Horizon 2020

Technical report data sources and data protection for algorithm development

Deliverable number: D2.2

Hecat

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870702

Project Acronym: HECAT
Project Full Title: Disruptive Technologies Supporting Labour Market Decision Making
Call: H2020-SC6-TRANSFORMATIONS-2018-2019-2020
Grant Number: 870702
Project URL: https://hecat.eu

Deliverable nature:	Report (R)
Dissemination level:	Public
Contractual Delivery Date:	M14
Actual Delivery Date	M14
Number of pages:	41
Authors[1]:	Ayo Næsborg-Andersen, Snorre Frid-Nielsen, Janine Leschke, Clément Brébion
Peer review:	Ray Griffin; Pavle Boskoski

Abstract

This report presents the legal considerations that one should take when using data and algorithms in public employment services. It largely draws on examples from statistical profiling as this is currently one of the most common algorithmic tools used in public employment services (PES). Importantly, the general lessons are applicable beyond profiling for example for the purpose of data-driven decision support systems in PES. The report discusses the legal basis for algorithmic decision making and among other discusses proportionality, discrimination, fairness and protection of sensitive data. It highlights the necessity to conduct a thorough assessment of issues such as data quality, accuracy and transparency during the whole life cycle of an algorithmic tool. The report also intends to provide lessons for the HECAT algorithmic pilot tool.

Contents

[1] Ayo Næsborg-Andersen was the main author of sections 1-3, Snorre Frid-Nielsen was the main author of section 4, Janine Leschke conceptualized the paper, wrote the conclusions and revised and fine-tuned drafts of all sections at several stages, Clément Brébion wrote parts of section 4 and 5 and revised and fine-tuned drafts at several stages.

1. Introduction: Why is data protection relevant?

The right to privacy has been debated since at least Warren and Brandeis' 1890 article of the same name extolling the dangers of the newfangled invention of the camera (Warren and Brandeis, 1890). While Warren and Brandeis advocated for limitations to be set on when and where a camera could be used without the consent of the portrayed, today we have established regulation and case law for exactly that purpose. While the taking of one's picture without consent can definitely still be considered an invasion of privacy, the modern-day threat of invasion of privacy holds a new angle which lawmakers are still grappling with, namely the copious amount of data available about individuals, be it from the internet and social media, or from the various registers developed and maintained by public authorities. As technology has developed, drastically expanding the possibilities of gathering information about individuals, so too has the protection afforded the same individuals, albeit not necessarily in the same tempo. The protection of the private life of citizens has been a foundational element of the various human rights treaties written since at least the United Nations' Universal Declaration of Human Rights was proclaimed in 1948. In Europe, article 8 of the European Convention on Human Rights (1952) sets out the right to a private life, and the EU Charter of Fundamental Rights (2000) repeats that same right in article 7 while adding a specific right to data protection in article 8. The General Data Protection Regulation (GDPR) aims to both protect the private life of citizens while also enabling the processing of personal data, a balancing act which necessitates careful weighing of the different aspects relevant to a particular process.

Combining data sources to glean new information about individuals can be considered an interference in their privacy. Such an interference is not necessarily illegal per se but needs to be carefully considered before being put into action, to justify its use. This consideration is necessitated both by the GDPR, and by the case law of the European Court of Human Rights as well as the Court of Justice of the European Union, whereby a measure – such as the introduction of a potentially privacy-violating profiling – needs to be shown to be both necessary and proportional in order to be in accordance with human rights law. If these considerations are not met, then the project runs a severe risk of being later found to be illegal, with, at best, changes required to the algorithm and/or the implementation, and, at worst, the entire algorithm being shut down because it is impossible to make it legal. See e.g. the Austrian algorithm for profiling jobseekers which was shut down before it was put into use (Szigetvari, 2020), or the Danish EFI-system ("Et Fælles Inddrivningssystem," roughly translated as One Common System of Collection) for collecting arrears owed to the state which was closed shortly after being put into use (Motzfeldt and Næsborg-Andersen, 2018).

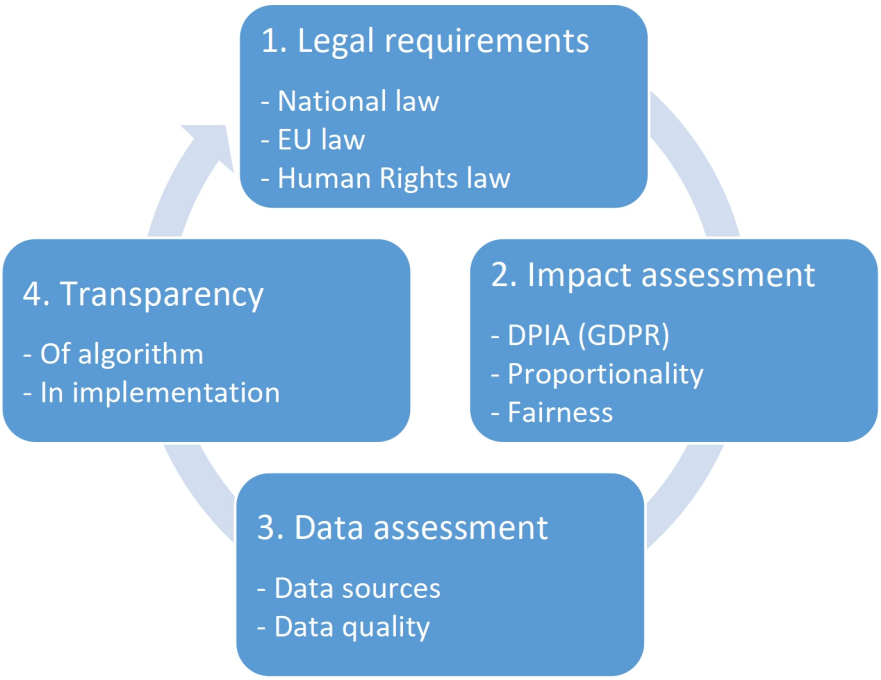
The introduction of algorithms in the administration is often justified by claims of the algorithm being "fair" and "objective", or at least *more* fair and objective than the individual case-worker. While the individual case-worker may very well be biased, and decisions made will invariably vary from one case to another as long as some degree of discretion is allowed, these are not unknown or indeed un-tackled issues. In fact, general principles of administrative law have been developed to counter exactly these problems, ensuring as fair and unbiased decisions as possible, while providing citizens with the opportunity to seek redress where the safeguards fail. Such principles have yet to be developed to counter the specific issues, such as "black box" decisions, or algorithmic bias, raised by the introduction of algorithms. Administrative law in general has developed over centuries, while the introduction of algorithms is a very recent development. Therefore, there is a particular need to be thorough when developing and implementing algorithms in the (public) administration, as many different aspects need to be considered, and no authoritative "roadmap" has been developed yet on which aspects to consider, or when. The following three sections of this paper collect and interpret principles from the GDPR, human rights law and general administrative law, to sketch out the

beginnings of such a roadmap. Where possible we relate these principles to data use and data practices in public employment services particularly with a view to statistical profiling. The focus on profiling in this report is due to the fact that this is one of the most common algorithmic tools currently used in public employment services. However, importantly, the general lessons are applicable beyond profiling for example for the purpose of data-driven decision support systems in PES. Section five aims at explicitly illustrating the potential pitfalls in view of our pilot case.

2. Proposal for a model for ensuring legality

First, it is important to note, that the GDPR holds no checklists for which data can be used, nor which algorithms can be implemented. Whether or not the use of an algorithm is legal must instead be continually assessed before, during and after development and implementation. The most important rule-of-thumb here is that as long as you have thoroughly considered the various aspects of both the data used, and of the impact the algorithm will have on the citizens in question, and can make a good argument for the justification of said impact, *and* you can point to a legal basis for both the use of the data and the use of the algorithm, then it is most likely to be legal. These considerations need to be *documented*, as it is the responsibility of the implementing authority (data controller in terms of the GDPR) to be able to demonstrate compliance with the GDPR (art 5(2)). In this process you also need to consider whether the use of the data and/or the algorithm will result in discrimination, and how to counteract this (see below sections 3.3.b, 4.6, and 5). You also need to ensure, in designing the algorithm, and especially in the implementation of it, that national administrative law is upheld. Therefore, the first step of development of an algorithm intended to be used on citizens should be an examination of the requirements and constraints of relevant laws, followed by an impact assessment. A proposed model for relevant considerations when designing an algorithm could look like this:

Figure 1 A proposed model for designing algorithms



Source: own depiction.

The model is circular to illustrate that the considerations are ongoing and should be continually updated to reflect the outcome of both the algorithm and the different steps of the considerations. Most of the steps should also happen more or less concurrently.

1. Legal requirements

Before designing an algorithm, it is important to map out the various requirements to both form and content of the algorithm, as well as requirements regarding the process of development and implementation. These are most frequently found in national law (particularly regarding administrative law), EU law (primarily the GDPR, but this could also include various other sources) and human rights law (the European Convention on Human rights and the Charter for Fundamental Rights). Some can also be found in the general principles of administrative law. The legal requirements can be seen as defining the next steps in the model. This paper describes many of the legal requirements that such a mapping should uncover, but does not make any claims to complete coverage. National law in particular has not been examined, and there may be other areas of EU law which are relevant, but not discussed here.

2. Impact assessment

The GDPR, in the form of the Data Protection Impact Assessment (DPIA), human rights law and general principles of administrative law all require a form of impact assessment, ideally before the development of an algorithm starts, or at the very least before it is implemented. The administration cannot make decisions that are not both proportional and fair, and therefore the impact of the decisions on the citizens in question needs to be carefully considered. See section 2.1, as well as the discussions on proportionality (section 4.2) and fairness (section 4.6).

3. Data assessment

Both the data sources, and the data quality need to be carefully considered, in order to ensure that the data is of sufficient quality, and that only the relevant data is used (see section 4.3).

4. Transparency

Any decision made by a public administration needs to be transparent to the citizen in question. The degree of transparency required is still not quite settled, but in essence, the citizen needs to be able to understand both why a decision was made, and what the basis for the decision was. Some transparency as to how the algorithm is implemented by the administration is also required (see section 4.5).

2.1. Impact assessment

To determine what data an algorithm can use, and how the algorithm can be implemented by the administration, the all-important first step is to make an impact assessment. This requirement follows directly from GDPR article 25, whereby data protection, including considerations for the citizens affected by the algorithm, must be built into the system from the start. This should be carried out ideally prior to development of the algorithm, and then updated concurrently with any changes. If the algorithm is developed as part of a research project, then at the very least the assessment must be made before the algorithm is implemented in administrative practice.

The assessment should focus on two different aspects: First of all ensuring the requirements of the GDPR, namely both the safety and proportionality of the data, as well as the necessity of using the

algorithm, including the impact of the algorithm on the citizens in question, particularly as regards human rights (see Data Protection Impact Assessment (DPIA), GDPR art 35), and second, examining the administrative impact of the algorithm on the citizen whom the decision concerns (Motzfeldt and Næsborg-Andersen, 2018, and the section on administrative law principles below). These aspects are somewhat overlapping, but it is important to keep track of both sides of the assessment, i.e. the data and the citizens.

