Meaningful Human Intervention

A tool for shaping and implementing meaningful human intervention

Consultation document

March 2025

Preface

Automated decision-making is used a wide range of sectors. Crucial to the implementation of algorithms for automated individual decision-making and surrounding processes are a number of concepts from the General Data Protection Regulation (GDPR) and the Law Enforcement Directive (LED), including a decision that is based "solely on an automated processing". [1] In other words: where there is no meaningful human intervention.

In the Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (hereinafter: the guidelines), the European Data Protection Board (EDPB) has clarified what is meant by this: "Solely automated decision-making is the ability to make decisions by technological means without human involvement" and "the controller cannot avoid the Article 22 provisions by fabricating human involvement." [2] Such human involvement is subject to requirements: "To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture." [3]

What makes human intervention meaningful has not yet been fully defined. Researchers of the Brussels Privacy Hub looked at the ways in which people would have to check algorithms, and would have to take action if something goes wrong. They concluded that: "Determining what could be meant precisely by meaningful is indeed an even more complicated -but necessary- task. The scarce precedents in the CJUE and national courts do not make it any easier." [4] ***Meaningful Human Intervention: a tool for shaping and implementing meaningful human intervention*** offers tools to data protection officers (DPOs), controllers and other parties involved to determine when human intervention could be meaningful. The guidelines, scientific literature, some court decisions and knowledge of AP employees who have dealt with automated decision-making form the basis for this document. In addition, this document was discussed with other European data protection authorities. The AI Act (Regulation (EU) 2024/1689) also offers clarification. In Article 14, in which the requirements for human oversight of high-risk AI systems have been formulated, we see human involvement defined as protection against negative effects of algorithmic decision-making: "Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights." [5]

Scope

This document is about meaningful human intervention, which ensures that there is no automated decision-making as referred to in Article 22 (1) GDPR and Article 11 (1) LED. These articles relate to "a decision based solely on automated processing." This refers to a decision that is entirely based on automated processing and that has legal consequences for the data subjects concerned or otherwise significantly affects them. If this involves meaningful human intervention, this means that the decision is not solely based on automated processing.

Additionally, "human intervention" is also an appropriate measure under Article 22(3) of the GDPR to protect the rights, freedoms, and legitimate interests of the data subject. This applies in cases where a decision based solely on automated processing is permitted because an exception under Article 22(2)(a) or (c) of the GDPR applies. The components discussed in this document are also relevant for regulating human intervention when

[1] GDPR Art. 22 (1).
[2] Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, p. 8-21
[3] Guidelines, p. 21.
[4] Lazcoz, G., & de Hert, P. (2023). Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities. *Computer Law & Security Review*, 50, p. 18.
[5] AI Act Art. 14 (2).

a data subject has the right to it after an automated decision has been made. However, this document is not written with that specific purpose in mind.

This document is intended as a tool for those within an organization who design and implement human intervention. In this document, they are referred to as "designers." It is also for those who carry out human intervention, referred to as "assessors" in this document. This document is not a checklist: not all questions and components will or can be applicable to every process. The context and individual circumstances of each case are, of course, relevant and decisive.

Algorithms

The term 'algorithm' is not used in the GDPR. Nevertheless, in this document, the term 'algorithms' is used when referring to automated processing that leads to a specific outcome in the context of a decision to be made. Furthermore, it is important to distinguish between so-called rule-based algorithms and machine learning algorithms. Rule-based algorithms follow a relatively simple decision tree (if X, then Y), formula, or step-by-step plan. For this type of algorithm, meaningful human intervention can generally be well organized. In contrast, with a machine learning algorithm, the exact link between input and output is determined by a computer. Machine learning is a part of artificial intelligence (AI). The complexity of machine learning has consequences for the transparency and explainability of how the outcomes of these algorithms come about. This can be a problem if data subjects are affected by the consequences.[6]

We want to make a few comments in advance regarding the use of algorithms. For instance, this document does not address whether the use of an algorithm in a specific process is appropriate, or whether the data processed by an algorithm is suitable for assessment by an algorithm (e.g., evaluating driving skills). Additionally, the way a human arrives at a decision is not always better or more transparent. That being said: the more responsible decision-making relies on human insight, experience, customization or intuition, the less appropriate it may be to have the decision taken exclusively by an algorithm.[7]

We also find it important to caution against tunnel vision when it comes to the process of human intervention. The many factors and questions mentioned above may mask a larger issue – namely, that the decision being made is inherently unethical. Or that the use of an algorithm is morally problematic. Therefore, it is important to consider the nature of the decision to be made as well, independent of whether human intervention is meaningful or not.

Example

The use of an algorithm can completely transform a decision-making process. In the past, a civil servant with a social background might have assessed a family's financial situation through an in-person conversation at their home. Now, however, both parties must rely on an online form. However, a situation can be too complex for a form; in such cases, it is wise to allow the assessor the flexibility to provide a tailored approach.

