NAV

Exit report from sandbox project with NAV

Themes: Legal basis, fairness and explainability

January 2022

Datatilsynet

Table of contents

Note

Technology and the law are constantly evolving, and adjustments and clarifications may have occurred after this report was written.

1. Summary

Objective of the sandbox project

NAV wishes to use machine learning to predict which users on sick leave will require follow-up two months in the future. This will help advisers to make more accurate assessments, which in turn will help NAV, employers and people on sick leave to avoid unnecessary meetings. The objective of this sandbox project was to clarify the lawfulness of using artificial intelligence (AI) in this context, and to research how profiling persons on sick leave can be performed in a fair and transparent manner.

Conclusions

- **Lawfulness.** NAV has a legal basis for using AI as support in making decisions about an individual's need for follow-up and dialogue meetings. There is uncertainty about whether the legal basis permits the use of personal information to develop the algorithm itself.
- **Fairness.** There is an important difference between using information that is already part of the model and utilising new information *not* used in the model, to check for discriminatory outcomes. A conflict arises between protection of privacy and fairness when the method for revealing and combating discrimination involves additional processing of personal information.
- **Transparency**. For the model to provide the desired value, it is essential that NAV advisers trust the algorithm. Insight into and understanding of the mode of operation of the model are important to evaluate the prediction on an independent and secure basis, irrespective of whether the final decision is to follow the recommendation of the prediction, or not.

The road ahead

The work on NAV's prediction model for sickness absence has highlighted a major and important challenge to public authorities seeking to utilise artificial intelligence: The laws that permit the processing of personal information are seldom formulated in a way that permits personal information to be used for machine learning in the development of artificial intelligence.

It is important that legislators facilitate future developments of AI in the public sector within a responsible framework. If NAV is to develop the model further, it will be necessary to have a clear and explicit supplementary legal basis, founded in legislation. A legislative process, with the associated consultations and reports, will help to ensure a democratic foundation for the development and use of artificial intelligence in public administration.

NAV's systematic work on the development of a model that meets the requirements for fairness and explainability shows that public sector organisations can serve as driving forces for responsible development in the field of AI.

2. About the project

NAV has a hypothesis that there are too many unnecessary meetings, and that the meetings steal time from employers, sickness certifiers (e.g. doctors), the person on sick leave and NAV's own advisers. This was the motivation for establishing the AI project which was to address predictions of the duration of sickness absence.

These meetings represent one of several legally-decreed waypoints in NAV's following up of sickness absence. Within seven weeks of sickness absence, the person on sick leave and the employer must hold a dialogue meeting. After eight weeks, NAV is required to check whether the person on sick leave is in activity, or whether he/she can be exempted from the activity requirement. And before sickness absence passes 26 weeks, NAV is required to evaluate the need for a further dialogue meeting with the person on sick leave, the employer and the sickness certifier. As early as in week 17, NAV must establish whether a new dialogue meeting will be necessary, i.e. whether the person on sick leave will be declared fit for work within week 26 or not. At each of the waypoints, NAV evaluates which type of follow-up the person on sick leave requires.

This project is based on the waypoint after 17 weeks of sickness absence and the decision to convene dialogue meeting 2. By using machine learning to predict the length of sickness absence, NAV wishes to support the adviser's decision regarding the necessity of holding dialogue meeting 2. The hope is to:

- Reduce time spent on assessing the need for a dialogue meeting for NAV personnel working with sickness absence.
- Save time for the parties involved in sickness absence, by a greater degree avoiding the need to convene unnecessary dialogue meetings.
- Provide better follow-up for persons on sick leave who require a dialogue meeting, by concentrating efforts on those who actually need it.

Presentation: This screenshot shows an early example from NAV of how a recommendation from the system could be presented to the adviser who will take the final decision on whether a dialogue meeting should be convened or not. The response is founded on three factors that indicate a longer duration and three that indicate a shorter duration. The NAV adviser also receives information about how employers and persons on sick leave evaluate the need for a dialogue meeting.

3. Sandbox objectives

NAV came to the sandbox with an AI tool more or less ready for use and having conducted thorough legal evaluations. This paved the way for the sandbox project to be more of a quality assurance of work performed than a joint innovation process. The overarching objective for the sandbox project was to help to create practices for how NAV can ensure control and responsibility over the course of an AI development process.

The project will:

- Clarify NAV's scope for utilising AI where doing so is legal and responsible.
- Shorten the path from idea to implemented AI in other areas in NAV and for other enterprises that see the potential in similar AI applications.

In other words, the project can be useful and have transferable value for NAV in general, but also for other enterprises, especially in the public sector.

In the sandbox we have discussed problem issues associated with legal basis, i.e. whether NAV has the right to use machine learning as planned. In addition, we have discussed the fairness of the model, including how discrimination can be revealed and counteracted in such a model . Finally, we have looked at the requirements for a meaningful explanation of the model, at both a system and individual level.