To be able to perform an impact assessment, it is necessary to determine first how the algorithm will be used, as this determines the level of impact on the citizen. If, as in the Belgian case, the purpose is to prioritize which jobseeker gets a phone call from the unemployment agency first (Desiere and Struyven, 2020), the impact is likely to be very low. After all, all jobseekers will get the same call, it is only a matter of timing. If, on the other hand, the purpose of the algorithm is to sort jobseekers into different categories with correspondingly vastly different types of services and/or differentiated access to active labour market measures (Allhutter et al., 2020), like in the Austrian case, the impact is likely to be moderate or high, depending on the various safeguards built into the system, and what, exactly, the consequences are. The levels of impact can be illustrated by tables such as the one below[2]:

Table 1 Impact assessment taken from the Canadian Directive on Automated Decision-Making

Level	Description
I	The decision will likely have little to no impact on: 1. the rights of individuals or communities, 2. the health or well-being of individuals or communities, 3. the economic interests of individuals, entities, or communities, 4. the ongoing sustainability of an ecosystem. Level I decisions will often lead to impacts that are reversible and brief.
II	The decision will likely have moderate impacts on: 1. the rights of individuals or communities, 2. the health or well-being of individuals or communities, 3. the economic interests of individuals, entities, or communities, 4. the ongoing sustainability of an ecosystem. Level II decisions will often lead to impacts that are likely reversible and short-term.
III	The decision will likely have high impacts on: 1. the rights of individuals or communities, 2. the health or well-being of individuals or communities, 3. the economic interests of individuals, entities, or communities, 4. the ongoing sustainability of an ecosystem. Level III decisions will often lead to impacts that can be difficult to reverse, and are ongoing.
IV	The decision will likely have very high impacts on: 1. the rights of individuals or communities, 2. the health or well-being of individuals or communities,

[2] Taken from the Canadian Directive on Automated Decision-making, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 (accessed Oct 29 2020). While this is not EU-law, it is very much in line with the requirements of the GDPR.

3. the economic interests of individuals, entities, or communities,
4. the ongoing sustainability of an ecosystem.

Level IV decisions will often lead to impacts that are irreversible, and are perpetual.

Source: Canadian Directive on Automated Decision-making: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

If the impact of the algorithm is determined to be level one, i.e. little to no impact, then the requirements for administrative safeguards, and restrictions on the data available, are very low. If, on the other hand, the impact of the algorithm is determined to be level four, i.e. very high, the requirements and restrictions will be proportionally higher. See the relevant legal texts in Appendix A2.

3. The legal basis

3.1. Choice of legal basis

According to article 5.1.a of the GDPR (see Appendix A2), processing of personal data needs to be lawful, i.e. have a legal basis. The various legal bases can be found in article 6 (and article 9 if processing sensitive data).

The relevant legal basis for profiling in the context of unemployment would be either article 6.1.c, whereby processing is necessary for compliance with a legal obligation to which the controller (in this case the public authority) is subject, such as the obligation to provide a legally mandated service to the jobseeker, or 6.1.e, where processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority, such as the tasks necessary to run a PES. Which of the two is relevant will depend on the national law of the country where the profiling is implemented.

One common misunderstanding is the idea that consent can be used as a legal basis in the context of unemployment services, as consent is one of the legal bases covered in article 6. The definition of consent varies from law to law. This makes consent very difficult to work with, as consent within e.g. health law or social law is not the same as consent according to the GDPR. In terms of the GDPR, consent is defined in art. 4(11): *"'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her".* In the context of unemployment profiling, the problem lies in "freely given". This means that the registered person must be in a position where the choice is actually free, and a lack of consent will not have negative consequences. In practice, in most institutional settings jobseekers are systematically profiled and consent is not ensured. But even in cases where profiling is not mandatory for the jobseeker (Denmark, the Netherlands, and to some degree Latvia, see here) authorities will almost always hold power over citizens who would not provide their personal information voluntarily. This creates an unequal relationship, meaning the choice is not actually free. An authority can use consent for inconsequential things such as subscription to a newsletter, but not for things that will almost inevitably impact a case, such as with unemployment profiling, where the outcome of the profile will inevitably influence the case-worker, as will the lack of a profile due to the citizen not giving consent. This situation, therefore, cannot be based on consent according to the

GDPR. See also preamble no. 43 of the GDPR and points 16-20 of Guidelines 05/2020 on consent under Regulation 2016/679 (source).

Whether or not the consent is implicit is irrelevant in this context, as it still must live up to the above-mentioned criteria of being "freely given". Consent must also be unambiguous, and therefore implicit consent is not likely to be considered legal. A pre-ticked check-box, e.g., is not considered unambiguous consent, as the data controller cannot demonstrate that the registered person actively consented (see recital 32 of the GDPR and case C-61/19, pr. 37 from the CJEU).

The profiling could be done on an "opt-in" basis, but that would not be considered consent in terms of the GDPR, and the profiling would therefore still have to have a legal basis in law and should therefore be considered to fall under the abovementioned article 6.1.c or 6.1.e. The choice of using profiling as "opt-in" would probably not significantly change the assessment of the impact of the profiling, as it is not, all things considered, much of a free choice.

Consent would potentially be a relevant legal basis if the resulting profile was made available exclusively to the citizen as a tool for self-assessment, and not to the caseworker and/or the authorities. This would to a large extent depend on the way it was presented to the citizen.

3.1.1. Principles of data protection

Article 5 of the GDPR describes the principles which all processing of personal data must uphold (see article 5 in Appendix A2).

Lawfulness refers to the fact that the processing must be in accordance with national law, as well as the GDPR. Therefore, there is a need to ensure the processing has basis in national law (see below 3.2).

Fairness is both a question of the processing being non-discriminatory, but also not going beyond the reasonable expectations of the persons whose data is being processed. As such, the principle of fairness also necessitates proportionality (see below 4.2 and 4.6).

Transparency is a matter of the registered persons being able to understand and follow what is happening to their personal data (see below 4.5).

Purpose limitation entails not using data for purposes which they were not originally intended to be used for. In the context of profiling this usually entails ensuring there is a legal basis for using the data specifically for profiling.

Data minimalization means you can only use the data which is actually necessary for fulfilling the purpose of the processing. Conversely, it also means you must use the data you deem necessary (provided it does not violate other principles such as fairness). "Need to have, not nice to have" is a popular way of phrasing this principle (see below in particular on proportionality, 4.1).

Accuracy is also sometimes referred to as data quality. This is the demand both that you know how accurate your data is, and that you do not rely on data of questionable quality (see below 4.3).

Storage limitation means not keeping data for longer than necessary. This is usually a question of setting up schedules for automatic or manual deletion.

Integrity and confidentiality as a principle demands of the controller that data is sufficiently protected, both from security breaches or accidental damage.

And finally, but perhaps most importantly, the principle of *accountability* demands that the controller is able to demonstrate compliance with all these principles. For this reason, all considerations on e.g. what data to use, how to use it, and when to delete it, must be *documented*.

3.1.2. Data protection by design and default

An important feature of the GDPR, especially relevant when developing and implementing algorithms, is the demand in article 25 that the requirements of the GDPR must be built into the system from the start (data protection by design).

Additionally, the system needs to be set by default to e.g. only gather the necessary data, so as to uphold the principles of the GDPR. It is, in other words, not enough to design the system to uphold the GDPR, the various settings need to be enabled, by default, to support this (data protection by default).

Read more in the European Data Protection Board's Guideline 4/2019 on Data protection by design and default here.

3.2. National law

One of the main principles of data processing is that all processing must be legal (GDPR art. 5.1.a.). This means that processing must happen in accordance with both the GDPR and national law. According to the GDPR art 6 and 9 you need a legal basis for using personal data. Consent is not a relevant legal basis here, as consent must be given freely in order to fulfil the GDPR's definition of consent (art. 4.11), and in cases where the giving or withholding of consent will impact decisions made regarding a citizen, consent is not considered to be given freely. Instead, the relevant basis would most likely be art. 6.1.e, i.e. the carrying out of a task in the public interest, or in the exercise of official authority. As a consequence hereof, the decisions made by the administration on the basis of processing need to have a legal basis in national law. The requirement of the amount of detail and precision of the national law will likely vary according to the impact the processing will have on the citizen. It is also necessary to consider what the relevant national administrative law procedures might be, such as the right to an explanation, or the right to be heard before a decision is made. The various requirements made by national law should be mapped out and communicated clearly to the stakeholders in the development and implementation of the algorithm. This is connected very closely to the general principles of administrative law, which will be described further in the next section. Motzfeldt and Naesborg-Andersen (2018) and Kuner (2020: 335–6) provide more details on this matter.

If the processing results in fully automated decision making, the GDPR requires such use to be regulated by national law, according to article 22.2.b. This law must include a careful consideration of the fundamental rights of the registered persons, and safeguards to protect them

3.3. General principles of administrative law

Unlike both human rights law and the GDPR, administrative law is mostly determined nationally. There is, however, some common ground amongst at least the European countries, dating back at least to two Council of Europe acts from 1977 and 1980, respectively. The first formulated the

following five principles on the protection of the individual in relation to the acts of administrative authorities (Council of Europe, 1977):

I
Right to be heard
1. In respect of any administrative act of such nature as is likely to affect adversely his rights, liberties or interests, the person concerned may put forward facts and arguments and, in appropriate cases, call evidence which will be taken into account by the administrative authority.
2. In appropriate cases the person concerned is informed, in due time and in a manner appropriate to the case, of the rights stated in the preceding paragraph.

II
Access to information
At his request, the person concerned is informed, before an administrative act is taken, by appropriate means, of all available factors relevant to the taking of that act.

III
Assistance and representation
The person concerned may be assisted or represented in the administrative procedure.

IV
Statement of reasons
Where an administrative act is of such nature as adversely to affect his rights, liberties or interests, the person concerned is informed of the reasons on which it is based. This is done either by stating the reasons in the act, or by communicating them, at his request, to the person concerned in writing within a reasonable time.

V
Indication of remedies
Where an administrative act which is given in written form adversely affects the rights, liberties or interests of the person concerned, it indicates the normal remedies against it, as well as the time-limits for their utilisation.

In relation to algorithm-supported decisions, all principles still apply, as long as the decisions have an impact on individuals. Of particular importance is principle IV on the statement of reasons, as such a statement requires some degree of transparency as regards the algorithm. If the decision is made by a so-called "black box" (Pasquale, 2015), then it is all but impossible to provide a relevant and correct statement of the reasons underlying the decision.

Also highly relevant is principle I on the right to be heard. In order for this principle to be upheld, the citizen needs to be consulted before a decision is made. The purpose of such a consultation is both to ensure that the administration bases its decisions on correct information, and that relevant arguments and information are taken into account. This principle serves, among others, to ensure that special consideration can be taken of cases which are not typical. In other words, this principle helps ensure that the administration both treats equal cases alike, and unequal cases differently, thereby preventing discrimination.

Whereas the above principles were formulated for all administrative acts, additional principles on the exercise of discretionary power by the administration can be found in the 1980 resolution (Council of Europe, 1980). Discretionary power is here defined as "a power which leaves an administrative authority some degree of latitude as regards the decision to be taken, enabling it to choose from among several legally admissible decisions the one which it finds to be most appropriate". The use of an algorithm to e.g. decide on a jobseeker's path through the unemployment system would be considered a use of discretionary power, unless both the algorithm and the possible outcomes are strictly and exhaustively defined by law. The principles are as follows:

An administrative authority, when exercising a discretionary power:
1. does not pursue a purpose other than that for which the power has been conferred;
2. observes objectivity and impartiality, taking into account only the factors relevant to the particular case;
3. observes the principle of equality before the law by avoiding unfair discrimination;

4. maintains a proper balance between any adverse effects which its decision may have on the rights, liberties or interests of persons and the purpose which it pursues;
5. takes its decision within a time which is reasonable having regard to the matter at stake;
6. applies any general administrative guidelines in a consistent manner while at the same time taking account of the particular circumstances of each case.

These principles emphasize the need for transparency, fairness/non-discrimination and proportionality to be a pivotal part of the development of an algorithm. As regards transparency, it is necessary to demonstrate that only factors relevant to a particular case were taken into account (principle 2). Fairness is reflected in both principles 3 and 4, which emphasize both the need to avoid unfair discrimination, and to carefully consider the proportionality and fairness of using the algorithm (see above).

Principles 5 and 6 are mainly concerned with the procedures adopted by the administration being timely and upholding general administrative guidelines, and as such warrants a close follow-up once implementation of any algorithm has begun.

Finally, the first principle underlines the necessity of any algorithm developed and implemented to have basis in law. The algorithm must thus pursue the same objective as the law which legitimizes it.

