides the possibility of forensic data analysis.	Because of process design and operational application, independent evaluations and explanations of decisions are often only possible, comprehensible or institutionalized to a limited degree.
Adaptability	ADM processes can be adapted to new conditions by using either new training data or self-learning systems.	The symmetry of the adaptability in all directions depends on how the process is designed. One-sided adaptation is also possible.
Efficiency	Having machines evaluate large amounts of data is usually cheaper than having human analysts evaluate the same amount.	Efficiency gains achieved through ADM processes can hide the fact that the absolute level of available resources is too low or inadequate.
Personalization	ADM processes can democratize access to personalized products and services that for cost-related reasons were previously only available to a limited number of people. For example, before the Internet, numerous research assistants and librarians were required to provide the breadth and depth of information that results from a single search-engine query.	When ADM processes are the main tools used for the mass market, only a privileged few have the opportunity to be evaluated by human decision makers, something that can be advantageous in non-standard situations when candidates are being preselected or credit scores awarded.
Human perception of machine-based decisions	ADM processes can be very consistent in making statistical predictions. In some cases, such predictions are more reliable than those made by human experts. This	People can view software-generated predictions as more reliable, objective and meaningful than other information. In some cases this can prevent people from questioning recommendations and predictions or

	means software can serve as a supplementary tool which frees up time for more important activities.	can result in their reacting to them only in the recommended manner.

An important – even definitive – quality factor must be stressed once again: An analysis of the opportunities, risks and societal consequences was only possible because independent third parties were able to verify the benefits of machine-based decisions. Institutions such as the investigative newsroom ProPublica, the US Government Accountability Office and the student-rights organization Droits des lycéens spent the time and financial resources needed to collect and evaluate data and to consider the relevant legal issues, allowing each of the algorithms to be explained and made transparent. Public debate on the impact of certain ADM processes thus depends completely on institutions of this sort – a situation that must change. It must be possible to verify and understand algorithmic decisions if an effective discussion is to take place, one which ensures that ADM processes actually increase participation and that machine-based decisions truly benefit people.

ADM processes will only contribute to the common good if they are discussed, criticized and corrected. We are still in a position to determine how we as a society want to make use of algorithms. We should not only consider how they are applied, but, in some cases, whether they should be used at all. For example, in those situations where society has chosen to promote solidarity and share risks, ADM processes cannot be permitted to individualize those risks. The guiding principle cannot be what is technically feasible, but what makes sense from a societal perspective – so that machine-based decisions truly do benefit people.

Address | Contact

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Telephone +49 5241 81-81216

Konrad Lischka
Taskforce Digitalisierung
Telephone +49 5241 81-81216
konrad.lischka@bertelsmann-stiftung.de

www.bertelsmann-stiftung.de

BertelsmannStiftung