Transparency and explainability for algorithmic decisions at work

Any employer or platform that uses AI or other algorithmic tools to manage their workers must be clear and open about how those tools work. Workers need to be able to understand decisions that affect them, especially when it's about important things like their working hours and pay. Otherwise, companies can unfairly leverage control over workers by withholding important information from them.

That's why workers must be given the right level and amount of information at the right time, including clear reasons for why decisions are made. This information should enable workers to understand which parameters are the most important in an automated decision-making process and what changes they can make to get the outcomes they prefer. They must also know how they can ask a human to review algorithmic decisions. Without parity in information, workers are left playing a game that they don't know the rules for.

Ultimately, this is about respect for the rights of workers who are subject to algorithmic management, in particular their right to privacy and dignity. To provide the foundation for that respect, we're calling on all employers and platforms to do the following when using AI and algorithms in the workplace.

1. Maintain a public register of the algorithms used to manage workers

The register must include all algorithms that affect how workers are managed. This includes any algorithm that produces outputs which influence decisions made about how workers access and are assigned work, how their performance is assessed, how much they are paid, when they must be available for work, what employment status they have, whether they continue to be offered work, and any other matters relating to their terms of work.

The public register is key to addressing the information imbalance of algorithmic management by allowing workers (and candidates) and their representatives to understand what algorithms are being used and how they work. In order to do this, the register must be in accessible non-technical language and kept up to date. For each listed algorithm, the following information must be included:

- The purpose and design of the algorithm

A short (two or three sentence) description to explain what purpose(s) the company uses the algorithm for and why it has been preferred to other options.

An overview of the algorithm's design should also be given including what sort of management decisions are made by the algorithm (and whether its outputs are advisory or decisive); whether the algorithm relies on neural networks, machine learning, probabilistic functions or other type of logic; what training data was used; and under what circumstances the algorithm is not deployed or has a failsafe.

- The relative importance of the algorithm's inputs and parameters

The register must explain, in an accessible and non-technical way, what data and ratings it uses to reach decision. This means providing an easy way to understand how important different inputs and parameters are to different decisions. This could be done in various ways: from a simple rating of 'high/medium/low importance', to giving more specific and granular detail of the weighting or impact each input or parameter has, or the overall distribution of the algorithm's inputs and outputs.

It should also explain the source of inputs (are they from the app, from customers, from the web, inferred, how long ago, while at work, from data brokers, etc) and how the accuracy and reliability of them is ensured. Where parameters (defined as complex input inferred or calculated by another algorithm, such as a risk score) are used, the company should provide a detailed description of the parameter and how it is calculated, including the data used to determine its value. Fundamentally, complex inferred parameters should not be used to obscure the source and weight of data used in algorithms.

Where decisions are based on the personal data or behaviour of workers and/or consumers this must be clear, and the source of any personal data used should also be set out. The register should also confirm that the algorithm uses only data that is strictly necessary for the purposes of the algorithm, and does not use any sensitive personal data, emotion recognition, data collected while not at work etc.

It is possible that AI algorithms will use parameters that are hard to give real-world human descriptions for. In such cases, the company must state this and thoroughly explain how the tool has been built, and how they monitor and audit its outputs to ensure that they do not result in bias or discrimination. Examples (or statistics) comparing different, but similar, inputs with differing outputs may also be valuable to explain which sorts of inputs tend to lead to which sorts of outputs.

- Human intervention

Where algorithms affect how workers are managed, there should always be a human either checking, and/or able to review, any decisions.

Companies must specify what level of decision-making authority oversight teams have, as well as the training that decision-makers have received particularly in respect of the design and potential impacts of the algorithm. The register should break down the level of decision-making authority across multiple stages of review and oversight, specifying at what points in the decision-flow a human makes a decision and how long they have to make that decision.

The register should also provide some operational information about how much staff capacity (in FTE) is dedicated to human review and how long a review is expected to take.

- Development history, updates, and impact assessments

The company should also state where responsibilities for the development and updating of the algorithm lie, especially where an external supplier has been involved. This does not require

identifying individuals, but rather relevant teams/departments/organisations and the nature of their different responsibilities. A log of updates should also be listed.

The register should also specify what, if any, consultation has taken place between the company and workers and their representatives with respect to the design or revision of the algorithm. Active consultation with workers helps prevent harmful practice and collective agreements ought to include provisions on transparency, explainability and consultation for algorithmic management.