3.4. Human rights law

Human rights in the EU are regulated by two main treaties: The European Convention on Human Rights (ECHR), and the Charter for Fundamental Rights (CFR). The ECHR is binding on member states of the Council of Europe, and the CFR is binding on member states of the EU. As all members of the EU are also members of the Council of Europe, both are binding on members of the EU. In addition, the EU has adopted the ECHR, and the jurisprudence of the European Court of Human Rights (ECtHR), as the minimum standard for the protection of human rights in the EU (TEU art 6 and CFR art 52.3). Therefore, all implementation of EU law must uphold both the CFR and the ECHR, and all national law must, at the very least, uphold the ECHR, if not also the CFR, depending on the degree to which the law in question is an implementation of EU law. In the following, the most relevant principles of human rights law in relation to the implementation of decision-supporting or decision-making algorithms are discussed.

3.4.1. Proportionality

The ECHR protects the right to the protection of private life (among others) in article 8. This is interpreted in ECtHR case law as also protecting personal data (see *inter alia* the case of Satakunnan Markkinapörssi Oy and Satamedia Oy v. Finland). The CFR also protects private life in art 7, with a special provision on the protection of personal data in art 8. The use of personal data, by someone other than the person in question, is considered an interference in the right to privacy, including especially if such use is made by a public authority in order to determine the allocation of funds and resources (Alston, 2019a). Such an interference can be justified if: a) there is a legal basis in law, b) the interference can be considered necessary in a democratic society, and c) the interference is proportional to the goal. The stronger the interference, the stronger the requirement to prove all three conditions are met (Gerards, 2019).

3.4.2. Discrimination

According to human rights law (ECHR art 14 and CFR chapter III) discrimination means treating somebody:

a) in comparable situations
b) differently
c) based on certain characteristics
d) without legitimate aim or proportionality

So, there are four criteria to assess when considering whether something is discriminating. First of all, *comparable situations*. In this case, that would be people who are unemployed. *The certain characteristics* could be all kinds of things but are typically the classical causes for discrimination; race, gender, sexuality, etc. – all personal characteristics, which cannot be changed (such as race), or which it is not reasonable to ask people to change (such as religion).

The list of characteristics protected by human rights law has a high degree of overlap with article 9 of the GDPR, but as article 14 of the ECHR is not exhaustive, human rights law is more inclusive. It may be more accurate to say that article 9 of the GDPR reflects some, but by no means all of the characteristics protected by human rights law.

If these characteristics are used as grounds for decisions without good reason, it is considered to be discriminatory (see also positive discrimination next paragraph). It is important here to consider whether things are actually comparable: The reason it is not discriminatory, e.g., to reserve the first part of the maternity leave for the mother is that her situation is not directly comparable to the father, as she will presumably be nursing the child. It is, however, discriminatory to exclude fathers from, say, parental leave later in the child's life, as there is no biological reason for why fathers cannot take equally good care of their children as the mothers, at that stage. So, the state can treat people *differently*, if the actual situation requires it, as that would be a *legitimate aim* – indeed it must do so, if the situations are significantly different. In the context of jobseekers this would mean that you can make a distinction between e.g. levels of education, as that is directly relevant to the types of job they can apply for. Likewise, an argument could be made for making a distinction on the grounds of previous work or unemployment experience (in months), as someone with many years of experience will present a different value for an employer than someone with no experience. It is, however, important to consider whether the reason you are distinguishing two persons or groups from each other is due to prejudice, as prejudice is never an acceptable reason for discrimination. Employers being reluctant to employ a certain category of people due to prejudice does not qualify, as a public authority must work against, instead of amplifying, any prejudice. In this case discrimination would be grounded in a personal characteristic, which has no influence on whether or not the person can fulfil the job and is therefore in violation of human rights law. Consequently, it is also a direct violation of the requirements of profiling in the GDPR, as noted above. It is, however, important to note that prejudice can be taken into account when it comes to so-called positive discrimination.

Another legitimate reason for treating people differently is, in fact, *positive discrimination*. You are allowed to treat people differently based on certain characteristics if you are trying to correct factual inequalities (sometimes you are even required to do so). You must, however, a) be very certain that there is an inequality (i.e. needs to be documented), and b) be fairly certain that the differential treatment will actually help correct the inequality (so follow-ups and course-corrections to the algorithm/implementation and use of the algorithm would be expected). In this context, it matters

greatly what the differential treatment consists of. If, as was intended in the Austrian case, falling in the category with highest employment barriers leads to a "moral training" course, then it is highly doubtful that such a course actually corrects anything. This would therefore be considered discriminatory and in violation of human rights law. There is also the question of whether the differential treatment attaches a stigma to the groups you are trying to help, which singling people out based on characteristics such as race invariably does, to some degree. In other words, positive discrimination needs to be considered very carefully, before it is put into action. Preferably, the national parliament needs to be making this decision, as this would give it the most democratic legitimacy possible, although this is not a guarantee that it will not be found discriminatory at a later time. If the main problem is prejudice on the employer side, then singling out the potential employees which the employers discriminate against is probably not going to help, unless you also (predominantly) actively work on challenging and changing the underlying prejudice. You need to attack the root cause of the discrimination, not the people being discriminated against, for positive discrimination to be considered positive (Desiere and Struyven, 2020).

In cases where statistics are used for decisions (or support for decisions), the question of *indirect discrimination* invariably poses an issue. If a measure can be shown to proportionally affect certain groups more than others, it is considered to be indirectly discriminating, regardless of whether that was the intent. This is especially the case if the affected group is one that is traditionally considered a minority, on the basis of, e.g., ethnic origin, gender or sexuality. Therefore, if some proxies correlate strongly with protected characteristics, it must be considered very thoroughly if they need to be included, and if so, what measures can be taken to counteract the correlation. Some data may need to be excluded on this account, even if this makes the algorithm less precise. One consideration here is whether the data reflects the information which one seeks in a precise enough way. If, say, country of residence is included, this poses a high risk of being discriminatory data. If the reason for including the information about country of residence is that it reflects whether the jobseeker has a temporary work permit or not, then that information should be the input instead. The information about the status of a work permit may still indirectly discriminate against jobseekers of other nationalities, but the justification for including it is much higher, as a) it is more accurate, b) it runs less of a risk of discriminating against specific nationalities, and c) it is directly relevant to the jobseeker's situation. The country of residence may also serve as a proxy for language skills, in which case information about the country of residence conflates two different issues, resulting in less precision, and less actionable information. Language skills can be worked on, but only if the jobseeker and/or the caseworker are aware of the potential problem. Therefore, using the input of country of residence instead of language skills and/or status of work permit, will disguise the relevant issues, leaving the jobseeker and the caseworker ignorant of potential areas for improvement. To completely ensure that the inclusion of a parameter such as country of residence is not discriminatory a final assessment of the relevance and necessity of including the information would need to be weighed against the measures imposed, to ensure proportionality.

Proportionality of the measure imposed is also a necessary consideration in assessing discrimination. The stronger the impact on the persons affected, the stronger will be the need to demonstrate the necessity of sorting people according to personal characteristics which have no impact on their ability to fulfill a job such as ethnic origin. A measure which has next-to-no impact, such as the ranking of which jobseeker will get a phone call from the unemployment agency first within a short timeframe, does not need strong justifications. In contrast, a measure such as sorting the jobseekers into different categories that determine both whether they get specific assistance from the public authority and which type, needs stronger justification.

Read more on discrimination in the ECtHR case-law guide on article 14 of the ECHR.

3.4.3. The right to an effective remedy

Article 13 of the ECHR provides the right to an effective remedy. Therefore, the citizen has the right to challenge an administrative decision. There is not currently any case law from either the ECtHR or the ECJ on what, exactly, the right to an effective remedy entails as regards assessments made by algorithms. Based on previous case law on effective remedies, however, some guidelines can be proposed.

A conclusive definition of an effective remedy does not exist. The process of attaining the remedy can take various forms, such as appealing a decision directly to the administration, or taking a case to the courts. The most important feature of the remedy is that it must be *effective*. According to Sürmeli v. Germany, 75529/01, 8/6 2009, § 99, remedies are effective "if they prevent the alleged violation or its continuation, or provide adequate redress for any violation that has already occurred."

In order for the right to be effective, the citizen about whom a decision was made must be able to assess whether the decision was correct, and whether they want to appeal it. Therefore, it follows that the citizen must have access to enough information to make this assessment. This does not necessarily entail access to the underlying codes of any algorithm, as that does not in-and-of itself provide useful information to anyone who is not trained in programming languages. Instead, an explanation of the underlying logic of the algorithm, as well as the deciding factors would be more useful (more on this in: Burrell, 2016 ; Goodman and Flaxman, 2017).

An explanation for all factors considered would also be relevant, as that enables the citizen to point out any omissions. Likewise, the importance attached by the administration to the outcome of an algorithm – in view of potential action – would be a useful explanation. These explanations must show a complete picture of the way the algorithm is implemented, and which data it assesses. In order to ensure that the citizen is given the complete picture the administration must also be able to verify that the explanations are in accordance with the truth and it would probably, as a consequence hereof, need access to the underlying algorithm. In other words, the right to an effective remedy, as regards the use of algorithms in administrative decisions, first-and-foremost requires *transparency (*see below 4.4).

A part of the right to an effective remedy is the requirement that the authority stops the continuation of any violations of human rights. Therefore, if an algorithm is found to be in violation of human rights in one decision, the authority must be able to make any necessary corrections before it is used in other decisions. The right to an effective remedy thus requires the ability of the administration to make continuous changes to any algorithm used.

Read more in the ECtHR case-law guide on article 13.

3.4.4. Necessary in a democratic society

In a democratic society, for an interference with the right to privacy to be considered legal, it must be *necessary.* One issue which is particularly relevant here is the consideration of who an algorithm targets. Does it primarily target so-called vulnerable groups, as is the case with algorithms detecting

social benefit fraud, or does it aim at uncovering, say, terrorism. The target of the algorithm therefore matters greatly. The more potentially vulnerable the target, the greater the justification must be. Conversely, the more urgent the problem is to the state, such as an acute threat of terrorism, the less justification is needed, although there must always be one.

Economic considerations, such as whether the caseworkers are overwhelmed by the case load, can be said to be relevant to some degree, but are not particularly weighty. A democratic society should not need to violate human rights for economic reasons. Algorithms targeting vulnerable groups purely to lighten the case load of caseworkers are therefore potentially very problematic (Alston, 2019b).

One way to strengthen the argument for an algorithm being necessary in a democratic society is to make the algorithm target the system, instead of the individual. If, e.g., the algorithm is used to predict how long a case will take to reach a final conclusion (see e.g. Holten Møller et al., 2020), then such an algorithm would be targeting (and exposing) the flaws inherent in the administration, rather than the inadequacies of the jobseeker.

If an algorithm targeting vulnerable citizens is considered necessary, then the implementation of it should be supported by national law. The greater the democratic legitimacy, the easier it is to argue that a measure can be considered to be necessary in a democratic society.

4. Relevant principles

In the following, the various principles as they relate to both human rights law, general principles of administrative law, and the GDPR are covered in further detail. When possible, practical considerations relating to the HECAT project are discussed. While the document mostly refers to profiling which is currently in focus in public employment services, including with a view to using algorithms, many of the considerations will also apply to automated decision support systems which might be a more useful and participatory way forward particularly, but not exclusively, from the viewpoint of jobseekers.

4.1. Necessity

According to GDPR article 5.1.c (see Appendix A2), usage of personal data must be limited to the strictly necessary data. Necessity is only justified on the basis of objective evidence. In turn, the option which is least intrusive must be used. National case studies of statistical unemployment profiling illustrate two potential avenues for objectively assessing the necessity of different input variables: a top-down and a bottom-up approach. During the development of the profiling tool in the Netherlands, for example, an extensive top-down process was applied to identify relevant input variables (Wijnhoven and Havinga, 2014). First, the researchers conducted a literature review identifying three relevant theoretical paradigms, as well as 500 questionnaire items covering hard and soft factors to explain return to the labor market. Secondly, the researchers conducted a cross-sectional study of short-term and long-term unemployed, reducing the number of features to 70 factors. Finally, the researchers tested the predictive value of these factors in a longitudinal study, resulting in 11 factors to be included in their profiling model. This approach thus can serve to document the necessity of the included factors, as long as the other principles discussed, such as non-discrimination, are still upheld.