[6] See Robbins, S. (2017). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, 29; Grant, D. G., Behrends, J., & Basl, J. (2023). What we owe to decision-subjects: beyond transparency and explanation in automated decision-making. *Philosophical Studies*, 182. This is not to say that meaningful human intervention will never be possible with case-based algorithms: there are techniques under development that may provide sufficient insight, see Molnar, C. (2024). Chapter 10 Neural Network Interpretation, Interpretable Machine Learning: https://christophm.github.io/interpretable-ml-book/neural-networks.html.

[7] For example, in France, the use of algorithmic decisions in the legal domain is prohibited under Art. 47 of the *Loi Informatique et Libertés*.

Reading guide

The components that make human intervention meaningful are divided into four chapters: human, technology and design, process, and governance. Each section consists of different subcomponents. Each section also includes questions that can help organizations guide the design of human intervention in a meaningful manner.

Consultation

The Autoriteit Persoonsgegevens (Dutch Data Protection Authority) (AP) invites you to respond to this document via email at ppa@autoriteitpersoonsgegevens.nl. You can submit your feedback until April 6, 2025. The feedback received will be summarized in a separate document, without mentioning names, organizations, or contact details. The summary will be published on www.autoriteitpersoonsgegevens.nl and used to improve this document. The revised document will be published later in 2025.

All feedback is welcome. We are particularly interested in insights from practice, for example: what makes the work easier for those who carry out human intervention? What obstacles do they face? How do these individuals cooperate with the algorithm?

Table of contents

1. Human

What does an algorithm lack that a human does have?
The GDPR does not allow people to experience legal consequences or be otherwise significantly affected by the outcome of an algorithm (with exceptions). A human is allowed to make such a decision. What human qualities does an algorithm lack when it comes to making a decision that takes human dignity into account? And what can a controller do to ensure those qualities are expressed?

Components

All relevant factors
According to the guidelines, assessors must consider all data relevant to the decision in their analysis. [8] This can include more than just the data processed by the algorithm, such as any additional information that the data subject may provide. Human assessors may assess whether it is relevant to use the same data that the algorithm uses or if they should take additional factors or information into account. If there is no room for making such judgments, it means that their assessment may not have sufficient meaning, and that the decision could ultimately still be considered as 'solely automated'.

Humans are also able to take data that is difficult or impossible to capture in an algorithm into account. Art. 22 GDPR implies a cooperation between humans and algorithms, where the human assessors pay attention to the individual circumstances of the case and algorithms reflect general patterns.[9] A human can serve as protection against machine errors and **digital rigidity**:[10] the indifference of algorithms to certain types of relevant information and the oversimplification of (complex) situations into a spreadsheet. For example, algorithms might signal an exceptionally high amount as an indication of fraud due to a misplaced comma, while a human assessor would quickly recognize it as a typo. Also, in certain exceptions, algorithms can draw incorrect conclusions, such as when an address field is empty due to a protected residential address of a data subject.

It is desirable that controllers with the help of assessors consider the aspects that the algorithm should take into account, as well as any additional aspects that the human assessors can consider before reaching their final decision. Certain factors are better suited to being assessed by an algorithm, such as a required minimum number of years of work experience for a job applicant. Other factors, such as the quality of writing in a cover letter, generally require a human perspective. A human assessor will also be able to extract more information from certain data than an algorithm can.

[8] Guidelines, p. 21.
[9] Binns, R. (2020). Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making, *Regulation & Governance*, 11 (1).
[10] For example, see Sztandar-Sztanderska, K., & Zielenska, M. (2022). When a Human Says "No" to a Computer: Frontline Oversight of the Profiling Algorithm in Public Employment Services in Poland, *Sozialer Fortschritt* 71 (6-7), p. 2.

Example: complaints department
At the complaints department of a delivery service, each complaint is jointly assessed by a human and an algorithm. Initially, both the human assessor and the algorithm see the same customer complaint about a delivery person. The algorithm selects whether the complaint should be further investigated based on the rating the customer gave to the delivery person. The assessor evaluates whether the complaint is specific enough and substantively significant, as a vague complaint could be a signal that it was filed just to "harass" the delivery person. A decision about the complaint is made only after this evaluation.

This does not mean that it is always better to let a human assessor consider (more) variables. Human assessors may view decisions through their own biases, and eliminating these biases may in some cases be an argument for using an algorithm instead.

The factors determined by the controller are not exhaustive. It is desirable that assessors are given the space to add other factors that they consider to be relevant. For example, it may sometime be necessary to contact a data subject to gain insight into an unusual variable, such as a wrongly filled field, or information that is missing for a valid reason, such as a protected residential address.

Example: warehouse employee
An algorithm flags warehouse employee as an underperformer because they clocked in later than their colleagues. The assessor might quickly notice that the employee had a dentist appointment that day.

The space given to the assessor does not have to be unbounded. In some cases, it is not desirable for an assessor to add information. In addition to the general principle of data minimization, specific restrictions also apply for special categories of personal data. The data subject may also experience a request for additional information as negative. In fact, such further investigation could significantly impact the data subject. The training of assessors can address these negative aspects of requesting information: the impact on the data subject, alternatives to contacting the data subject, and situations where further investigation is useful.