Any impact assessments of algorithmic systems (including Data Protection Impact Assessments and/or Algorithmic Impact Assessments) should be published either in redacted form or in full. And where there are redactions, these must be justified. When made public impact assessments should incorporate process transparency so that changes over time are documented and explained.

With regards to how this information is shared, we encourage the use of a variety of means such as flow charts, FAQs, or short form videos to accompany textual explanations. The selected mediums should make understanding how the algorithm operates as accessible and simple as possible.

Demand 1 Examples

Case Scenario 1: Worker identification system

Purpose: Ensure the person logged in and using the service to work is the person registered for this account. This system aims to confirm the account is used by the registered worker and not used by a different person. This was the preferred solution to keep costs manageable given the high number and high turnaround of workers using the platform. A third-party service was selected to avoid developing an in-house solution which would have to reach a high standard with regard to potential bias and false positives.

Design: Deterministic system relying on a third-party facial recognition service. This system captures photos of the users' face and match it against previously stored photos of the account owner.

Parameters and importance:

- Biometrics data captured in the photo (high importance, normally decisive)
- Metadata including device used to capture the photo, time, date (low importance)
- Previously recorded account information such as ID, previously captured photos (supporting importance)

Teams involved:

- Human Intervention: 5 members of staff trained to review appeal by workers.
- In house Customer Identification team: notified of algorithmic decisions and challenges brought by workers. Can override a decision after investigation. Keeps a record and report bugs and other issues detected to the development team in charge of the system. Minimum training of 2 weeks required to be part of the team.
- Third party FRT development team: maintain and update the system, informed about bugs and issues.
- Data Protection team: Audits algorithm to ensure data collected and processed adheres to Privacy Policy and Data Protection Laws.

First deployed: 01/03/2022
Last major update: 24/07/2023
Impact assessment: Completed on 15/02/2022
Engagement with workers and workers representative: None

Case Scenario 2: Account/contract termination

Purpose: Company policy is to terminate a worker's contract if their feedback ratings and efficiency fall below a certain level. The aim of this system is to identify accounts that do not match the standards set by the company and should have their account reviewed for termination.

Design: Deterministic system that monitors reviews, reports and timings to automatically flag workers who reach a defined threshold

Parameters and importance:

- Number and quality of reports from clients having interacted with the worker - High
- Feedback from clients having interacted with the worker - High
- Number of hours active on the platform - medium
- Number of jobs performed by the worker - medium
- Geolocation data - low
- …

Teams involved and means of contact:

- Engineering team: [role, responsibilities and means of contact]
- HR team: [role, responsibilities and means of contact]
- Data Protection team: [role, responsibilities and means of contact]
- …

Human Intervention: The decision taken to terminate a worker's contract can only be taken by a human based on their interpretation of the information provided by the algorithm. There is no human involvement in the algorithm's decisions to flag a worker to the human decision-maker.

Staff allocated: 10 people with specific training

Decision overrun by human review: 10% in 2022

First deployed: 25/10/2021

Last major update: 14/11/2023

Impact assessment: Completed on 01/02/2022

Engagement with workers and workers representative: Yes. Met with Union 1 between November and December 2021. Circulated survey to workers during November and December 2021.

2. Accompany all algorithmic decisions with an explanation of the most important reasons and/or parameter(s) behind the decision and how they can be challenged

Any company using algorithms should accompany all management decisions with a statement of how they were made (for example, 'fully-automated', 'algorithm-supported' or 'no algorithmic involvement'). When a human has been involved in a decision, it must state who has done what, when they did it, and on what information their decision was based. Where decisions have relied on algorithms, workers should be notified of which algorithm it was, including a link to its description in the public registry, and how to ask for a human review.

A reason for the decision should always be made available to the worker, including with reference to the inputs, including worker personal data, and parameters that were decisive to the outcome or that, if changed, would have resulted in a different outcome. Sources of particular parameters and inputs must also be provided and explained – for example in the event that a decision is based on a customer feedback rating. Reasons given for a particular decision must be specific and personalised rather than wholly generic and should not be provided in overly technical language (for example, stating that 'on this date you were expected to make X deliveries, but you only made Y', rather than 'your deliveries are slower than expected').