Alternatively, a bottom-up approach would not be based on a theoretical-deductive approach, but rather allow important variables to emerge from the data itself. This would entail using all available data as features in a predictive model, whether or not they seem relevant from a theoretical or expert basis, and identifying relevant variables on the basis of whether or not they contribute to the predictive power of the model. The method would be considered as research and therefore be allowed as long as the original data fulfills the usual conditions otherwise (quality, purpose limitation, etc.) and the resulting variables respect the conditions otherwise mentioned in this document.

4.2. Proportionality

To assess proportionality, the importance of the overall objective must be considered, as well as whether the included measures meet the objective. In short, the assessment of proportionality relates to whether the ends justify the means. For predictive analytics, as used in labor market profiling systems, the "means" are twofold: 1) the algorithm that learns to map specific inputs to predicted outcomes, 2) the historical data that provides examples of inputs and outcomes. To assess proportionality from a data protection standpoint, it is necessary to evaluate both the opportunities and threats of using different types and sources of data. If a type of variable infringes upon fundamental rights, it must be strictly necessary for the functioning of the model to be included. If a variable infringes upon fundamental rights and does not substantially contribute to the functioning of the model, it should be removed from the training data.

This raises the following questions: at what threshold is a variable considered to contribute substantially, and likewise, at what level of sensitivity is a variable considered to be too sensitive for inclusion? This motivates the establishment of a clearly defined balancing point between these thresholds of predictive contribution and sensitivity. For predictive analytics, the added value of utilizing sensitive data is generally measured in terms of a model's ability to correctly predict a chosen outcome, as judged by a chosen evaluation metric. Therefore, to evaluate the added value of including sensitive or risky data, it is necessary to consider both the definition of the outcome which the model is tasked with predicting, as well as how the models' predictions of this outcome are evaluated. Further evaluation on the impact of the program (e.g. how effective is the use of the algorithm in helping jobseekers to find a new job of decent quality) should also be taken into consideration.

A useful factsheet on necessity and proportionality can be found here.

4.2.1. The model outcomes in the case of profiling

In the case of profiling, the outcome variable of interest is whether a given person who has entered unemployment will remain in this state for a long enough period to be classified as long-term unemployed. Table 2 demonstrates varying definitions of long-term unemployment across national profiling systems. Notably, these different definitions of long-term unemployment will likely lead to different models and different success rates. Namely, it will likely be easier to achieve a higher level of accuracy when predicting 12 months of long-term unemployment than 6 months of long-term unemployment, since more individuals will re-enter the labor market over time and the model will only be tasked with identifying a limited number of high-risk individuals. In addition, the definition of the outcome variable, i.e. what is considered to be long-term unemployment, has substantial importance for proportionality because the influence of input variables will likely vary depending on

the time scale. The individual characteristics, personal skills, experiences and external labor market conditions that hinder or expedite returning to the labor market in the short-term will likely differ greatly from those that are at play for extended periods of unemployment. For example, (Meyers and Houssemand, 2010) compare the contribution of various "socio-professional" and psychological variables in order to assess which unemployed persons will return to the labor market, at various points in time. (Meyers and Houssemand, 2010) find that besides age socio-professional variables such as allowances, loans, and training can correctly predict employment status at 6 months in 74% of cases. In turn, demographic variables such as age and gender, combined with psychological variables, such as openness, self-efficacy, symptom-reduction coping, social anxiety and intelligence, allow the model to correctly predict employment status at 12 months in 75% of cases. These findings illustrate that there are different time-dependent factors at play depending on the chosen operationalization of long-term unemployment. The implication for proportionality is that the marginal utility of sensitive data may fluctuate over time, complicating the justification of the use of sensitive data. This might be less of an issue with regard to a decision support system where jobseekers receive recommendations for relevant occupations given their profile (e.g. previous occupation, skills, geographic constraints) and expectations (e.g. wages, working-time, type of contract). These factors might however come into play if the tool works at a very detailed level including conditioning on the time a jobseekers has spent in unemployment.

Table 2 Predicted outcome of profiling models across OECD countries

Country	**Predicted outcome of profiling model**
Australia	Long-term unemployment (12 months)
Austria	Short-term unemployment (3 months of unsubsidized employment within 7 months) Long-term unemployment (6 months of unsubsidized employment in 24 months)
Belgium (Flanders)	Long-term unemployment (> 6 months)
Denmark	Long-term (>26 weeks unemployment)
Ireland	Probability to exit at 12 months
Italy	Long-term unemployment (12 months)
Latvia	Long-term unemployment (12 months)
Netherlands	Long-term unemployment (12 months)
New Zealand	Lifetime income support costs
Sweden	Long-term unemployment (6 months)
United States	Exhausting 26-week entitlement to unemployment insurance benefits

Source: Table drawing on Desiere et al. 2019.

4.2.2. Choosing an evaluation metric in profiling models

In addition to considering how to define the desired outcome that a model is tasked with predicting, it is relevant to consider how to evaluate the quality of these predictions. There is no "golden rule" for the choice of evaluation metric by which to assess predictive models. In fact, certain evaluation metrics may favor certain algorithms, where some may favor prediction, recommendation, or utility maximization tasks (Gunawardana and Shani, 2009). Therefore, the choice of evaluation metric should depend on e.g. the characteristics of the problem at hand, the distribution of the data, or the needs of stakeholders and end-users. As we see in much of the preexisting reporting on statistical profiling (Desiere et al., 2019) "accuracy" is the golden standard for evaluating model performance.

In predicting a binary or multiclass outcome, as is generally the case in statistical profiling of long-term unemployment, accuracy is simply defined as the share of correctly predicted outcomes.

Accuracy as an evaluation metric has some serious shortcomings, urging a move "beyond accuracy" within the decision-support systems literature (McNee et al., 2006; Cremonesi et al., 2010; Ge et al., 2010; Kaminskas and Bridge, 2016; Nilashi et al., 2016). For example, accuracy does not take into account differences between false positives and false negatives in the assessment of predictions. If there are different costs associated with false negatives and positives, this can have serious impacts for the individual-level outcomes generated by the interaction with a model. In terms of labor market profiling, for example, false negatives would imply individuals who are categorized as low-risk of long-term unemployment, who actually become long-term unemployed. Conversely, false positives would imply individuals categorized as high-risk of long-term unemployment, who actually return to work quickly. While overinvesting in the low-risk group will lead to inefficiency at the administrative level, underinvesting the high-risk group may fail to reduce unemployment durations, thereby missing to reduce caseworkers' load and to improve the fate of the jobseekers most likely to be trapped in negative feedback loops involving poor metal health due to long-term unemployment (Oberholzer-Gee, 2008; Paul and Moser, 2009; Strandh et al., 2014). Therefore, the costs associated with false negatives and false positives within unemployment profiling are not symmetrical.

Predictive models can be assessed on numerous other criteria beyond their predictive skill, such as their efficiency, transparency and fairness. This complicates the assessment of the proportionality principle since the risk of using sensitive data should be assessed with regards to these different criteria which are not always met concomitantly. In particular, the recommender systems literature suggests that model accuracy does not always correlate with user satisfaction, urging a more user-centric approach (McNee et al., 2006). For example, mediating factors such as language and cultural background can influence users' interaction and satisfaction with decision-support tools (Rashid et al., 2002).

4.3. Data quality, accuracy, minimization and purpose limitation

According to the GDPR, the project should only collect and use data that is up to date and relevant (see section 3.1.1). This therefore calls for an assessment of the quality of data at the core of profiling models. The frameworks recognized useful for such assessment are few. An interesting example was published by Smith et al. (2018). In short, they distinguish five dimensions to assess the quality of administrative data (the complete framework can be found in Appendix A3).

(i) *Accuracy*: gives account of the completeness (how important are missing values), correctness (how important are values wrongly formatted, impossible values, or outliers), measurement error (wrong answers), level of bias (of the reported values relative to the real values) and consistency (would repeated measures provide the same table?) of the database.

(ii) *Internal validity*: measures the extent to which values of two different fields of the dataset are compatible (over time or across variables at a given time).

(iii) *External validity*: measures the extent to which elements of the dataset contradicts other data sources.

(iv) *Timeliness*: refers to whether the data is up-to-date and, for databases including waves, how quick it is to update.

(v) *Interpretability*: measures how easy it is to understand the content of the database, including with its documentation.

As the HECAT project is among others utilizing registry data from the PES in Slovenia, all (probably except for the last) dimensions should be assessed. While registry data has advantages over other forms of data, such as surveys, in terms of objectivity and coverage, it is not without its challenges and, if found to occur, limited quality of the administrative data would undermine the results of the project. For example, in the case of the Austrian profiling system, Holl et al. (2018) highlight missing registrations for young persons, recent migrants, and persons with discontinuous employment histories. The lack of completeness that is biased against this population impacts the performance of the model estimates and risks to limit the fairness of the algorithm (ibid).

Eurostat's labour force survey data will be another source of data inspiring the HECAT project. It needs to be carefully checked for accuracy (e.g. role of missing values, outliers and consistency) but, given its comparative nature which implies the use of different – though coordinated – questionnaires and the application of international standard classifications external validity can also pose a challenge. Breaks in the time series due to changes in the questionnaire as well as necessity to regularly update standard classifications can jeopardize internal validity. A well-known problem is the timeliness of data availability and the constraints in accessing the micro-data in a timely fashion. Deliverable 2.1 (Report on Supply and Demand Statistics in the Labour market) will provide a comprehensive census of relevant data sources for the HECAT pilot tool with a view on job quality items.

Further principles should be evaluated regarding the data. In particular, one should only use data that is necessary to sufficiently serve the quality of the profiling or model, and in a way that is compatible with the specific purpose for which the data was collected in the first place. This implies that it is relevant to reflect on how much of the jobseekers' (employment) history should be included in the model; that is, is there a point where it is no longer useful to "go further back in time"? For example, according to (Frid-Nielsen, 2019) information pertaining to the individuals' most recent occupation contributes a high level of accuracy for predicting the next occupation, with diminishing returns when using information on occupations further in the past. Limiting the data in terms of the time scope may not only help comply with the legal basis, but also offer an avenue to prune data and improve the efficiency of the modelling. Similarly, one should carefully consider whether information included in the model might carry the risk of discriminating jobseekers. The Danish Institute for Human Rights, for example, has recently put forward a complaint on the grounds of labour market discrimination against the Danish Agency for Labour Market and Recruitment due to the inclusion of information on ethical origin in their profiling tool (Institut for Menneskerettigheder 2020).

4.4. Transparency in the case of profiling

Transparency is discussed across many different sources of legalese. Section 3.4.3 has shown that human rights law compels the administration to provide enough non-technical information for the unemployed to understand or assess: (i) the level of importance that the model prediction plays in the final decision; (ii) the quality of the prediction; (iii) which features are predominantly used. The GDPR adds that that consent is necessary and must be explicit. In particular, it requires an explanation of the underlying logic and consequences of profiling models to be provided to the person in question (article 14 and 15). To some extent, the same kind of explanation is needed for

decision-support systems. General principles of administrative law also require a statement of reasons for decisions affecting the jobseeker. National law may have similar requirements.

The principle of transparency should lead us to take a number of decisions with regard to the HECAT algorithm and visualization tool. Note that these considerations need to be taken into account from the start of the development of the algorithm (GDPR, article 25). First, jobseekers should be able to access information collected about them and be notified about the collection of data (GDPR, article 13 to 15). Second, jobseekers and caseworkers should both be aware of the reason and the purpose of data processing by the algorithm. Third, they should know why the model provides specific predictions to their relevant case(s). While it is probably not necessary to provide caseworkers and unemployed with the full code, the main deciding factors should be clear. Jobseekers and caseworkers should also understand that changing specific inputs (say one's municipality of living) impacts the probability to return to employment.[3] Specifically for caseworkers, it matters that they know how the algorithm works, and are familiar with the various types of limitations and biases connected to the use of decision-supporting algorithms[4]. Training sessions on the fallacies of the algorithm should therefore be considered (e.g. in the case of a profiling algorithm, if the algorithm has a 80% rate of accuracy, then the case-worker should potentially correct 1 in 5 cases; an algorithm tends to be less accurate when it comes to minority groups).