Implementation questions

- Do assessors include all relevant information in their assessment?
    - Can assessors consider specific circumstances in their assessment that an algorithm does not take into account?
- Do assessors have more information at their disposal than the algorithm?
    - If not, can assessors gain access to that information?
- On what basis should assessors evaluate a decision, or potentially go against the algorithm?
- In which way(s) can assessors go against the algorithm?
- Can assessors disregard, complement or correct data in the algorithm?
    - For example when information is missing or it is clear that something was formulated incorrectly.
- Is it clear what is expected from assessors?

Human discretion

What makes human intervention ***human***? This question touches on the essence of human intervention. Simultaneously, it is perhaps the question with the least clear answer. In addition to the fact that humans and algorithms can assess different data, humans and algorithms assess data differently. The human way is less tangible than the algorithmic way. Various attempts have been made in the literature to define this: ethics, a

human measure, empathy, capturing unpredictable factors, (emotional) intelligence, moral ability, conscience, interpretation, experience, creativity, or non-linear reasoning. This idea is most often described as **'discretion'** (in our Dutch document, this has been translated as 'menselijk inzicht' or 'human insight'). [11] If implemented in the process, human discretion may protect human dignity.

Human discretion does not have to be unbounded. We cannot explain how a human reaches a decision as clearly as we can for an algorithm. For humans, sometimes it is nothing more than a feeling, whereas for algorithms, it is as a precise numerical score. This is why it can be difficult to use human discretion as the basis for rejecting the outcome of an algorithm. This is especially true for decisions that are better suited for human judgment rather than for consequences based on an algorithm's outcome.

Example: job application
Algorithms that could infer a job applicant's leadership qualities from their writing style are in development. These algorithms can assign a score to the applicant. An assessor who, after an interview, believes that the candidate lacks sufficient leadership qualities based on experience might start to doubt their judgment when the algorithm has rated the candidate a 9 based on the cover letter. The assessor's judgment is less clearly to explain than the algorithm's numerical score.

Sometimes, nuance is lost in the development of an algorithm. The assessor can reintroduce this nuance into the decision. When translating policy for a specific decision into code for the algorithm, developers and/or controllers make choices: which factors should be considered? Which individual situations and exceptions do we account for in advance? The assessor can, when necessary, deviate from the algorithm's rigid perspective on the data subject's situation.

Different interests play a role in the decision-making process. It is important that the assessor takes the interests of the data subject into account—such as risks to health, safety, and fundamental rights, as outlined in the AI Act. Organizational interests often also influence the use of an algorithm, as it can make processes faster, more efficient, and therefore cheaper. However, such interests can have a significant impact on the decision and may hinder meaningful human intervention.

Example: insurance claim
An intermediary assesses whether a claim should be paid to a policyholder on behalf of insurers. From a financial point of view, it is in the insurer's best interest to pay out fewer claims. The intermediary instructs the algorithm's designers and the assessors that decisions should not be in favor of the policyholder too often.

It is important to reiterate that adding human discretion to the process can also have drawbacks. People may have biases that an algorithm does not have. They may also be less capable of evaluating certain variables that an algorithm ***can*** assess accurately. Additionally, humans may struggle to consider all relevant factors simultaneously when making a decision.

[11] For example, see Solove, D. J., & Matsumi, H. (2024). AI, Algorithms and Awful Humans, *Fordham Law Review*, 92 (5); Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet*, 11(1); Lazcoz, G., & de Hert, P. (2023). Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities. *Computer Law & Security Review*, 50; Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45; Palmiotto, F. (2024). When Is a Decision Automated? A Taxonomy for a Fundamental Rights Analysis. *German Law Journal*, 25 (2).

Implementation questions:

- Do assessors apply human discretion in every case by, for example, evaluating aspects of the decision in their own way?
- To what extent does the decision require an assessment of individual factors?
- What requirements are placed on assessors?
- What other interests, other than those of the data subject, does the assessor take into account?
- To test whether there is enough room for human discretion, an assessor can be asked to make several decisions before an algorithm does or before an algorithm provides input. The assessor's decisions can then be compared with the algorithm's output. This process is known as *cognitive forcing*.

Competence

The assessor is expected to be competent. [12] This means they must have the knowledge and skills required to make the decision. Additionally, they must have an understanding of the algorithm. An assessor should at least have a general understanding of how an algorithm arrives at its results. If an assessor lacks insight into the algorithm, they would essentially have to re-evaluate the entire process when weighing its outcome. The data controller is responsible for providing information and training to ensure the assessor has sufficient understanding of the algorithm.

Occasionally, the skills required to assess an algorithm's outcome may differ from those needed to make a decision without the algorithm. Evaluating an algorithm can sometimes be more complex than the decision-making process itself. Some decisions may also require highly specialized knowledge, such as in a medical context. In such cases, an assessor may not be able to specialize in both the decision-making process and the algorithm's technical details. When this happens, it may be necessary to establish a **team-in-the-loop** approach instead of relying on a single individual. In this setup, the assessors who make the substantive decisions about the data subjects work together with colleagues who have expertise in evaluating the algorithm.