4. Evaluations and conclusions

4.1 Problem issues

The work in the sandbox has revolved around three problem issues associated with AI: legal basis, fairness and explainability. In the first section we examine the legal challenges associated with NAV's legal basis, i.e. the legality of processing personal information to develop and use a machine learning model. The second section is an assessment of NAV's approach to the requirement that this type of model must be deemed to be fair, and in the final section we discuss issues of transparency and how the function and outcomes of the model can be explained.

4.2. Legal basis

4.2.1 Introduction

Public authorities process a great deal of personal information, and this processing is often sanctioned by laws or regulations. This means that authorities avoid having to obtain consent from or draw up an agreement with every single person whose personal information is processed, but rather receive permission–a legal basis–to do so via specific laws and regulations.

New technology can lead to new ways of processing personal information that were not taken into account when the laws that govern NAV's processing of personal information were made. The development and use of artificial intelligence requires the processing of large amounts of data–often personal information–which is compiled and analysed on a scale that is not possible by other means.

Clear legal authority is required for the development of artificial intelligence in the public sphere. Efforts are made to safeguard this aspect via requirements for clear statutory authority in the General Data Protection Regulation's (GDPR) articles 5, 6 and 9, Article 102 of the Norwegian Constitution, and Article 8 of the European Convention on Human Rights, in addition to case law associated with these provisions.

4.2.2 In general concerning legal basis

The legal basis which is most appropriate to consider for NAV's prediction model is Article 6 (1) (e). This provision states that personal information can be processed if processing is necessary in the exercise of official authority vested in the data controller. In addition, legal authority is required pursuant to Article 9, if special categories of personal data are processed. NAV's prediction model does this, particularly in regard to health information. NAV therefore applies Article 9 (2) (b), which provides a basis for processing special categories of personal information in the exercise of rights and obligations in social security law.

Both Article 6 (3) and Article 9 (2) (b) require a supplementary legal basis in national law. No explicit or specific statutory authority is required for the exact processing. The *purpose* of processing must be founded in national law *or* it must be necessary for the exercise of official authority.[1]

The legal authority must nevertheless be sufficiently clear to ensure predictability for those affected and prevent arbitrariness in the exercise of official authority.[2] This means that the law must define how the information can be used and set limits on the way in which the authorities are able to use the information. A specific evaluation must be made as to whether the provision is adequate for the processing in question. The more intrusive the processing, the clearer the statutory authority must be.

4.2.3 NAV's supplementary legal basis

NAV expands on the supplementary legal basis in the National Insurance Act Section 8-7 a, seen in conjunction with Section 21-4 of the same Act and the Public Administration Act Section 17. In addition, NAV has the authority to process personal information in the Act relating to the Labour and Welfare Administration (NAV Act) Section 4a first paragraph.

The National Insurance Act Section 8-7a regulates some of NAV's obligations to follow up persons on sick leave. Section 8-7a second paragraph contains regulations pertaining to dialogue meeting 2 that must be held in week 26 of sickness absence–except "when such a meeting is assumed to be clearly unnecessary".

The regulation must be viewed in context with the general regulation in the Act's Section 21-4. This gives NAV general legal authority to collect information in order to exercise its duties. As an administrative agency, NAV is also covered by the general provision in the Public Administration Act Section 17. This requires that "the administrative agency shall ensure that the case is clarified as thoroughly as possible before any administrative decision is made".

The development phase

It is natural to split the question of legal basis in two, based on the two main phases in an AI project; the development phase and the application phase. The two phases utilise personal information in different ways.

In the development phase, NAV uses a large amount of historical data–personal information concerning those previously on sick leave–from numerous registered individuals, to train a model that will predict the duration of sickness absence of other persons in the future. In the development phase, no personal information is used from people who will be followed up in the future.

The question will then be whether the relevant provisions of the Act (the National Insurance Act Section 8-7a and Section 21-4) that provide statutory authority to process personal information to evaluate whether it is clearly unnecessary to hold a dialogue meeting 2 in a specific case, also allow for the processing of personal information in connection with the development of an AI tool for use in case processing.

A natural interpretation of the wording indicates that these provisions do not provide such statutory authority. Compared with the current evaluations of dialogue meeting 2, the development of a prediction model will process a far larger volume of personal information belonging to persons that are no longer on sick leave. An important aspect is also

---

[1] GDPR Article 6 (3)

[2] cf. The Norwegian Constitution's Article 102 and ECHR Article 8.

that this information to a major degree will be special categories of personal information such as diagnoses, sickness absence history and information from the free text field in the medical certificate.

The invasive nature of the processing in the development phase also indicates that clear and explicit statutory authority must be required. It is doubtful whether the Social Security Act Section 8-7 (a), cf. Section 21-4 and the Public Administration Act Section 17 are sufficiently specific to represent a clear and explicit supplementary legal basis according to Article 6 (1) (e) and Article 9 (2) (b). It is not sufficiently evident in the statutes applied by NAV as supplementary legal basis, that information from previous users can be used in the development of artificial intelligence.

The application phase

For the application phase, NAV has carried out a thorough evaluation of the supplementary legal basis for use of the prediction model as support in decision-making. The evaluation expands on the supplementary legal basis in the National Insurance Act Section 8-7a, seen in conjunction with Section 21-4 of the same Act and the Public Administration Act Section 17. In addition, NAV has the statutory authority to process personal information in the NAV Act Section 4a first paragraph.