The predominant level of decision should also be explicitly stated and systematically respected. What is the exact decision power of caseworkers and can they exert it with no constraint? In particular, the extent to which each caseworker sticks to or corrects the predictions of the profiling model should not be used in assessments of her productivity or affect in any way her career. This must be explicitly stated when profiling is rolled out to ensure that caseworkers are not afraid to decide against the model predictions. It matters as well that jobseekers know that they can appeal against the decision made by the caseworker on the basis of the prediction with no subsequent impact on the quality of services the PES will provide them later. A question remains of whether this is possible given that profiling is inherently designed to offer different services to jobseekers.

Finally, researchers, politicians and data protection agencies will need to be able to examine the algorithm in detail if willing to assess it or to build on it. There are several data science approaches for opening up "black box" machine learning models. Traditionally, feature importance measures (Dash and Liu, 1997) can shed light on the influence of individual input variables on the models' predicted outcomes. More recent approaches further increase model transparency by building surrogate models that are more easily interpreted by humans, and can illustrate how classification decisions are made for specific individuals (Ribeiro et al., 2016; Ghorbani and Zou, 2019).

More inspiration regarding transparency can be found in ICO's guide to explaining decisions made with AI here.

4.5. Fairness

Different groups may be statistically disadvantaged or challenged in their labor market opportunities, this, in turn, can lead to differing distributions of risk of long-term unemployment (e.g. gender,

[3] On the question of transparency of the variables and their explanatory power, note the potential disincentive effect for jobseekers whose fixed (i.e. non-changeable) characteristics are associated with low probability of re-employment.

[4] Such as decision-automation bias and automation-distrust bias. More on this here.

ethnicity etc.). Machine learning algorithms learn to map relations between input features to outcomes on the basis of historical data. This implies that profiling tools based on these approaches may learn to replicate inequalities present in society (Dwork et al., 2012; Feldman et al., 2015; Barocas and Selbst, 2016; Corbett-Davies and Goel, 2018). In this case, since the outcome variable of interest for algorithmic profiling tools is to predict the probability of long-term unemployment, which may vary substantially across vulnerable labour market groups, there is the risk that labor market profiling tools may lead to discriminatory treatment of individuals. (Desiere and Struyven, 2020), for example, demonstrate for the Belgium case that an advanced AI-based profiling model is 2.6 times more likely to misclassify foreign-born jobseekers as high-risk compared to natives. This problem arises since there is a statistically higher risk of long-term unemployment among foreign-born jobseekers. Since the probability threshold for when to classify an individual as high-risk is based on the majority group, this leads to a greater level of misclassification among the high-risk minority group. Interestingly, the authors also show that using this variable (place of birth) is useful in terms of accuracy since it is a good predictor of individuals' likelihood to fall in long-term unemployment. They therefore highlight an equity-accuracy tradeoff.

On a more practical note, the predictions of profiling algorithms can be linked to differential administrative treatment by caseworkers, such as whether a client is to be referred to interventions focused on job training or job-seeking motivation (Allhutter et al., 2020). Therefore, there is evidence that statistical disparities in data can have direct impact for the treatment of individuals in job centers.

Assessing the fairness of a profiling algorithm is therefore crucial. But how to do so? The literature has used different fairness criteria that can be mutually exclusive, therefore this choice must be considered carefully based on the task at hand and knowledge of the protected groups (Kleinberg et al., 2016). (Gajane, 2017) provides a starting point by contrasting and comparing the technical formalization of key definitions of fairness in algorithmic predictions, highlighting trade-offs in their implementations. Gajane (2017) contends that no definition of fairness is a panacea, and therefore urges consideration of relevant social issues such as access to resources and social norms. The fairness criteria of group blindness, for example, simply removes the protected class as a model input, generating fairness through unawareness by avoiding the explicit use of sensitive information in the prediction process. However, Gajane (2017) argues that group blindness may nonetheless lead to unfairness if the outcomes of profiling are left unchecked. In the case of labor market prediction, for example, protected classes tend to correlate with other key variables such as wages and occupational groups, where their inclusion can lead to discrimination without the algorithm being made aware that different groups exist. Group fairness, on the other hand, explicitly utilizes information about protected classes in the modelling process, making these two fairness definitions incongruous. Group fairness imposes a condition on the model stipulating that it must predict the positive class (i.e. long-term unemployment) with a similar probability across protected and unprotected groups—a form of affirmative action. For example, (Russell et al., 2017) proposes the injection of fairness constraints into the optimization of the model during the training phase. However, group fairness will likely be inefficient for predicting long-term unemployment, where the distribution of risk varies among groups, as highlighted by (Desiere and Struyven, 2020).

4.6. Protection of sensitive data

A proposed solution for reducing data privacy concerns is to remove sensitive features from training data. This can include personal characteristics such as gender, age and ethnicity. As mentioned in

section 4.5, the added benefit of removing personally-identifiable features from training data is that it may reduce the extent to which the trained model will discriminate users on the basis of these features, by e.g. learning to map relations between sensitive features and certain outcomes. However, removing features for the purpose of increasing privacy, fairness and transparency also presents several limitations.

First, after removing such sensitive variables from training data, it may not be possible to identify bias or discrimination anymore – and thereby correct it. One practical possibility is to exclude sensitive data from the training data but use it, in a second step, to check whether the algorithm is unintentionally discriminatory. Second, as mentioned in section 4.5, there is evidence that sensitive variables can increase equity and efficiency, sometimes drastically (Kleinberg et al., 2018; Desiere and Struyven, 2020). Our model will therefore perform less well than the state of the art that is currently using such sensitive data. Note as well that removing the most obvious candidates in terms of sensitivity of the variables may not be enough to mitigate privacy concerns. Even seemingly neutral features such as geography may lead to discriminatory profiling (Goodman and Flaxman, 2017). For example, postal codes may correlate with higher minority ethnic groups (Calders and Verwer, 2010), meaning that profiling models utilizing geographical data as inputs may be indirectly discriminatory.

The literature has proposed computational methods more advanced than the simple removal of variables to measure and deal with the risk of data protection (Calders and Verwer, 2010; Hajian et al., 2011; Žliobaite et al., 2011; Berendt and Preibusch, 2012; Feldman et al., 2015; Goodman and Flaxman, 2017). Note however that these methods may decrease the transparency of the algorithm and thereby negatively affect its legality. Further, modifying sensitive data for the purpose of depersonalization may come at the cost of overall decreased accuracy.

5. A short illustration of potential pitfalls in the case of the HECAT algorithmic pilot tool

In this section, we focus on some practical considerations from the previous sections. In the most recent version of the algorithmic pilot tool, some variables had larger importance than others. We list them in table 3 and provide an assessment of the associated risks in terms of discrimination and privacy.

As an illustration we tentatively apply some of the learning points from the above sections to the outcome of this specific version of the algorithm. The approach chosen by work package 3 was bottom-up (see section 4.1). Six variables of the list provided the most explanatory power for the model: (i) age; (ii) work experience; (iii) reason for contract termination[5]; (iv) whether one receives social support benefits; (v) whether one receives unemployment benefits; (vi) employability assessment by PES counsellor[6].

[5] This variable can take a very large number of values. The ones with the highest frequencies are: end of a fixed-term contract, financial difficulties of the firm, bankruptcy, mutually agreed termination, own firm closure, layoffs on personal grounds. Note that the list also includes some more problematic items such as termination due to disability.

[6] This variable can take six different values: directly employable; employable with additional actions; temporary unemployable; permanently unemployable; employable with intense and deep support; no assessment.

Using work experience seems highly justified on the basis that it is likely to be directly relevant for labor market re-integration and the risk with respect to data privacy is limited. The other variables might put us on a more slippery slope, as they can be directly linked to an individual and thereby used for re-identification in the case of a data breach. With the exception of the employability assessment by PES counsellors, they also raise the risk of discrimination within the labor market on the basis of a specific category.

To limit the risks, a first concrete step to take would be to explore the extent to which the most disputable variables correlate with others that are less problematic. Using aggregate measures at a local level (e.g. number of job openings at the occupational code level, unemployment rates, etc..) may also be an alternative way to proxy the probability of receiving unemployment benefits and social assistance as well as to predict layoffs in a less problematic and more targeted way. Replacing these variables might reduce the accuracy of the model but it would also render it more in line with the above-mentioned principles.

The final choice of whether one should include a variable is the result of a cost-benefit arbitrage, where the discrimination and privacy risks are weighted against the accuracy of the model. Given the very high predictive power of age in the most recent version of the algorithmic pilot tool, one may advocate for its use based on the medium risk assessment pertaining to privacy. Basing a decision on age is, however, discriminatory (see section 3.4.2), and such impact assessment must therefore be conducted.

Note that even more problematic variables could have emerged among the most predictive ones (e.g. the country of origin as in the Austrian case, gender, municipality, etc.). We are surely on a safer side here, but one should keep in mind that these variables are still being used in the background. To follow principles of proportionality & necessity, one would probably rather exclude these variables from the model given the low benefit they bring in terms of accuracy and the privacy risk they contain.

AI algorithms are black boxes that allow gains in accuracy to the detriment of transparency. As much as possible should be done to open this black box and increase transparency, especially when the algorithm directly impacts peoples' lives as is the case with our pilot which, moreover, targets vulnerable labour market groups among others. From the start, avoiding the use of the most problematic input variables could be a solution. Absent these efforts, an impact assessment is hardly achievable, and the algorithm-based tool may thus prove illegal.

This section is only intended for illustrative purposes; a detailed assessment of issues such as data quality, accuracy and transparency would have to be performed on the final version of the model (see section 2).

Table 3 Overview of the most important variables used in the process of building the algorithm piloting tool in Slovenia

Feature's importance for model - ranking	Variables	Risk with respect to discrimination	Risk with respect to privacy	Comment
1	Age	Medium	Medium	High importance sufficient to justify the privacy/discrimination risk?
2	Work experience	Low	Low	Relates tightly to labour market attachment, it is easier to justify
3	Reason for contract termination	Medium	Medium	Some minor categories relating to health issues are problematic
4	Social support benefits	Medium	Medium	Problematic to some extent as benefit receipt identifies vulnerable individuals and here additionally distinguishes between two groups. As it relates tightly to labour market attachment, it's easier to justify though.
5	Unemployment benefits	Medium	Medium	
6	Employability assessment by PES counsellors	Low	Medium	Emerges from the discussion with the PES counsellor, which limits the risk of discrimination. However, privacy risks are not absent as this variable is highly sensitive.

Authors' own assessment.

6. Conclusion

This document has presented the legal considerations that one should take when using data and algorithms for profiling jobseekers in public employment services. Many of the considerations are also relevant with regard to use of individual-level data in PES beyond profiling for example for the purpose of data-driven decision support systems. It is of utmost importance to carefully follow these legal considerations as it regularly happens that experimentations are aborted or challenged for non-respect of the legalese presented here. To give but one example, in 2020, the Austrian algorithmic profiling tool intended to classify jobseekers was shut down by the data protection authority due to a lack of a sufficient legal basis, even before it was rolled out. It had previously been highly criticized on grounds of potential discrimination and lack of transparency.

The most important conclusion from the above analysis is that the HECAT algorithm and the underlying data (irrespective of the type of data – administrative, survey, user subscription) need to be continuously assessed with a view to the right to privacy of the jobseeker. Such impact assessment is relevant both for the design as well as the implementation phase of the project. Crucially, the requirements for administrative safeguards and restrictions to the underlying data will vary with the expected impact of the algorithm – the higher the expected impact on the jobseeker the stricter the requirements. Legal requirements to consider are to be found in national (administrative) law, in EU law, particularly the GDPR, as well as Human Rights law. Importantly, all considerations vis-a-vis the data and the algorithm need to be documented and there needs to be a legal basis for the algorithm and the underlying data. The legal requirements will shape the subsequent impact assessment of the algorithm. As administrations such as the PES need to make decision that are both proportional and fair, these two principles need to be at the core of such an assessment.