Implementation questions

- Do assessors understand how and based on what data an algorithm arrives at a result?
- Do assessors have basic knowledge of statistics?
- Do assessors know which factors the algorithm takes into account?
    - Do assessors have access to the data used by the algorithm?
- Would assessors be able to make the decision without the algorithm?
- Can assessors make the decision independently, or does this require a team in which various specialist areas are represented?
    - How does that team cooperate?
- Is there sufficient insight into the algorithm being used?
- Is it clear what is expected of an assessor in terms of human judgement and organizational interests?

[12] Guidelines, p. 21.

2. Technology and design

How does the algorithm influence the human assessor?

When talking about human intervention, it is important to also consider at the technology itself. Technology is never neutral and can affect the extent to which human intervention is meaningful. For example, an algorithm's interface design and the data on which an outcome is based can influence assessors. However, remember that these are human choices – after all, humans design the technology. In general, the more a human adapts (or has to adapt) to an algorithm, the more automated a decision becomes. An algorithm may for instance limit an assessor's options (e.g., by providing only a binary yes/no choice), whereas before the algorithm's implementation, the assessor might have arrived at a tailored solution. If the algorithm dictates how an assessor must act, this can come at the expense of their autonomy. Human intervention can become more meaningful through appropriate technical measures.

Relevant concepts

Automation bias

Automation bias refers to the tendency of people to overestimate the performance and accuracy of algorithms. We often place too much trust in algorithms, even when they make mistakes. In other words: people tend to accept algorithmic output as truth too quickly. This can lead them to ignore their own knowledge or observations. For example, a British study found that police officers in London overestimated the reliability of real-time facial recognition technology three times more often than was actually justified. [13] Algorithms are often framed as more reliable or precise than humans, reinforcing this bias.

This framing can contribute to **automation bias**. It is therefore useful to examine how an algorithm is talked about within an organization. Is it described as highly reliable and a crucial part of a process? Or is it emphasized that the algorithm only plays a supportive role? It is important that assessors are aware of automation bias so they can recognize it. Training sessions can address this issue.

Algorithmic aversion

On the other hand, people can also underestimate an algorithm's performance, even when it is known to be more accurate. This is called **algorithmic aversion**. This tendency often arises when algorithms make decisions about people, such as granting a visa or approving a loan. Algorithmic aversion can cause various problems, including human bias.

These two concepts are relevant to meaningful human intervention. **Automation bias** and **algorithmic aversion** show that adding a human does not always lead to desirable results. On the one hand, people may be inclined to quickly accept the output of algorithms as the truth, which makes human intervention less meaningful. On the other hand, people may unjustifiably have less trust in algorithms. This critical attitude could result in missed opportunities that a responsibly designed algorithm could offer.

[13] Fussey, P., & Murray, D. (2024). Policing Uses of Live Facial Recognition in the United Kingdom. AI Now Institute. https://ainowinstitute.org/wp-content/uploads/2023/09/regulatingbiometrics-fussey-murray.pdf.

Components

Interface

Design can influence our behavior. Elements of an object can either encourage or hinder certain behavior. The same applies for computer interfaces. Consider the use of specific colors in pop-up windows for example. The design of an algorithm's interface can influence the assessors. Ideally, an interface is designed with the end users in mind. This sounds logical, but does not always go well in practice.

Example: aircraft radar

In 1988, the U.S. Navy warship USS Vincennes shot down an Iranian passenger aircraft as a result of a bad interface design. Air-traffic controllers were under the assumption that the aircraft was flying towards the ship, while in reality it was flying away from it. The direction of the aircraft was not clearly visible on the screen. The screen also did not show the speed of the plane. As a result, employees had to compare data manually and make calculations in their heads, on scrap paper, or on a calculator. [14]

An interface can also affect neutrality. For example, colors may evoke certain associations. Consider a red risk score in the interface of a fraud detection algorithm. A red signal can imply that someone has committed fraud, even when this is not the case. A more neutral score gives the assessor a better chance to make an objective decision.

Humans communicate differently from computers. Well-designed interfaces ensure better communication between humans and computers. Design elements can be taken into account in the interface, such as providing explanations for certain data. For example, the interface could indicate how and based on what factors a risk score was generated. Elements like color, font, or pop-ups can also help make the interface more understandable. [15] Therefore, design elements can be used to encourage meaningful intervention. For example, an assessor could be encouraged to check certain input data by the interface.

Amount of data

The amount of data an assessor is shown when an algorithm generates a result is crucial for the process and for the added value of human intervention. You can imagine that it is difficult to make the right decision when you have too little information. At the same time, an excess of data can make it difficult to arrive at a well-informed decision. Too much data can be overwhelming. Algorithms often generate output based on hundreds or thousands of data points, which is difficult for a human to grasp. So, which data *should* the assessor be shown? Organizations must think carefully about this.

Data in context

Furthermore, data without context has little meaning to a human. For example, an oxygen meter in an office space tells us little if we do not know the desired oxygen level. In some processes, human lives are reduced to data, and the outcomes of algorithms may have significant consequences. It is therefore important that data is placed in the correct context so that the assessor can make a well-informed decision. One could argue: the

[14] Cummings, M. L. (2006). Automation and Accountability in Decision Support System Interface Design. *The Journal of Technology Studies*, 32 (1).