NAV has found that no special statutory authority is required for the method itself, including the use of the prediction model; however, an evaluation must be made as to whether the method is proportional, in order to determine whether the person on sick leave should be called in to dialogue meeting 2, or not.

Decisive for this evaluation is whether the use of the prediction model can be considered to be more intrusive for the user. Moreover, an evaluation has been made as to whether the planned use of personal information, both in terms of volume and how the information is used, can be considered necessary in order to comply with the requirements of the Act.

NAV has concluded that the processing of personal information is both proportional and necessary in order to achieve the objective and will therefore have a supplementary legal basis for using the prediction model as support for decision-making in the application phase itself, provided it has a legal basis for the development.

4.2.4 Conclusion concerning legal basis

In our assessment, NAV may have a legal basis for processing personal information in using AI in this context. However, it is doubtful whether the legal basis stated by NAV can represent a legal basis for using personal information to develop a prediction model–even though the model will later be able to contribute to better follow-up of persons on sick leave. A legal basis for development and the associated processing of personal information is a prerequisite for NAV to use the prediction model as support for decision-making in decisions that concern whether dialogue meeting 2 shall be held.

It could be argued that there are societal benefits in NAV developing artificial intelligence to improve and increase the efficiency of its work. At the same time, the development of artificial intelligence is a process that challenges several important personal data protection principles. In order to safeguard the rights of registered persons, clear and explicit statutory authority in laws or regulations for this type of development will be necessary. A legal process, with the associated consultations and reports, will help to ensure a democratic foundation for the development and use of artificial intelligence in public administration.

The conclusion above is based on discussions held between the Data Protection Authority and NAV in the sandbox project and is therefore for guidance only, and not a decision by the Data Protection Authority. The responsibility for evaluating the legal basis for the relevant processing lies with NAV as the data controller.

4.2.5 Automated decision-making processes

Even if processing is legal, GDPR gives the registered person the right not to be a subject of automated, individual decision-making, i.e. decisions taken without human intervention, if the processing produces legal effects concerning

him or her or similarly significantly affects him or her.[3] The human involvement must be genuine and not fictive or illusory.[4]

If the prediction model is used only as support for decision-making, the prediction concerning the length of sickness absence will be one of multiple elements in the NAV adviser's assessment of whether the person on sick leave will be called in to a dialogue meeting. In such a case, the human evaluation will mean that the processing is not defined as fully automated. However, it could be reasoned that the decision in practice *is* fully automated. The advisers' workload and knowledge of the algorithm, and the perceived and actual accuracy of the predictions, will influence the risk that the person in the loop–the adviser–will accept any results generated by the prediction model more or less without thinking.

There are various measures that can mitigate this risk. Good routines and training of advisers will be fundamental. The information they receive in connection with the use of the tool must be comprehensible and allow them to evaluate the prediction against other aspects. In addition, routines must be introduced to reveal whether decisions are fully automated.

Admittedly, in the longer term, NAV wishes to fully automate the process of convening dialogue meeting 2. There are exceptions to the prohibition against fully automated decisions; however, this assumes that the decision does not "produce legal effects concerning him or her or similarly significantly affects him or her".[5] So, does the model produce these effects?

NAV's prediction model, which will estimate the length of sickness absence, involves profiling[6] and is an automated process. If the model in actuality is used as *support* for decision-making, the decision itself is not automated. It is the decision on whether or not to convene a dialogue meeting that has the potential to produce legal effects for or similarly significantly affects the registered person–not the prediction itself.

The question is thus whether the invitation to dialogue meeting 2 has legal effects or similarly affects the user. A decision has 'legal effects' if it affects the person's legal rights, such as the right to vote or has contract law-related effects. An invitation to a dialogue meeting is not encompassed by this. Thus what remains is to evaluate whether the decision concerning a dialogue meeting significantly affects the user similar to a legal effect.

The answer is yes, if the decision has the potential to affect the individual's circumstances, conduct or choices, has a long-term or permanent effect, or leads to exclusion or discrimination.[7] Decisions that influence the financial circumstances of a person, such as access to health services, can be considered an effect similar to a legal effect.

A decision concerning dialogue meeting 2 is not an individual decision; however, an argument can be made that it 'significantly affects', and in a fully automated version will fall within the scope of Article 22. In the case of a public sector activity, the scope of Article 22 takes in more than solely individual decisions, a view that is supported by the preparatory works for the new Public Administration Act. What exactly falls within 'legal effect' or 'similarly significantly affects' must be evaluated substantively, based on the consequences the decision has for the registered person. For NAV's prediction model, a distinction might be envisaged between a situation in which a dialogue meeting 2 is convened, and one in which no meeting is convened.

If the person on sick leave is not convened to a dialogue meeting, no obligation arises for the person concerned. However, the person on sick leave retains the right to request a dialogue meeting. In such situations, the decision will have a less invasive effect on the registered person, as long as the possibility of requesting a dialogue meeting remains genuine. At the same time, dialogue meeting 2 is designed to help the person on sick leave to return to work. Not all persons on sick leave will have the resources to assert the right to request a dialogue meeting. This can perhaps be partially safeguarded by providing good information to registered persons.