The data sources and variables considered for the algorithm need to be carefully checked for quality; this implies considerations of accuracy, internal validity, external validity and timeliness. The importance of certain variables for the model needs to be weighed against the potential discrimination that could arise from including them. People can be treated differently if the actual situation requires it – a case in point would be to distinguish between different levels of education or previous (un)employment experience as these are known to impact the likelihood for labour market re-integration. Positive discrimination is possible, and people can be treated differently, if a case can be made that this will contribute to correct factual inequalities; this needs to considered carefully and documented.

Accuracy of the model in terms of predictive power is often in focus; this said accuracy is an evaluation metric with important shortcomings which has led to a move to go beyond accuracy within the decision-support systems literature and assess predictive models on additional grounds such as efficiency, transparency and fairness. Statistical profiling can for example imply equity-accuracy trade-offs in terms of disadvantaged labour market groups. It is therefore crucial to assess the fairness of a profiling algorithm and the administrative decision that follows from it.

Transparency is of key importance. Caseworkers need to know how the algorithm works and need to be made aware of the limitations and biases of decision-supporting algorithms (i.a. disproportional mis-classification of certain groups of jobseekers) so that they can correct for this where relevant. It is thus of prime importance to grant caseworkers discretion to over-rule automated decisions; in fact according to the GDPR individuals shall have the right to not be subject to a decision based solely on automated processing, unless national law explicitly allows it, and said law also considers how to protect the fundamental rights of these individuals. Ensuring transparency with regard to the

algorithm, the underlying data and the way the algorithm is implemented in the public employment service is also of prime importance to the concerned jobseeker. The jobseeker needs to understand why a decision was made and on which grounds. In the case of profiling, this puts him/her into a position to challenge an administrative decision (the right to effective remedy).

While this paper primarily draws on profiling as the currently most commonly used activity in terms of algorithmic decision making in public employment services, many of the above considerations will also apply to data-driven decision support systems as currently envisaged in HECAT. Again, a careful evaluation of the tool and underlying data is of prime importance. The more detailed and individual the underlying data, the more problematic. We explore this in more details in deliverable 2.1 (Report on Supply and Demand Statistics in the Labour market) focusing on different types of data including survey, administrative and user provided data with a view to job quality items.

In contrast to allocating jobseekers to different support categories as commonly done in profiling, data-driven decision support systems aim to visualize a range of alternative employment opportunities rather than prescribing certain activities. Overall, such algorithms will therefore likely be less intrusive (lower impact level on the jobseeker) which, in turn, implies fewer necessity for administrative safeguards and restrictions on the data to be used. It is, however, still important that these considerations are performed, and documented, to ensure compliance with the relevant regulations.

Appendix

A1. Definitions

This appendix is intended to introduce the reader to selected definitions made by the GDPR. As such, it shortly describes the various terms, bullet points enabling the reader to quickly find specific information.

A1.1. Personal data

- GDPR: "any information relating to an identified or identifiable natural person (data subject) (GDPR Article 4)
    - Highly sensitive: health, genetic, biometric data
    - Processing of sensitive data is allowed under GDPR if it is "necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical" (GDPR: art 9.2.j) and assuming safeguards are in place (GDPR article 89)
- Data which allows for *direct* or *indirect* identification of the data subject, through e.g. triangulation (Forcier et al., 2019)
- Name, identification number, location data, "factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"
- Three definitions for personal data (Pangrazio and Selwyn, 2019)
    - Data given to devices/systems by users: consciously given by the user to the system, uploaded, typed in, etc.
    - Data extracted from devices/systems: brought into existence by being collected—organizations and institutions control them
    - Data processed by devices/systems on behalf of users: dashboards, visualizations, etc.
- De-identified data is still considered identifiable data according to GDPR (Forcier et al., 2019)
    - "To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."

A1.2. Profiling

'Profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements; (GDPR article 4(4))

A1.3. Processing

'Processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction

A1.4. "Special categories" of data according to GDPR (Article 9)

- Racial or ethnic origin

- Political opinions, religious beliefs
- Trade-union membership
- Genetic data
- Biometric data
- Health (potentially including pregnancy, and disabilities)
- Sex life or sexual orientation

A1.5. The "data controller"

The natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data (GDPR article 4.7)

A1.6. Safeguards

The GDPR often requires so-called "safeguards" to be put into place, in order for data processing to be legal. The term "safeguard" is a catch-all for formal, technical and procedural measures put into place in order to ensure no violation of the GDPR takes place. Therefore, a safeguard can take many forms, such as the individual right to judicial remedy, encrypting sensitive data, or ensuring all relevant personnel know the relevant procedures to follow should a security breach happen. The exact safeguards required in a specific situation should be determined by the data controller prior to processing, e.g. as a consequence of a DPIA being performed.

A1.7. Necessity

According to the European Data Protection supervisor (quoted from here), ***"Necessity*** is a fundamental principle when assessing the restriction of fundamental rights, such as the right to the protection of personal data. According to case-law, because of the role the processing of personal data entails for a series of fundamental rights, the limiting of the fundamental right to the protection of personal data must be strictly necessary.

Necessity shall be justified on the basis of objective evidence and is the first step before assessing the proportionality of the limitation. Necessity is also fundamental when assessing the lawfulness of the processing of personal data. The **processing operations, the categories of data processed and the duration the data are kept shall be necessary for the purpose of the processing**."

A1.8. Proportionality

According to the European Data Protection supervisor (quoted from here), ***"Proportionality*** is a general principle of EU law. It restricts authorities in the exercise of their powers by requiring them to strike a balance between the means used and the intended aim. In the context of fundamental rights, such as the right to the protection of personal data, proportionality is key for any limitation on these rights.

More specifically, proportionality requires that advantages due to limiting the right are not outweighed by the disadvantages to exercise the right. In other words, the limitation on the right must be justified. Safeguards accompanying a measure can support the justification of a measure. A pre-condition is that the measure is adequate to achieve the envisaged objective. In addition, when assessing the processing of personal data, **proportionality requires that only that personal data which is adequate and relevant for the purposes of the processing is collected and processed**."

A2. Relevant legal texts

a. For the Impact Assessment

- GDPR art. 5. More details below
- GDPR art 35(1) – "Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data."

The GDPR does not provide an exhaustive list of high risk processing. For a practical assessment, the article 29 of the Data Protection Working Party (Article 29 Data Protection Working Party, 2017) mentions that if two of the following nine criteria are met, then a Data Protection Impact Assessment must be performed (we highlighted the criteria that are most likely to be met by the project).

  - Evaluation or scoring
  - **Automated decision-making—with "legal or similar effect"**
  - Systematic monitoring
  - **Sensitive data or data of a highly personal nature**
  - **Data processed on a large scale**
  - **Matching or combining datasets**
  - **Data concerning vulnerable data subjects**
  - Innovative use of technological or organizations solutions
  - Preventing data subjects from exercising rights

- GDPR art 25 – data protection by design and default
- ECHR art 8, CFR art 7, 8 and 52 – profiling is considered an interference in the right to privacy, and the right to data protection. This can be justified, if a) basis in law, and b) proportionate and necessary in democratic society. To show that it is proportionate, the impact must be considered, and mitigated if necessary.
- General administrative law principles

b. GDPR article 5 - Principles relating to processing of personal data

1. Personal data shall be:

   (a) processed lawfully, fairly and in a transparent manner in relation to the data subject **('lawfulness, fairness and transparency')**;

   (b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes **('purpose limitation')**;

   (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed **('data minimisation')**;

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay **('accuracy')**;

(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject **('storage limitation')**;

(f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures **('integrity and confidentiality')**.

2. The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 **('accountability')**.

N.B. emphases added by the authors.

A3. Data quality framework

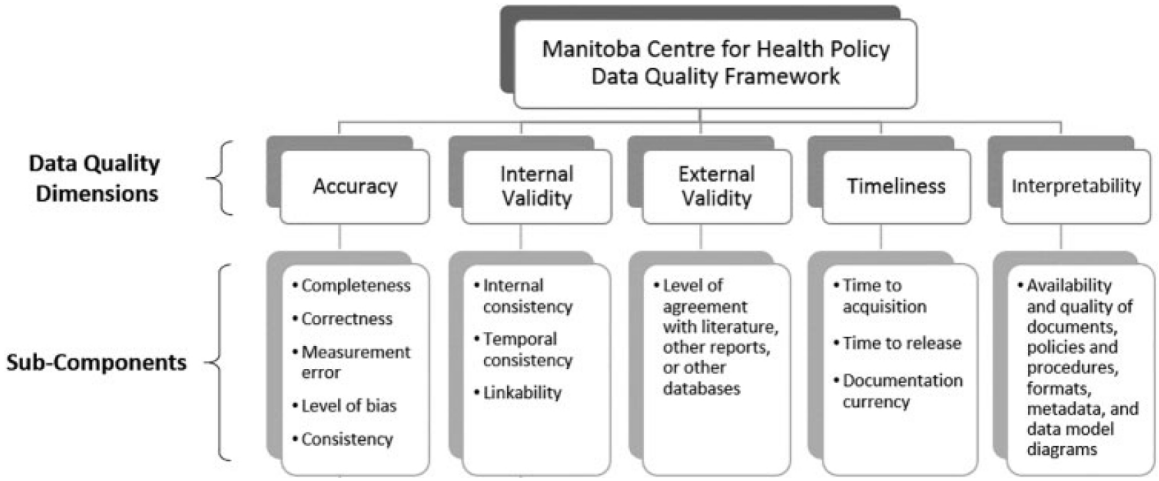


Figure 2 Smith et al's (2018) framework to assess the quality of administrative data

References

Allhutter, D., F. Cech, F. Fischer, G. Grill and A. Mager (2020), Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective, Frontiers in Big Data 3, 5.

Alston, P. (2019a), Amicus brief in the case of NJCM c.s./De Staat der Nederlanden (SyRI): Implications of the use of digital technologies in welfare states.

Alston, P. (2019b), Report of the Special Rapporteur on extreme poverty and human rights nr. A/74/493 Extreme poverty and human rights.

Article 29 Data Protection Working Party (2017), Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679.

Barocas, S. and A.D. Selbst (2016), Big Data's Disparate Impact, California Law Review 104, 671.

Berendt, B. and S. Preibusch (2012), Exploring Discrimination: A User-centric Evaluation of Discrimination-Aware Data Mining, December.

Burrell, J. (2016), How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society 3 (1), 2053951715622512.

Calders, T. and S. Verwer (2010), Three naive Bayes approaches for discrimination-free classification, Data Mining and Knowledge Discovery 21 (2), 277–292.

Corbett-Davies, S. and S. Goel (2018), The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.

Council of Europe (1980). Rec(80)2 11/03/1980 concerning the exercise of discretionary powers by administrative authorities.

Council of Europe (1977), Res(77)31 28/09/1977 on the protection of the individual in relation to the acts of administrative authorities.

Cremonesi, P., Y. Koren and R. Turrin (2010), Performance of recommender algorithms on top-n recommendation tasks.

Dash, M. and H. Liu (1997), Feature selection for classification, Intelligent Data Analysis 1 (1), 131–156.

Desiere, S., K. Langenbucher and L. Struyven (2019), Statistical profiling in public employment services.

Desiere, S. and L. Struyven (2020), Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off, Journal of Social Policy, 1–19.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold and R. Zemel (2012), Fairness through Awareness.

Feldman, M., S.A. Friedler, J. Moeller, C. Scheidegger and S. Venkatasubramanian (2015), Certifying and Removing Disparate Impact.

Forcier, M.B., H. Gallois, S. Mullan and Y. Joly (2019), Integrating artificial intelligence into health care through data access: can the GDPR act as a beacon for policymakers?, Journal of Law and the Biosciences 6 (1), 317–335.

Frid-Nielsen, S.S. (2019), Find My Next Job: Labor Market Recommendations Using Administrative Big Data.

Gajane, P. (2017), On formalizing fairness in prediction with machine learning.

Ge, M., C. Delgado and D. Jannach (2010), Beyond accuracy: Evaluating recommender systems by coverage and serendipity.