[15] Such design components are often employed in *dark patterns*: tricks used in the design of websites and apps to make people do things they did not intend (such as accepting cookies). In dark patterns, design components are deployed in a negative way and not for the benefit of users. But, design components can also be deployed *for* the benefit of users.

more abstract the data, the less meaningful the human intervention will be. Abstract data consists of numbers without clear explanation, such as a risk score of 0.5. Without context, it is not clear what this risk score is based on or how it compares to other scores.

Implementation questions

- What is visible on the interface when an assessor evaluates the outcome of an algorithm? [16]
    - To what extent has the interface been designed in a clear and understandable way?
- Are assessors involved in the design phase?
- Does the interface make the decision clearer, for example by providing explanations for numbers and graphs or a reliability score for the result?
- Are there any design elements that could affect the neutrality of assessors?
    - To what extent do assessors understand what is displayed on the interface?
    - How is data presented to assessors?
- Is data presented in a clear and understandable manner?
- Is it clear what certain data means?
- Which data do assessors need to make a decision?

Order of data

It is also important which data an assessor sees first. The information that a person sees first often forms the basis for later decisions. Our brain tends to latch onto a certain reference point, regardless of what that reference point is. This is called **anchoring**. By showing certain information first, subsequent decisions can be influenced. In practice, this could happen when someone is marked as a 'potential fraudster or risk.' This will influence the assessor's judgment.

In general, people tend to pay more attention to factors that are emphasized. Therefore, an assessor can be (overly) influenced by the presentation of data or the result of an algorithm. For instance, it has been shown that assessors are more likely to assess something as a 'risk' when a risk score produced by the algorithm is part of the result.[17]

The data that an assessor sees first thus affects their judgment, but the order in which data is presented also has an effect. For example, sorting data alphabetically can have undesirable effects. When people are sorted alphabetically by location, it may happen that someone from Amsterdam is more likely to be subject to further investigation than someone from Zoetermeer.

Finally, it may be necessary for an assessor to check whether the information used by the algorithm is correct. This is especially true when it concerns exceptional data. There may be an option in the interface to check, correct, request additional data, or add data.

[16] Go through the assessment process with an assessor to see what they see when making a decision. Pay attention to what data is shown, but also to *how* the interface is designed. For example, does the interface look understandable and clear?
[17] Green, B., & Chen, Y. (2021), Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts, Proceedings of the ACM on Human-Computer Interaction, 5.

Implementation questions

- What data is shown to assessors when assessing an outcome?
    - What effect does this have?
- How much data is shown to assessors when making a decision?
    - Do assessors see enough data to make a well-informed decision?
    - Do assessors receive the same amount of data for each decision?
- If not, what happens when an assessor has (too) little data?
    - Is there an option in the interface to request or add additional data?
    - Is there also an option in the interface to correct data?
- Is data presented in a specific order?
- What possible effects might this have?

Routine

If the work of an assessor becomes routine, without having an influence on the outcome, it loses its function.[18] This could be taken into account in the design of the algorithm. For example, an outcome may be presented in different ways, and control questions may be built in. There can also be variation in the presentation of the output. For instance, an algorithm might assign a risk score to individuals, but it could also explain why an individual poses a risk, without attaching a risk score to it. Think, for example, of 'risk of theft', 'risk of burglary' and 'risk of vandalism', instead of a general risk score. This variation in output requires more thought from the assessor to come to at a well-considered decision. This may lead to meaningful human intervention. It is important that the controller ensures equal treatment of the individuals involved. Note: Be aware that different presentations of a decision could cause the decision itself to change.

The assessor may also regularly be presented with a decision that must be made without the algorithm. Finally, the data subjects selected by the algorithm for further review, for example, may be supplemented with random data subjects who were not selected. This ensures that the assessor cannot blindly trust the algorithmic output.[19]

Implementation questions

- Has the prevention of routine been considered in the design of the algorithm?
    - How is the outcome presented?
    - Are there control questions?
    - Is data or the output presented in a different order? Or does the order in which a partial decision is made vary?
- Do all results that assessors see come from the algorithm?
- Do assessors often have to make decisions manually?
- Are assessors kept alert?
- Is there a check for errors?

[18] See the Guidelines, p. 21.

[19] See also AI & Algorithmic Risks Report Netherlands - Summer 2024, Dutch DPA, p. 4. https://www.autoriteitpersoonsgegevens.nl/en/documents/ai-algorithmic-risks-report-netherlands-summer-2024.

3. Process

How do the organizational choices influence the human assessor?
When assessing the meaningfulness of human intervention, it is important to look at the process surrounding it. Ideally, an organization has a clearly described process for meaningful human intervention, and the involved employees can explain this well. For example, thought should be given to ***how*** an assessor should evaluate the outcome of an algorithm. Additionally, specific requirements for meaningful human intervention should be established, and it should be well justified based on which data an assessor makes a decision.