---

[3] GDPR Article 22 (1)

[4] Article 29 Data Protection Working Party – "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" p. 20-21.

[5] GDPR Article 22 (1)

[6] GDPR Article 4 (4)

[7] Article 29 Data Protection Working Party – "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" p. 21-22

In situations where a person on sick leave is called in to dialogue meeting 2–which is the principal rule according to the National Insurance Act Section 8-7a–an obligation will arise for the person on sick leave to attend the meeting. Failure to comply with this obligation will–ultimately–lead to the termination of sick pay. In such cases, the obligation to attend dialogue meeting 2 will potentially have a major effect on the person on sick leave and may fall within the scope of Article 22.

Summarised briefly, decisions to convene a dialogue meeting may reach the threshold in Article 22, which triggers a prohibition. Decisions to not convene a meeting may fall short of the threshold, provided that the right of the person on sick leave to request a dialogue meeting is genuine. Whether it is possible in practice to separate the decisions in this way, will be a matter for NAV to consider.

4.3 Fairness

4.3.1 Introduction

When we have discussed fairness in this sandbox project, we have taken as our starting point three main principles for responsible artificial intelligence: lawful, ethical and robust. These main principles are based on the "Ethics guidelines for trustworthy AI", prepared by an expert group appointed by the European Commission. The same principles are also reflected in the National strategy for artificial intelligence.

In its guidelines for integrated personal data protection, the European Data Protection Board (EDPB) lists several aspects that are included in the fairness principle, among them non-discrimination, the expectations of the registered person, the process' broader ethical issues and respect for rights and freedoms.[8] The fairness principle contains several more aspects in addition to non-discrimination. Discrimination in algorithms is a familiar challenge in artificial intelligence and the sandbox work has therefore had focus on this. A major public body such as NAV has a particular responsibility to be aware of the imbalance in power manifested in interactions between users and NAV's systems.

The fairness principle is a central element in other legislations, among them various human rights provisions and the Equality and Anti-Discrimination Act. These statutes could also have a bearing on the question of fairness and their requirements might also be more or less stringent than the provisions of data protection legislation.

4.3.3 NAV's model

NAV has developed methods that enable the fairness of the model to be tested. The main focus has been on the *bias* of the model, i.e. potential biases in data collection, the choice of variables, model selection or implementation and how these are manifested in skewed outcomes and possible discriminatory effects. Machine learning models will inevitably treat persons differently, as the desire for a more user-adapted differentiation often motivates the development of a machine learning model. Avoiding arbitrary discrimination was one of the central themes in this sandbox project. NAV does not wish to reproduce or strengthen existing biases, but risks doing exactly this if bias is not analysed and addressed.

To support this analysis, NAV wishes to evaluate what a fair algorithm outcome involves in a legal sense. Developing a machine learning model that addresses several legal requirements[9] for fairness involves operationalising legal and ethical principles. To evaluate whether the model is consistent with the concepts of fairness in the legislation, it is useful to clarify how the model will function when it is put into production. What kind of outcome might, for example, groups with special requirements for protection against unfair discrimination, expect to see?

NAV itself points out that this kind of analysis does not cover all the ways in which the processing of personal information can be unfair or discriminatory. However, focusing on the outcome (regardless of issues associated with, for example, data collection, processing and practical model application), facilitates a discussion of how the fairness concept must be interpreted and how it can be operationalised.

---

[8] Guidelines 4/2019 on Article 25 Data Protection by Design and by Default | European Data Protection Board (uropa.eu)

[9] In addition to GDPR, NAV must comply with regulations in the Public Administration Act, NAV Act and the Equality and Anti-Discrimination Act.

In the operationalising of the fairness evaluation, NAV has elected to focus on outcome fairness, i.e. whether the outcome of the model is distributed fairly across various groups. The evaluation is comparative, i.e. it examines how various groups that are part of the model are processed compared to each other, rather than measured against a standard or norm. NAV has also concluded that model errors resulting in the convening of unnecessary dialogue meetings are less serious than the contrary. One of the starting points for evaluating fairness in the prediction model is the National Insurance Act Section 8-7a, which instructs NAV to hold a dialogue meeting "except where such a meeting is considered to be clearly unnecessary". This type of requirement suggests that in cases of doubt, one dialogue meeting too many should be held rather than one too few.

From a personal privacy perspective, fairness must be evaluated both at a group level and an individual level. The model may also conflict with the fairness principle if only individuals are negatively affected to a significant degree and not solely if group discrimination occurs–for example, if there are rare combinations of factors that lead to very negative effects for the registered person.

Moreover, one can envisage that the prediction of the length of sickness absence for certain groups will be erroneous in terms of evaluating when a dialogue meeting should be convened. For example, this could apply in circumstances where the future length of sickness absence is not the best evaluation factor for a decision as to whether a dialogue meeting is 'clearly unnecessary' and where based on a fairness perspective, such case histories might need to be identified to avoid this kind of imbalance. For example, one can envisage situations in which several pregnant women have long periods of sickness absence where it is still clearly unnecessary to hold dialogue meeting 2. The same might apply in the case of partially disabled persons who will be on sick leave for one year from confirmation of their residual work ability percentage, with the future objective of full disability pension.