Gerards, J. (2019), General Principles of the European Convention on Human Rights, Cambridge University Press.

Ghorbani, A. and J. Zou (2019), Data Shapley: Equitable Valuation of Data for Machine Learning.

Goodman, B. and S. Flaxman (2017), European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation", AI Magazine 38 (3), 50–57.

Gunawardana, A. and G. Shani (2009), A Survey of Accuracy Evaluation Metrics of Recommendation Tasks, Journal of Machine Learning Research 10, 2935–2962.

Hajian, S., J. Domingo-Ferrer and A. Martínez-Ballesté (2011), Discrimination prevention in data mining for intrusion and crime detection, April.

Holl, J., G. Kernbeiß and M. Wagner-Pinter (2018), Das AMS-arbeitsmarktchancen-Modell.

Holten Møller, N., I. Shklovski and T.T. Hildebrandt (2020), Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services, conference paper, New York, NY, USA.

Institut for Menneskerettigheder (2020), Klage til Ligebehandlingsnævnet, JR. NR. 19/02577, 3 JUNI 2020.

Kaminskas, M. and D. Bridge (2016), Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems, ACM Trans. Interact. Intell. Syst. 7 (1).

Kleinberg, J., J. Ludwig, S. Mullainathan and A. Rambachan (2018), Algorithmic Fairness, AEA Papers and Proceedings 108, 22–27.

Kleinberg, J., S. Mullainathan and M. Raghavan (2016), Inherent Trade-Offs in the Fair Determination of Risk Scores.

Kuner, C., L.A. Bygrave, C.A. Docksey and L. Drechsler (2020), The EU General Data Protection Regulation (GDPR): A Commentary, Oxford University Press.

McNee, S.M., J. Riedl and J. Konstan (2006), Accurate is not always good: How Accuracy Metrics have hurt Recommender Systems.

Meyers, R. and C. Houssemand (2010), Socioprofessional and psychological variables that predict job finding, European Review of Applied Psychology 60 (3), 201–219.

Motzfeldt, H.M. and A. Næsborg-Andersen (2018), Developing Administrative Law into Handling the Challenges of Digital Government in Denmark, The Electronic Journal of e-Government 16 (2), 136–146.

Nilashi, M., D. Jannach, O. bin Ibrahim, M.D. Esfahani and H. Ahmadi (2016), Recommendation quality, transparency, and website quality for trust-building in recommendation agents, Electronic Commerce Research and Applications 19, 70–84.

Oberholzer-Gee, F. (2008), Nonemployment stigma as rational herding: A field experiment, Journal of Economic Behavior & Organization 65 (1), 30–40.

Pangrazio, L. and N. Selwyn (2019), 'Personal data literacies': A critical literacies approach to enhancing understandings of personal digital data, New Media & Society 21 (2), 419–437.

Pasquale, F. (2015), The Black Box Society, Harvard University Press.

Paul, K.I. and K. Moser (2009), Unemployment impairs mental health: Meta-analyses, Journal of Vocational Behavior 74 (3), 264–282.

Rashid, A. et al. (2002), Getting to Know You: Learning New User Preferences in Recommender Systems, International Conference on Intelligent User Interfaces, Proceedings IUI.

Ribeiro, M., S. Singh and C. Guestrin (2016), "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Russell, C., M.J. Kusner, J. Loftus and R. Silva (2017), When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness, in: Guyon, I. et al. (eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 6414–6423.

Smith, M. et al. (2018), Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy, Journal of the American Medical Informatics Association 25 (3), 224–229.

Strandh, M., A. Winefield, K. Nilsson and A. Hammarström (2014), Unemployment and mental health scarring during the life course, European Journal of Public Health 24 (3), 440–445.

Szigetvari, A. (2020), Datenschutzbehörde kippt umstrittenen AMS-Algorithmus - derStandard.at, https://www.derstandard.at/story/2000119486931/datenschutzbehoerde-kippt-umstrittenen-ams-algorithmus, retrieved 1.3.2021.

Warren, S.D. and L.D. Brandeis (1890), Right to Privacy, Harvard Law Review 4, 193.

Wijnhoven, M.A. and H. Havinga (2014), The Work Profiler: A digital instrument for selection and diagnosis of the unemployed, Local Economy 29 (6–7), 740–749.

Žliobaite, I., F. Kamiran and T. Calders (2011), Handling Conditional Discrimination, December.

Supporting Material

«The Limitations to the Use of Data for Profiling Purposes with Regard to Data Protection Rules, in Particular for Research»

Authors: Natasha Hauser[7], Viktor Györffy, Roxana Paz

I. Abstract

This paper analyses the limitations of data for profiling purposes with regard to the applicable law. First, the legal framework in general is provided. Secondly, the general principles of data processing in Switzerland are outlined. Different legal principles are relevant for data processing by a private person (entity) and for data processing by Regional Employment Agencies (RAV resp. PES). Therefore, the important and relevant legal provisions with regard to the planned research work of a private person, using data from unemployed persons (using questionnaires) will be considered first, followed by the legal basis for processing of data provided by the PES.

AI. Legal Framework

Data protection rules apply when personal data are processed. Personal data means any data relating to an identified or identifiable natural person, Art. 3 lit. a of the Swiss Data Protection Act (DSG). DSG does not apply when anonymous or statistical data (as far as no reference to persons is possible any more) are processed. Pseudonymized data are considered as personal data within the meaning of Art. 3 lit. a of the DSG according to federal court rulings and in accordance with further explanations in this paper. Therefore, all relevant data protection provisions are relevant in this case.

1. Applicability General Data Protection Regulation

According to Art. 3 of the General Data Protection Regulation (GDPR), the GDPR is applicable to Swiss companies if the processing activities are related to

- the offering of goods or services to data subjects in the Union; or
- the monitoring of their behaviour as far as their behaviour takes place within the Union (observation of conduct in the EU).

Swiss companies that process data of EU citizens, which are located in Switzerland, are therefore not subject to the GDPR. Accordingly, the DSG is generally applicable to data processing by a private company. Thus, the nationality of a person (EU citizen or non-EU citizen) is not relevant in this context.

2. Applicability Swiss Data Protection Act

[7] Natasha Hauser and Viktor Györffy were the main authors of the paper. Roxana Paz conceptualized the paper, revised and fine-tuned the drafts.

The Swiss data protection law will be completely revised. The applicable DSG in Switzerland will be adapted to the GDPR. The date on which the revised DSG (rev-DSG) will be in force has not yet been determined. Relevant for the data processing in the present case is Swiss law (DSG / rev-DSG).

We are assuming that the GDPR will not be applicable 1) because it is not the characteristic of citizenship of the unemployed and employees that is relevant, but the activity of the company (market place principle of the GDPR) and 2) as long as there is no processing concerning the observation of conduct in the Union.

BI. «Personality Profile» and «Profiling»

In the still applicable DSG, in force since 1 July 1993, the term "personality profile" is used. According to Art. 3 lit. d of the DSG, a personality profile is "a compilation of data that allows an assessment of essential aspects of the personality of a natural person". This term is known only in Switzerland and will be replaced by the term "profiling" in the revision of the DSG (rev-DSG).

The terms are not congruent. A personality profile is the result of an editing process; therefore, it comprises something static. Profiling according to the GDPR is a dynamic process that is geared to a specific purpose. The term "personality profile" in the DSG is to be replaced by the term "profiling" in the rev -DSG and thus, adapted to European terminology. In the future, "profiling" in the rev-DSG will therefore mean the assessment of certain characteristics of a person based on automatically processed personal data. It is a completely automated processing. For “profiling” to take place, the data must be evaluated, e.g., for the purpose to analyse or predict certain behaviours of a person.

IV. Data Processing by a private person

If non-anonymised or pseudonymised data are processed by private parties, the DSG applies. In the present case, the DSG applies as the legal basis, respectively the rev-DSG once it will be in force.

1. Collected Data

Data from social insurances are personal data in accordance with Art. 3 lit. as of the DSG (Art. 5 lit. a rev-DSG). A person can no longer be determined if personal data has been anonymized. The personal reference has thereby been irreversibly removed. The decisive factor here is that the personal reference can no longer be made with reasonable effort. Anonymized data is not subject to the DSG. With pseudonymization, however, the identifying data is merely replaced by a neutral dataset and a key for re-identification is stored. This data remains personal data. Such data is no longer personal data only for outsiders who do not have access to a key. However, there is a risk that the combination and analysis of larger amounts of data may make the data subjects identifiable again (even if found by accident). Only completely anonymous data is no longer personal data. For this reason, pseudonymized data is in principle still considered personal data and therefore, data processing must be in accordance with the regulations of the data protection law to ensure sufficient protection of fundamental rights.

Highly sensitive personal data pursuant to Art. 3 lit. c of the DSG is data about the religious, ideological, political or trade union views or activities; health, privacy, or racial affiliation; measures of social assistance;

administrative or criminal prosecutions and sanctions. Art. 5 lit. c of the rev-DSG additionally includes data about ethnic affiliation, genetic data and biometric data that uniquely identify a natural person.

The DSG provides for personality profiles the same qualified legal consequences respectively stricter regulations as for highly sensitive personal data.

Art. 3 lit. c number 2 and 3 of the DSG (Art. 5 lit. c number 2 and 6 of the rev-DSG) includes in particular data that provide information about health, privacy or race, or information about social assistance measures. Data that is provided to the PES belongs to this category. Consequences of qualification as data requiring special protection are qualified legal requirements such as explicit consent if consent is necessary (Art. 4 para. 5 DSG), registration in the register of data collections (Art. 11a of the DSG) and obligation to inform the concerned person (Art. 14 of the DSG).

The rev-DSG defines as profiling any form of automated processing of personal data consisting of using such data to assess certain personal aspects relating to a natural person, in particular to analyse or predict aspects relating to that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, whereabouts or change of location (Art. 5 lit. f of the rev-DSG). Profiling that entails a high risk for the personality or the fundamental rights of the data subject by leading to a linkage of data that allows an assessment of essential aspects of the personality of a natural person is considered high-risk profiling (Art. 5 lit. g of the rev-DSG). High-risk profiling requires explicit consent. Since it is not yet clear what will be considered high-risk profiling in practice, it is recommended to obtain explicit consent for profiling in the specific context discussed here.

In the context discussed here, a "profiling" with pseudonymised, particularly sensitive data for research purposes takes place. The data protection regulations must be complied with.

2. General Principles of Data Processing

Any use of personal data, regardless of the means and procedures used, is considered as processing; included is especially the obtaining, storing, using, transforming, disclosing, archiving, or destroying of data.

The following principles, that are embodied in Art. 4 ff. of the DSG (Art. 6 ff. of the rev-DSG), must always be observed and complied with when processing data.

a. Principle of Legality (Art. 4 para. 1 of the DSG; Art. 6 para. 1 of the rev-DSG)
b. Good Faith (Art. 4 para. 2 of the DSG; Art. 6 para. 2 of the rev-DSG)
c. Proportionality (Art. 4 para. 2 of the DSG; Art. 6 para. 2 of the rev-DSG)
d. Principles of Purpose (Art. 4 para. 3 of the DSG; Art. 6 para. 3 of the rev-DSG)
e. Transparency and Recognizability (Art. 4 para. 4 of the DSG; Art. 6 para. 3 of the rev-DSG)
f. Data Correctness and Security (Art. 5 and 7 of the DSG; Art. 6 para. 5 and Art. 8 of the rev-DSG)
g. Privacy by Design and Privacy by Default (Art. 7 of the rev-DSG)

3. Data disclosure to foreign countries

Data processing also includes the disclosure of data (c.f. Art. 3 lit. e and lit. f of the DSG; Art. 5 lit. d and lit. e of the rev-DSG). The core provision on the transfer of data abroad is Art. 6 of the DSG. Other provisions of the DSG and the VDSG supplement or elaborate on this article. Further information can be found at: https://www.edoeb.admin.ch/edoeb/de/home/datenschutz/handel-und-wirtschaft/uebermittlung-ins-ausland.html

Data disclosure to a foreign country must be distinguished from data processing by third parties according to Art. 10a of the DSG. If personal data is processed when using cloud computing, this is normally data processing by third parties within the meaning of Art. 10a of the DSG. Nevertheless, it is possible that the rule above (Art. 6 of the DSG) additionally applies. Further information can be found at: https://www.edoeb.admin.ch/edoeb/de/home/datenschutz/Internet_und_Computer/cloud-computing/erlaeuterungen-zu-cloud-computing.html

4. Requirements for Highly Sensitive Personal Data and Profiling

a. Governing Law (DSG)

The General Principles of Data Processing must be fulfilled. Additionally, the qualified legal consequences respectively stricter regulations for highly sensitive personal data and profiling will occur.