Example: content moderation
Researchers looked at employees who moderated content on a large social network after an algorithm had flagged the content as potentially problematic. [20] The workers "often work in very difficult conditions with low pay and are typically given only a few seconds to decide about each piece of content…As the workers are paid very badly, these jobs do not attract skilled workers and particularly not those who have any kind of legal or technological qualifications." The research concludes: "While this might suggest the presence of human agency at some levels, humans are so deeply embedded in the algorithmic systems that decide things around them that it would be hard to call them actual decision makers per se. Rather, the analysis seems to indicate they are actually embedded there to suggest a limited degree of human agency within a highly complex system. While human labor may be necessary in some areas, part of the reason for human involvement seems to suggest human agency when there is actually very little."

Components

Timing
Human intervention can occur at various points in the process before the final decision-making. Roughly, human intervention can take place at the following points: [21]

a) An algorithm provides the necessary information for a decision, but the assessor makes the decision. The timing of human intervention in this case occurs at the end of the process, but before the final decision is made.

b) An algorithm provides information for a (sub) aspect of the decision or is used to describe relations between data, make a diagnosis, or predict events. The assessor makes the decision. In this case, the algorithm provides information that, after further examination, can lead to a decision. Here, human intervention occurs earlier in the decision-making process.

c) Human intervention occurs at different points in the process.

You can imagine that human intervention is less meaningful if there is only a check on the output of an algorithm. Especially when it is not clear how the algorithm arrived at this output. In fact, the decision has

[20] See Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet*, 11(1).
[21] See also Raad van State. (2021). Digitalisering: Wetgeving en bestuursrechtspraak. https://www.raadvanstate.nl/@125918/publicatie-digitalisering/.

essentially already been made. In such cases, the algorithm could be seen as having a decisive role. However, even then, a thorough analysis by a human before the final decision is made can still be meaningful. For human intervention to be meaningful, an assessor must be able to influence the outcome of an algorithm. The degree of influence partly depends on where in the process human intervention takes place.

An algorithm that labels a pupil as a 'slacker' differs in timing from an algorithm that provides various data points from which the assessor draws a conclusion. In the first example, human intervention takes place afterward. In the second example, human intervention takes place earlier in the process.

This is an example of a supporting algorithm. This difference may also be evident from the reason for the use of the algorithm. Is work taken out of the hands of experts for greater efficiency? Or is expertise supported, as in the case of diagnostic tools?

Human intervention can also take place at multiple points in the process. For instance, the input from an algorithm may first be checked by an assessor. Then the algorithm provides information for certain aspects of the decision. The assessor evaluates these, and the final output is also assessed by the assessor. Since different people often perform these tasks, this is also known as team-in-the-loop approach.

Implementation questions

- At what point in the decision-making process does human intervention take place?
    - Does the algorithm provide a complete outcome?
    - Or does the algorithm provide information about a (sub) aspect of the decision to be made?
    - Do assessors have to check the decision of an algorithm?
- Does human intervention take place at various points in the process?
- Which role does the outcome of the algorithm play in the decision?
- What is the role of the algorithm in the process (decisive, supporting, or controlling)?

Workload

Assessors often make decisions about people with potentially far-reaching consequences. It is important that that assessors have enough time to do this. The less time an assessor has to make at a decision, the less meaningful the human intervention will be. It is important here to remember that there is no ideal time unit, as this depends on the context. Ideally, the entire process is reviewed to ensure that assessors have enough time for their decisions.

An organizational culture that puts emphasis on efficiency may negatively impact human intervention. For example, if a certain number of decisions must be assessed per day. It may even occur that (commercial) organizations work with targets. As a result, assessors may not have or take enough time to make a well-informed decision.

Implementation questions

- How much time do assessors usually have for assessing the outcome of an algorithm? [22]
    - How does this relate to the nature of the decision to be made? [23]
- How many assessments do assessors make on average each day?
- Is there a minimum number of decisions that assessors have to assess each day?

Authority

When an assessor assesses an outcome of an algorithm, it is important that they are ***allowed*** to overrule the outcome. Additionally, it is important that the assessor actually does this when necessary. An assessor might be formally authorized to go against the algorithm, but may encounter obstacles in practice. These obstacles can make human intervention less meaningful.

Example: work monitoring

An organization uses a monitoring algorithm to track employee performance. A manager receives an email from the monitoring algorithm stating that one of the employees is underperforming. The manager immediately sees that the alert was linked to the wrong employee, but has never received instructions to make changes in the system.

It is important to check whether an assessor feels free to challenge the algorithm. Organizational culture plays a role here. For example, in a strong hierarchic culture, assessors may not feel free enough to go against an algorithm. Or they may be reluctant to go against the algorithm when they are harshly punished for mistakes.

The consequences an assessor experiences from their own mistakes or the algorithm's mistakes also play a part. If an assessor is held accountable for making an incorrect decision, this may discourage them from challenging the algorithm. This is why an assessor should be authorized to go against the algorithm, but not be punished for making a mistake. This does not mean that an assessor should not face consequences for incorrect decisions. The controller should able to assess the quality of the human intervention (also see the section on monitoring).