4.3.4 Other aspects

The model that has been discussed in the sandbox is a decision-making support system. This means that the prediction will be one of multiple information elements that form part of the adviser's evaluation. If a fully automated decision is made, a new fairness evaluation must be carried out. At the same time, it is important to remember that humans also discriminate. It is therefore by no means certain that the actual outcome for the registered person will be made fairer by the presence of a person in the loop. Nevertheless, it can be experienced as more intrusive to be unfairly treated by a machine learning model than by an adviser. In addition, any unfair practices exhibited by the model will scale in a completely different way than the current system and lead to systematised unfairness. A new evaluation of the registered person's rightful/reasonable expectations of processing will likely become even more important in a fully-automated model. This also applies to revision and control of the algorithms.

4.3.5 Who has a right to special protection?

The method that has been chosen to evaluate the machine learning model's outcome fairness, requires NAV to define which groups should be evaluated against each other. As a starting point, there are an arbitrary number of user groups that can be defined based on the user mass that forms the data basis for training the model. Which groups should be included in a fairness evaluation of the model is a question with several different social, historical and societal dimensions. NAV exists for everyone; however, it is neither technically nor practically possible to perform an evaluation for all group identities in Norwegian society. Who has the right to or a particular need for protection against biased model outcomes, is thereby a key issue.

A large part of this question falls more naturally within the realm of the Equality and Anti-Discrimination Act, and as part of the sandbox work, we invited the Equality and Anti-Discrimination Ombud to discuss these issues.

In principle, the groups that NAV utilises–including gender, age and diagnoses–are well founded in the Equality and Anti-Discrimination Act. It is possible, that in addition to the defined groups, complex discrimination bases will also occur, in which a combination of group identification generates a particularly biased result. There are also other vulnerable groups that it might be be useful to include, such as persons dependent on intoxicants, persons with care duties and persons with a low economic status.

A central question in connection with discrimination is whether this type of prediction model differentiates in such a way that it can be called discrimination. As the specific model being evaluated is concerned with the length of sickness absence and deals with whether a dialogue meeting should be held or not, this discrimination threshold will not necessarily be reached. The situation is likely to be different in the case of a model for other types of benefits with greater consequences for the registered person.

4.3.6 The conflict between personal privacy and fairness

In all machine learning models, a tension can arise between the model's mode of operation and several personal privacy principles. In the NAV project, this type of tension arises when NAV must fulfil its obligation to check whether the model functions in a biased manner or discriminates. In principle, personal information needs to be processed to both uncover and correct outcome bias. Admittedly, uncovering bias in the model's outcomes can be done regardless of whether group identification is part of the model. However, to carry out an evaluation of the model's outcome, group identification must be used. Finally, it can be possible to comply with other requirements for information fairness without this type of processing of personal information. These questions are key for developers of responsible AI, and the EU's proposal for new AI legislation touches on these questions.[10]

NAV's services must be accessible to the entire population, and NAV must therefore navigate the tension between personal privacy and biased outcomes in each model that is developed. In addition, there is a major overlap between groups that personal privacy regulations define as vulnerable and groups that are covered by the Equality and Anti-Discrimination Act.

When considering the fairness of the model, there is, viewed from a personal privacy standpoint, a difference between utilising information that is already part of the model and utilising new information that in principle is not used in the model, but that is incorporated in the analysis in order to check for discriminatory outcomes. A tension arises between protection of privacy and fairness when the method for uncovering and combating discrimination involves complex processing of special categories of personal information. Information that is already included in the algorithm forms part of the decision-making basis in the follow-up of sickness absence. Entirely new information, on the other hand, requires a new assessment of legality. In addition, it is likely that the registered person has a rightful expectation that information that is irrelevant to an evaluation of whether a dialogue meeting should be held, will not be utilised in the model. One can envisage that the use of anonymised or synthetic data could offer a solution that could uncover outcome bias, whilst at the same time safeguarding personal privacy. Fully anonymised data is not considered to be personal information and therefore personal privacy legislation does not apply. However, this is something that we have not discussed extensively in the sandbox.

There is not necessarily an adequate answer to the question regarding the conflict between personal privacy and fairness in a machine learning model. Equally, however, it is a central part of the discussion about and the work towards responsible artificial intelligence.

4.3.7 Tolerance for discrimination?

The objective of the prediction model is to support a type of differentiation: to assist the adviser in the evaluation of who should be offered a dialogue meeting. The central issue will therefore not be whether the model differentiates, but rather whether it differentiates correctly, and that the differentiation is not unreasonable and/or discriminatory.