Where an algorithmically generated score was used as a parameters in relation to a decision – companies should provide workers with its distribution ratio of the score – i.e. what percentage of workers fall into the same category within a given geographic area (for example the city in which the worker is operating). The purpose of this is to provide the context behind decision-making, which would in turn uphold algorithmic accountability and enable workers to challenge inaccurate parameters and inputs. This information could be provided through an aggregated percentage of workers with a certain score or rating. Given that this information is likely to change over time, the ratio should be provided to workers at the time that they face a particular decision, such as termination. The company should also provide information about the prevalence of any issues that prompted a particular automated output at the time of the decision. For example, if a company flags a worker on suspicion of GPS spoofing and suspends his account as a result – it should provide information on whether other workers also experienced technical issues relating to the app's collection of locational data.

The purpose of this is not just to address information asymmetry and allow decisions to be challenged, but also to allow workers to understand why they are being treated a certain way and what changes they can make to get a better outcome. This doesn't necessarily mean going into the details of the algorithm, but rather providing insight into what change(s) a worker could make to receive a more desirable outcome in the future.

A worker should be able to challenge any decision they think is wrong or unfair. Contact details of a human must be provided for this, as well as information on how to request a review and which teams have what oversight over the algorithm's outputs.

Demand 2 Examples

Case Scenario 1:

A driver is refused access to their account after taking a photo of themselves following a prompt by the platform. The driver should be provided with the name of the algorithm that assessed the photo, and the key parameters that led to this decision, for example: what existing data the photo was compared against, match percentage and metadata (such as device maker and model).

In this case scenario, they also should be informed about what match percentage is required to provide authorisation, and their historical match percentage rate. The average match percentage rate of access attempts approved by the algorithm could also be provided.

The worker should be given the possibility to immediately contest this decision and have it reviewed by a human or be provided with an alternative way to verify their identity.

Case Scenario 2:

A courier is notified that their account has been de-activated. The notification should provide all the relevant information that led to this decision, for example that their conduct has been flagged and reported by multiple customers following an identified pattern and that a human staff member acted on the report. The courier should be provided with the relevant data, including number of reports, time, date, general reason for this report, their risk score (where applicable) and the percentage of other workers falling in the same risk category (where applicable.

The notification must also identify any review and oversight team(s) involved in the de-activation determination. This information should also include the relative seniority and job titles of any human agents involved in the decision as well as the length of time taken to review the de-activation decision by each respective team(s) and an explanation of any escalation process between teams (if applicable).

Within the notification, or easily accessible, should be a mean to contest this decision with adequate inputs allowing the worker to provide additional information.

Case Scenario 3: (poor practice)

A driver is assigned a 'medium risk rating' by a fraud detection algorithm after 4 failed trips. Following a further 3 failed trips, he is elevated to a 'high risk' and informed via SMS that he is above the local fraud threshold. The notification includes the percentage of workers also falling in that category.

3. Allow workers, their representatives and public interest groups to test how the algorithms work

To allow workers to more deeply understand the impact on those affected by algorithms, companies should make a version of their algorithm available for testing. This could be done by providing API access to a sandboxed version of the system, making open source the key algorithms used on the platform, or providing access to anonymised/synthetic data accurately reflecting the behaviour of the system. Companies should also consider sharing their source code and training datasets directly to further improve transparency and accountability.

While public access would be a gold standard, a more limited approach may be appropriate, in which only worker representative bodies or recognised academic or civil society organisations can access the testbeds, potentially via collective agreements or licensed access. The ability for workers to test how the algorithms work cannot be confined to a single instance and instead there should be an ability to repeatedly trial the system.

Adequate documentation should be provided to make use of these resources. Different algorithms may require different processes: no specific auditing regime should be mandated as the focus should be on providing a pragmatic, flexible and effective way of allowing engagement with how the algorithms work in practice.

Demand 3 Examples

Case Scenario 1:

A company offers a taxi driver and passenger matching service. To allow drivers and their representatives to understand how the matching algorithm functions, they provide a sandboxed version of the algorithm over an API. The sandboxed version allows authenticated users to simulate client requests and the location of drivers, and evaluate how these demands propagate to drivers, including how this information is presented to each driver individually.

The API allow users to import data for easy testing of complex scenarios and allows export of data to analyse how the algorithm reacts to different conditions. The company provide a feedback mechanism for API users to report bugs or issues.

To enable better auditing of the algorithm, the company provides access to the three latest version of the system so that users can compare results.

Individuals and organisations must apply to be authorised to access the API and must sign a contract before doing so.