Example: benefits

Employees at Public Employment Services in Poland received a handbook during the implementation of an algorithm. The handbook stated that any change in the system (such as correcting data) was recorded. [24] Out of fear of being punished for errors, assessors only rarely went against the algorithm's outcome.

[22] If possible, go through the process several times with assessors to get an estimate how much time they typically spend on a decision. Note that this may require working with synthetic datasets.

[23] A decision with less far-reaching consequences on an individual may require less time.

[24] Sztandar-Sztanderska, K., & Zielenska, M. (2022). When a Human Says "No" to a Computer: Frontline Oversight of the Profiling Algorithm in Public Employment Services in Poland, *Sozialer Fortschritt* 71 (6-7).

Implementation questions

- Are assessors authorized to override the outcome of an algorithm?
    - And how has that process been structured? Is it documented in policy?
- How often do assessors go against the outcome of an algorithm?
- Do assessors experience obstacles to going against the algorithm?
    - Do assessors face negative consequences when they go against the algorithm?
    - How easy is it for assessors to go against the algorithm?
    - Does the organizational culture influence the work of assessors?
- Are there quality checks on the work of assessors?
    - What are the consequences of these checks?
    - Is feedback provided from these checks?
- Does the organization take responsibility for incorrect decisions made about a person?

Support

The work of an assessor requires a lot of attention. After all, they make decision with far-reaching consequences for individuals in a relatively short time. When assessors are provided with sufficient support, this may be an indication of meaningful human intervention. Support can take different forms, such as through a confidential adviser/officer. Assessors can seek support from one another. Typically, assessors make decisions individually, but a team-in-the-loop approach can help distribute the mental load. In a team-in-the-loop approach, no single assessor is solely responsible for a decision about an individual. Instead, multiple assessors are involved in the decision-making process.

Implementation questions

- Do assessors have access to sufficient (mental) support when needed?
    - Is there a confidential adviser, officer or trusted person?
    - Can assessors ask each other for help?
    - Is attention given to resilience during training courses?

4. Governance

How does the organization maintain ultimate responsibility?
It is crucial for an organization to retain responsibility for the use of an algorithm. Human intervention ensures that the outcome of an algorithm does not lead to a decision that is based solely on automated processing. This responsibility should not lie with the assessor alone. Various researchers studying automated decision-making have warned against this. [25] How can an organization ensure that ultimate responsibility for the outcomes of the process remains with the right persons?

Components

Implementation
It is desirable for an organization to document its policy for meaningful human intervention clearly within procedures, such as decisions about the implementation, for which all of the above components can serve as guidelines. In a Data Protection Impact Assessment (DPIA), the data controller should identify and record the degree of any human involvement in the decision-making process and at what stage this takes place. [26] For a system that produces outcomes that may significantly impact data subjects, it is likely that a DPIA will need to be conducted. [27]

It is good practice to involve assessors in the design of the process and the development of the algorithm. Managers and developers responsible for decisions made during this may be further removed from the process and may lack the necessary expertise.

Assessors may be able to provide advice on which aspects can be assessed by an algorithm, and which should be handled by a human. They can also indicate whether the interface has been designed clear enough. Before an algorithm is deployed, it is advisable to test it with human intervention. The results of these tests and any adjustments resulting from them can be incorporated into a DPIA. During this test phase, it is essential to evaluate, among other things, how much time an assessor needs for the intervention, what knowledge is missing for a proper assessment, and which variables are missing in the algorithmic part of the decision-making process, according to the assessors.

[25] Discussed by Green, B., & Wagner, B. during the IPEN event on "Human oversight of automated decision-making," 3 september 2024. Assessors can be blamed for harmful effects caused by the choice to deploy an algorithm, or by improper design. Human intervention then acts as a band-aid on a process that is already fundamentally broken.
[26] Guidelines, p. 21.
[27] From Lazcoz, G., & de Hert, P. (2023): "Article 35(3)(a) is however crystal clear by requiring a DPIA in all cases of systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person."

Implementation questions

- Is policy for meaningful human intervention documented?
- Who were involved in drafting this policy?
    - Has the perspective of assessors been included in this process?
- Who were involved in the design of (the interface of) the algorithm?
    - Has the perspective of assessors been included in this?
    - Has the algorithm been tested with assessors?
- Has a DPIA been carried out on the process?
- Has the perspective of assessors and data subjects been included in this DPIA?

Training

To ensure meaningful human intervention, assessors need training and information.[28] Many of the components listed above require providing assessors with instructions.

Aspects that may be important for the training: [29]

- **All relevant factors**:
    - Assessors understand how their expertise complements the algorithm and know which factors must be considered in the decision.
        - The controller may draw up a list with factors of importance.
        - In addition to IT skills, the necessary (social) skills for making the decision must be taught.
    - Assessors know when and in which way additional information may be requested, for example from the data subject.
        - Assessors are aware of impact of requesting additional information on the data subject.
- **Human discretion**:
    - Assessors understand the possibilities for tailoring decisions to specific situations.
    - During the training, assessors are not encouraged or instructed to adjust their decision to organizational interests, such as retaining customers.
    - The training addresses human bias.
    - The algorithm is not presented as infallible or better than humans during the training.
- **Competence**:
    - Assessors understand how the algorithm arrives at its outcome.[30] Some example questions to test this knowledge:
        - How does the output of the algorithm change if certain variables change?
        - What are the most important inputs for the algorithm's outcome?
        - What choices were made regarding the fairness metrics of the algorithm? In other words: which groups may be flagged more frequently by the algorithm?
        - Which rules does the algorithm follow?