The model, intended to predict sickness absence, is in practice an automated contribution to the many thousands of evaluations that are made each day by NAV advisers. There are methods to evaluate how fair the outcomes of a prediction model will be, making it possible to quantify fairness in a way that is impossible at present. Consequently, the use of a machine learning model allows discriminatory outcomes to be revealed that at present are hidden behind the daily workflow at Norway's NAV offices. This opens the way for a difficult discussion concerning how much unfairness should be accepted and what the approach to this type of quantified unfairness should be. No-one would argue that all NAV clients are treated fairly; however, a machine learning model will mercilessly quantify the rate of unfairness.

It is unlikely to be possible to set a percentage rate for an accepted degree of tolerance of discrimination, given the way in which the Equality and Anti-Discrimination Act is set out. Which type of practices lead to the actual greatest discrimination effect is equally something that Norwegian and European equality and anti-discrimination authorities must consider when dealing with this type of technology.

---

[10] The European Commission's proposal for a new regulation pertaining to artificial intelligence Article 10 no. 5. Extract from EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu).

4.4 How to explain the use of artificial intelligence?

4.4.1 Transparency and explainability

Transparency is a fundamental principle of GDPR.[11] In addition to being a prerequisite for uncovering errors, discriminatory treatment or other problematic issues, it contributes to increased confidence and places the individual in a position to be able to assert their rights and safeguard their interests. In connection with AI, the concept of 'explainability' is often used, which addresses AI-specific problem issues associated with transparency, which can be said to be a part of the concretising of the principle of transparency. Traditionally, transparency has been about showing how different personal information is used; however, the use of AI requires other methods that can explain complex models in an understandable way.

Explainability is an interesting topic, both because explaining complex systems can be challenging and because the way in which the requirement for transparency is to be implemented in practice will vary from solution to solution. In addition, machine learning models permit explanations that appear radically different than those we are used to, generally based on advanced mathematical and statistical models. This opens the way for an important trade-off between a more correct, technical explanation or a less correct, but more understandable explanation.

In this section of the report, we share evaluations and conclusions from the discussions we held concerning transparency and explainability in NAV's solution for predicting the length of sickness absences. Advisers at NAV offices and the person on sick leave as an individual are the two most central target groups for explanation in this case.

4.4.2 Transparency requirement

Regardless of whether you use artificial intelligence or not there are certain requirements for transparency if you process personal data.[12] Briefly summarised these are:

- The registered person must receive information about how the information will be used, whether the information is obtained from the registered person themselves or from others.[13]
- The information must be easily accessible, for example on a web page, and must be written in clear and understandable language.[14]
- The registered person has the right to know whether information is processed about him/her and the right of access to his/her own information.[15]
- It is a fundamental requirement that all processing of personal information shall be carried out in a transparent manner. This means that it is a requirement to assess which transparency initiatives are required in order for the registered person to be able to safeguard his own rights.[16]

In the first bullet point, there is a requirement for information to be provided about *how* the information will be used. This includes the contact details of the data controller (in this case NAV), the purpose of the processing and which categories of personal data will be processed. This is information that is typically provided in the privacy statement.

In regard to artificial intelligence, it may be useful to note the requirement that the underlying logic of the algorithm must be explained. There is a specific requirement to provide "relevant information concerning the underlying logic and the significance and the envisaged consequences of such processing ".[17] It is not necessarily self-evident as to how these requirements should be interpreted. One should strive to ensure that the information given is meaningful, rather than using complicated explanation models based on advanced mathematics and statistics.[18] It is also highlighted in the GDPR's foreword that technological complexity makes transparency additionally important.[19] The expected consequences should also be exemplified, for example with the help of visualisation of previous outcomes.

---

[11] GDPR Article 5 (1) (a) and Recital 58.
[12] Detailed information concerning the requirement for transparency in AI solutions can be found in the report Artificial intelligence and personal privacy (2018)
[13] GDPR articles 13 and 14
[14] GDPR Article 12
[15] GDPR Article 15
[16] GDPR Article 5
[17] ICO, GDPR articles 13 and 14
[18] Article 29 Data Protection Working Party – "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" p. 31
[19] GDPR Recital 58

It is specified that this should in any event be done in cases of automated decision-making or profiling according to Article 22. Whether information about the logic must be provided if there are no automatic decisions or profiling must be considered from case to case, based on whether it is necessary for the purpose of securing fair and transparent processing.

4.4.3 Automatic, or not?

If processing can be categorised as automated decision-making or profiling according to Article 22, there are additional requirements for transparency.[20] You have, among other rights, the right to know whether you are the subject of automated decision-making, including profiling. There is also a specific requirement that the individual is provided with relevant information concerning the underlying logic and the significance and the envisaged consequences of such processing, as stated above.

However–do you have the right to an individual explanation about how the algorithm reached the decision? The wording of the legislation itself does not state this; however, the recitals state that the registered person has the right to an explanation of how the model arrived at the result, i.e. how the information has been weighted and evaluated in the specific instances, if one falls within the scope of Article 22.[21] The recitals also state that the registered person should "be informed of the existence of profiling and the consequences of such profiling".[22] The recitals themselves are not legally binding and do not of themselves grant the right to an individual explanation.