Pursuant to Art. 4 para. 5 of the DSG, increased requirements apply to the processing of highly sensitive personal data or personality profiles. The explicit (written) consent of the person concerned must be obtained. If personality profiles are regularly processed, they must be registered (Art. 11a para. 3 lit. a of the DSG).

Anyone who processes personal data must not unlawfully infringe the personality of the persons concerned, Art. 12 of the DSG. Examples can be found in Art. 12 para. 2 of the DSG.

In the absence of consent, the violation of personality rights is unlawful under Art. 13 para. 1 of the DSG, unless it can be justified by a legitimate private or public interest or by law. According to Art. 13 para. 2 lit. e of the DSG, private Institutions may process personal data for non-personal purposes, in particular in research, planning and statistics. The results must be published in such a way that the persons concerned cannot be identified (anonymized). Researchers must meet certain requirements. They must give preference to the use of anonymized data, and where the purpose of the research does not permit this, they may work with pseudonymized data. However, the research result must be published anonymously in any case. If the requirements of Art. 13 para. 2 lit. e of the DSG are fulfilled, no consent is necessary according to DSG. If the requirements are not fulfilled, another justification as mentioned above, is necessary, e.g., consent.

Compulsory consent must be obtained by researchers if data is subject to professional secrecy or if special laws require it.

Under Art. 14 para. 1 of the DSG there is an obligation to inform the persons concerned. The duty to inform exists only in connection with the collection of highly sensitive personal data in connection with a data collection pursuant to art. 11a of the DSG. Thus, if answers to the questionnaire are processed by a private person that is anonymous to this private person, the duty to inform is not relevant. On the other hand, the duty to inform applies directly to surveys that involve highly sensitive personal data and are therefore not pseudonymized for this private person. Pursuant to Art. 14 para. 2 lit. a-c of the DSG, the person concerned must be informed at least about the owner of the data collection, the purpose of the processing and the

category of the data recipients if data disclosure is envisaged. If the data is not collected from the data subject, the data subject must be informed at the latest when the data is stored or if the data is not stored, on its first disclosure to a third person (Art. 14 para. 3 of the DSG). Further information should be provided on: Type and scope of the data collected; disclosure to third parties / cross-border disclosure of data / address categories; voluntary participation in the project and right of withdrawal at any time; consequences for persons concerned in the event of withdrawal (no significant disadvantages); right of access and right to correction; anonymisation/pseudonymisation of the data; storage of the data (form and duration) and further use; contract with third parties and their duty of confidentiality; possibility of being informed about research results.

According to Art. 14 para. 4 lit. b of the DSG, the duty to provide information does not apply if information is not possible or only with disproportionate effort. In the present case, it does not represent an increased effort to include the most important information in the questionnaire and thereby maintain the transparency of the data processing as far as possible.

b. Law not yet in Force (rev-DSG)

Under the rev-DSG, explicit (written) consent of the person concerned is required in the case of high-risk-profiling (Art. 6 para. 7 of the rev-DSG). The obligation to register a data collection no longer exists under the new law.

Anyone who processes personal data must not unlawfully infringe the personality of the persons concerned, Art. 30 of the rev-DSG. Examples can be found in Art. 30 para. 2 of the rev-DSG.

In the absence of consent, the violation of personality rights is unlawful under Art. 31 para. 1 of the rev-DSG, unless it can be justified by a legitimate private or public interest or by law. According to Art. 31 para. 2 lit. e of the rev-DSG, private Institutions may process personal data for non-personal purposes, in particular in research, planning and statistics. The results must be published in such a way that the persons concerned cannot be identified (anonymized). Researchers must meet certain requirements. They must give preference to the use of anonymized data, and where the purpose of the research does not permit this, they may work with pseudonymized data. However, the research result must be published anonymously in any case. As far as highly sensitive personal data is concerned, the researchers shall disclose the data to third parties in such a way that the data subject cannot be identified; if this is not possible, it must be ensured that the third parties process the data only for non-personal purposes. If the requirements of Art. 31 para. 2 lit. e of the DSG are fulfilled, usually no consent is necessary. If the requirements are not fulfilled, another justification as mentioned above, is necessary, e.g., consent. Compulsory consent must be obtained by researchers if data is subject to professional secrecy or if special laws require it.

Under Art. 19 para. 1 of the rev-DSG there is an obligation to inform the persons concerned. The duty to inform exists only in connection with the collection of personal data. Thus, if answers to the questionnaire are processed by a private person that is anonymous to this private person, the duty to inform is not relevant. On the other hand, the duty to inform applies directly to surveys that involve personal data and are therefore not pseudonymized for this private person. Pursuant to Art. 19 para. 2 lit. a-c of the rev-DSG, the person concerned must be informed at least about the owner and the contact address of the responsible person, the purpose of the processing and the recipient or the category of the data recipients if data disclosure is envisaged. If the data is not collected from person concerned, the person must also be informed about the categories of personal data processed (Art. 19 para. 3 of the DSG). If the personal data will be disclosed to someone in a foreign country, the concerned person must be informed about the name

of the foreign country or the international organ and if necessary, the guarantees according to Art. 16 para. 2 of the rev-DSG or the application of an exception according to Art. 17 of the rev-DSG.

Further information should be provided on: Type and scope of the data collected; disclosure to third parties

/ cross-border disclosure of data / address categories; voluntary participation in the project and right of withdrawal at any time; consequences for persons concerned in the event of withdrawal (no significant disadvantages); right of access and right to correction; anonymisation/pseudonymisation of the data; storage of the data (form and duration) and further use; contract with third parties and their duty of confidentiality; possibility of being informed about research results.

According to Art. 20 para. 2 rev-DSG, the duty to provide information does not apply if the data is not collected from the person concerned and the information of those person is not possible or only possible with disproportionate effort. In the present case, it does not represent an increased effort to include the most important information in the questionnaire and thereby maintain the transparency of the data processing as far as possible.

V. Data Processing by PES

When a private person processes data that it receives directly from PES, the legal requirements set out above apply.

DSG applies in the fields of AVIG and AVG directly in case of processing by federal organs, private unemployment insurance fund, private companies and private persons which are entrusted with public affairs of the federal organs. The processing by cantonal and communal authorities (such as PES, LAM, KAST) or private persons which are entrusted with public affairs of the cantons or communities is regulated by cantonal data protection law, even if they implement federal law. If data of social insurance are processed by a cantonal public body such as the PES, several federal and cantonal acts are relevant. Due to the cantonal autonomy, the cantonal data protection regulations (Canton Zurich: IDG, IDV) and the specific federal data protection regulations (AVIG, AVG, ATSG), as lex specialis, must be observed.

Data processing also includes the disclosure of data (c.f. Art. 3 lit. e and lit. f of the DSG; Art. 5 lit. d and lit. e of the rev-DSG).

1. Cantonal Law of the Canton of Zurich (IDG and IDV)

According to § 18 of the IDG, the public body may disclose data for non-personal purposes, such as research, unless this is excluded by a legal provision. In doing so, the recipient must ensure, in accordance with § 18 para. 2 of the IDG, that the personal data is made anonymous, that no conclusions about data subjects can be drawn from the evaluations and that the original personal data is eliminated as soon as it has been evaluated.

In the case of cross-border disclosure, a distinction must be made between receivers that are subject to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data and those that are not. If the receiver is not subject to the aforesaid Convention, cross-border disclosures are only permitted if the general conditions for data disclosure are met and additionally if adequate protection for the data transfer is guaranteed in the recipient state, if a legal basis permits this to protect certain interests of the persons concerned or predominantly public interests or if adequate contractual security measures are taken by the public body (§ 19 of the IDG). In addition, Art. 22 of the IDV must be observed, e.g.: The list of the Federal Data Protection Commissioner on the status of global data protection indicates which countries guarantee adequate protection (Art. 7 of the DSV), Art. 22 of the IDV.

If the receiver is subject to the aforesaid Convention, the general conditions for data disclosure must be met. The disclosure of personal data for non-personal purposes, namely for research, planning, statistics or expertise independent of persons, in accordance with § 21 of the IDV, requires a written application with details listed in § 21 para. 2 of the IDV. The application must contain the name of the recipient (lit. a), a brief description of the project (lit. b), a description of the personal data required (lit. c), the procedure and type of data processing (lit. d) and details of the measures to be taken to protect the information, in particular with regard to the storage, anonymization and destruction of the personal data (lit. e). According to § 21 para. 3 of the IDV, the public body must issue a written decision, to which it may attach conditions for the protection of personal data.

In advance, the public body must carry out a data protection impact assessment in accordance with §10 of the IDG in conjunction with § 24 of the IDV, assessing the risks to the fundamental rights of the data subjects and submitting it to the public data protection officer for review. Special risks occur when a large number of special personal data are collected (§ 24 para. 1 lit. b IDV) or when the processing involves a large number of persons (§ 24 para. 1 lit. e IDV).

2. Federal Law (ATSG; AVIG, AVG)

Art. 33 of the ATSG, Art. 34 of the AVG lay down a general obligation of confidentiality. According to Art. 33 of the ATSG, the obligation of confidentiality applies towards third parties to persons which are involved in implementation, control, or supervision of the laws according to social insurance. The obligation of confidentiality of Art. 34 of the AVG applies towards third parties concerning details on job seekers, employers and job vacancies to persons which are involved in implementation, control, or supervision of public employment services.

The aim of the social insurance in the context discussed here is to integrate the persons concerned into the labor market and to secure their livelihood. To achieve this goal in the best possible way, close cooperation between the systems is necessary (interinstitutional cooperation = IIZ). This principle is embodied in Art. 85f of the AVIG and 35a of the AVG. However, this mainly concerns complex individual cases and not general cooperation, as the wording of the Art. 85f of the AVIG and Art. 35a of the AVG might suggest.

In principle, a distinction is made between the disclosure of data in response to a written substantiated request or without. Without a written request, a one-time disclosure of a large amount of data pursuant to Art. 97a para. 1 lit. a-e^{bis} of the AVIG and Art. 34a para. 2 of the AVG is only possible to public bodies that pursue specific and important purposes. A private person does not fall under any of these categories in the context discussed here, which is why it can only request data by means of a written application. In accordance with Art. 97a para. 4 of the AVIG and Art. 34a para. 4 of the AVG, a distinction is made between personal and non-personal data when it is passed on. Only the data necessary for the purpose pursued (para. 5) may be disclosed. Personal data may be disclosed to third parties if the disclosure corresponds to an overriding interest. Personal data may only be disclosed if the person concerned has given his or her consent in the individual case, moreover, if such consent is not possible, such disclosure may be assumed to be in the interest of the job seeker under the circumstances. If a private entity that aims to optimize the reintegration of unemployed people into the labour market, this serves a predominant public interest.

An automatic, repeated disclosure of data is not possible in the context discussed here, as this is only permissible under the conditions of Art. 96c of the AVIG, Art. 35 para. 3 AVG, Art. 75b para. 2 ATSG.

Here again there is a duty to inform, which the public body must comply with.

VI. Conclusions

Since pseudonymized data are considered personal data under data protection law, the provisions of data protection law must be observed when a private person (entity) receives pseudonymized data from the PES. However, the transfer of data may be based on the provisions on disclosure for non-personal purposes, namely for research. As described above (V. 1.), a written application is required. The protection of the data during processing for research purposes is sufficiently guaranteed by the prior pseudonymization. The results must be published in such a way that the persons concerned cannot be identified.