[28] In an Austrian case, the Federal Administrative Court (BVwG) ruled that the controller must provide evaluators with training and instruction so that they do not uncritically adopt the results of the algorithm (ECLI:AT:VWG-H:2023:RO2021040010.J09)

[29] Partly from Information Commissioner's Office (2020). Guidance on the AI auditing framework. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

[30] From Lazcoz, G., & de Hert, P. (2023): "The French Conseil Constitutionnel (Conseil Constitutionnel, Décision n° 2018-765 DC du 12 juin 2018, §71) declares human intervention a fundamental safeguard in the design and development of AI algorithms, see Malgieri (n 16) 15. And further recognises the link between this safeguard and the ability to explain, in detail and in an intelligible form, how the processing has been carried out to data subjects. "

- What exactly is being assessed? The controller should emphasize that the decision impacts a person and explain what the impact is.
- Assessors recognize the limitations of the algorithm.
- Can assessors recognize when the output of the algorithm is (likely to be/possibly) incorrect?
- Do assessors know how often the algorithm makes mistakes?

- Assessors are able to provide a justification for accepting or rejecting the outcome of the algorithm.
  - The controller may provide guidelines for cases in which correcting the algorithm is necessary.

- **Technology and design**:
  - The training addresses **automation bias** and **algorithmic aversion**.
  - Assessors understand how to use the algorithm's interface and the meaning of the displayed information.
  - The training warns against routine-based assessments of algorithmic outcomes.
- **Process**
  - Assessors know how to reject an algorithmic outcome.
  - Assessors are confident that they can reject the algorithm's outcome without facing negative consequences. This is explicitly covered in the training. [31]
  - Assessors know where to go if they have any concerns, questions or feedback.
  - The training pays attention to resilience.

Implementation questions

- Which of the above components are relevant to the training?
- How does the organization keep the knowledge up to date (e.g., through refresher courses)?

Testing and monitoring

Even after the implementation phase, it is desirable that the process can be adjusted based on signals, as described above. Since research into the human role in automated decision-making is still ongoing, it is crucial that the controller keeps an eye on how the process functions. This can be done by testing and monitoring.

In order to test if human intervention is actually meaningful, the organization can track various metrics. A simple method is to monitor how often an assessor rejects the outcome of an algorithm (or changes a "yes" to a "no" and vice versa). This can serve as a starting point for further investigation. Assessors can also receive feedback on their accuracy in dealing with algorithmic outcomes.

A controller may also carry out recurring ***mystery-shopping*** actions, during which misleading data or algorithmic outputs are deliberately introduced. The assessor should disagree with this and verify that they detect the errors. The assessor may also regularly be presented with a decision that must be made entirely without the algorithm. Another indication that the process should be reassessed could come from data subjects. The controller should track the number of complaints, objections, or requests for human intervention and ensure proper follow-up.

[31] In Sztandar-Sztanderska, K., & Zielenska, M. (2022). When a Human Says "No" to a Computer: Frontline Oversight of the Profiling a low amount of corrections was partly caused by information assessors had received from the trainer: "As one of the client advisors from the pilot study recalled, PES staff were warned by the trainer (who took part in designing the profiling tool) not to correct profiles, "because the Ministry will check it" and the employees will "have to explain themselves"."

In evaluations, the connection between individual decisions and the entire process is also important. For example, the process and the algorithm should be adaptable based on feedback from assessors. Since assessors are closest to the decisions, they play a crucial role in ensuring that the process remains fair. If a assessor notices that people with a certain educational level are disproportionately flagged by the algorithm, this could indicate a bias that requires further investigation. However, it is important that the data controller does not shift the responsibility for overseeing the entire process onto the assessor. It goes without saying that the controller must adjust the process as needed based on testing and monitoring.

Monitoring assessors can also have a ***chilling effect***, making them less inclined to openly question an algorithm. The way monitoring is conducted and communicated can help prevent this effect.

Implementation questions

- Is the algorithm adjusted after feedback from assessors, data subjects or monitoring?
    - Do the individuals who evaluate the feedback have the right skills for this?
- Do assessors feel free to criticize the algorithm and the surrounding process?
- Does the organization monitor the degree of meaningful human intervention?
    - Consider mystery shopping. Evaluate how often data subjects file objections and how often assessors override algorithm.
- How does the organization handle system errors?
    - How does the organization deal with false-positives and false-negatives? [32]

[32] Consider, for example, wrongly flagging someone as a "fraudster" in a fraud detection algorithm. This would be a "false positive."

5. Conclusion

With this document, the concept of 'meaningful human intervention' has been made more tangible for data controllers who want to implement it. We will update this document as needed.