The requirement for transparency does not necessarily mean that the source code must be made available; however, the explanation must enable the registered person to understand why a decision was what it was. This applies where the decision falls within the scope of Article 22 concerning automated individual decision-making. One can also imagine circumstances where the fairness and transparency principle places higher demands on explanation, for example in profiling that does not comply with the conditions of Article 22, but where sound reasons indicate that the registered person should receive such information.

A meaningful explanation will depend not only on technical and legal requirements, but also linguistic and design-related considerations. An evaluation must also be performed of which target group the explanation is aimed at–something that may mean a difference for advisers and users. The practical application itself of the explanation model in advisers' daily working day may also mean that trust and the sense that the adviser is receiving a meaningful explanation may vary, in that explanations provided appear to be standardised and therefore offer little guidance over time. Social factors such as trust in the enterprise, the significance of the decision and trust in AI systems in general may also influence the experience of a meaningful explanation.

A key question for NAV has been whether the prediction model for the length of sickness absence is an automated decision and therefore invokes these extra requirements, or not. In this case, there is little doubt that the prediction model does not constitute fully automated processing. The prediction will be one of several information elements that an adviser must evaluate before a decision is made.

However, there are reasons for information to be provided about the logic and mode of operation in models that are not fully automated. The prediction model carries out profiling,[23] and a meaningful explanation contributes to building trust and is an expression of responsibility. Additionally, a meaningful explanation will put an adviser in a better position to evaluate how much weight they shall place on the recommendation generated by the algorithm.

Regardless of whether this concerns fully automated decision-making or not, the data processor is required to provide sufficient information so that the user has the information necessary to safeguard his/her rights. NAV's central role in public administration leads to an asymmetric power relationship between user and government body, which is also an argument in favour of striving for as meaningful an explanation as possible, despite the fact that the model is not fully automated.[24]

---

[20] GDPR Article 13(2)(f) and 14(2)(g), see also "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" and Article 4(4).
[21] GDPR Recital 71
[22] GDPR Recital 60
[23] GDPR article 4 (4)
[24] Artificial intelligence and personal privacy | Norwegian Data Protection Authority

4.4.4 Can we trust the algorithm?

Good explanations of the algorithm and its predictions increase trust in the systems on the part of its users, which is fundamental in achieving the desired value. The several thousand NAV employees that work in user guidance therefore play a decisive role.

The system that predicts the length of sickness absence is a decision-making support system; but what happens if the system in practical use becomes a decision-making system? A NAV adviser reviews many cases in the course of a normal working day. If it appears that the algorithm provides consistently sound recommendations, it can indeed be tempting to always follow these recommendations. The adviser might perhaps believe that the algorithm holds so much data that it knows best, so why should the recommendation not be followed? How easy is it for a recently hired employee not to follow the recommendation of the algorithm?

Or, what if the adviser believes that the algorithm is generating strange recommendations and does not trust them? A consequence of this would be that the adviser does not consistently use this as decision-making support. This will also be unfortunate, as the entire intention of the solution is to help advisers to make good decisions, so that invitations to dialogue meetings are more often appropriate. Ideally, this type of model will reduce arbitrary variations among advisers and lead to more uniform practice, in addition to reducing costs.

In the sandbox we discussed the risk that an adviser will rely too much or too little on the decision-making support system and how to ensure that the system is experienced as providing genuine support for the adviser and is used in a sound and correct manner. That an adviser must receive proper training and instruction in how the algorithm functions and is used, and a meaningful explanation in individual cases is important in order to reduce the risk of “automation by stealth”, or that it is not included in the evaluation at all. When a NAV adviser understands the construction of the model, its mode of operation and behaviour, it will be simpler to evaluate the prediction on an independent and secure basis. Additionally, the explanation can help the adviser to uncover discrimination, undesirable differentiation and

Explanation of the model:

This screenshot shows an example of how a general explanation of the operation of the model can appear for an adviser using the system.

It includes the data used in the model, explains that the model is based on similar data from everyone who has previously been on sick leave for at least 17 weeks, and it explains that the supervisor will now see the three most important factors that increase the probability, and the three most important which reduces it.

errors. In such a case, an explanation associated with individual decision-making will be supplemented with information associated with the outcome for certain naturally comparable groups.

4.4.5 What does a meaningful explanation look like?

An issue we have discussed in the sandbox is how a meaningful explanation would look in practice in NAV's case. The target group for transparency in the solution are those on sick leave and NAV advisers. The explanations are both global, i.e. at a system level, and local outcome explanations. The two different levels will therefore have partially different target groups, and differing requirements apply as to how they are organised.

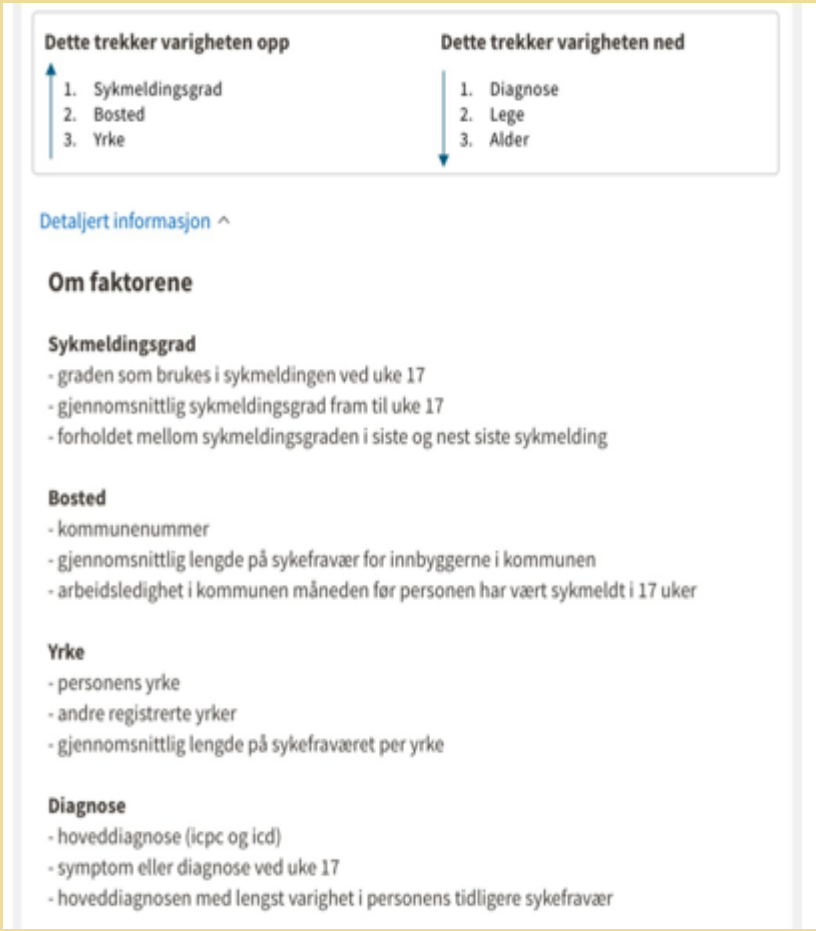
NAV wishes to provide information in advance of processing that the user has the right to protest entirely against that a prediction will be made based on profiling at all. They also wish to inform about how the model is constructed and which variables are incorporated. NAV also considers informing the individual user about the most important factors that extend the predicted sickness absence period and the most important factors that reduce it.

A meaningful explanation does not only depend on technical and legal requirements, but also linguistic and design-related considerations. The explanation must be adapted to the specific target group. For example, NAV advisers require explanations that can be used in practice in a hectic working day. NAV must therefore balance and make a trade-off between depth and simplification that make it possible to use the explanation. Moreover, the explanation must also be integrated with other information to which the adviser has access. A specific example is that NAV cannot present information about how 100 variables have contributed to a prediction. NAV must group these together and make a selection. Particular vigilance is required if the explanation is aimed at children or vulnerable groups. NAV's model may include several special categories of personal information about vulnerable groups, and NAV must therefore consider adapting language, content and form based on this.

Information concerning data:

This is how NAV envisages that their advisors can be presented with which data will be used and applied as a basis, in which manner, in the model.

Here, the factors that increase or decrease the probability of permanent sick leave are presented, together with a key word explanation of what is included in the factors.



In regard to the advisers, NAV plans to explain how the model works in general terms and to describe how the results produced by the model should be used in case processing routines. In addition, advisers will receive explanations at an individual case level as well as information components the model has learned from, as part of the information basis to make the final decision as to whether or not a user will be invited to a dialogue meeting. The prediction will constitute one of several elements that are available to the adviser, including the information on which an adviser bases a decision at present.

In addition to the two main target groups (users and advisers) discussed here, NAV has identified the business side/management, those responsible for the model and supervisory authorities as other target groups that will have a need for, and the right to, an explanation of how the algorithm functions.

NAV wishes to shoulder its share of the responsibility in regard to transparency around the use of algorithms. One possible initiative being discussed is to provide general information on how NAV wishes to utilise artificial intelligence. NAV also seeks to contribute to the broad dissemination of information and an informed debate about the use of artificial intelligence through the media. A final measure is to inform and involve a user panel in advance of, and during, the development of services based on artificial intelligence.

5. The road ahead

The work on NAV's prediction model for sickness absence has highlighted a major and important challenge to public authorities seeking to utilise artificial intelligence: The laws that permit the processing of personal information are seldom formulated in a way that permits personal information to be used for machine learning in the development of artificial intelligence.

It is important that legislators facilitate future developments of AI in the public sector within a responsible framework. If NAV is to develop the model further, it will be necessary to have a clear and explicit supplementary legal basis, founded in legislation. A legislative process, with the associated consultations and reports, will help to ensure a democratic foundation for the development and use of artificial intelligence in public administration.

NAV's systematic work on the development of a model that meets the requirements for fairness and explainability shows that public sector organisations can serve as driving forces for responsible development in the field of AI.

Norwegian Data Protection Authority regulatory sandbox for responsible artificial intelligence

Office address:
Trelastgata 3, Oslo

Postal address:
PB 458 Sentrum
0105 Oslo, Norway

sandkasse@datatilsynet.no
Telephone: +47 22 39 69 00

datatilsynet.no/sandkasse
personvernbloggen.no