Chapter 8
Implementing employee interest along the Machine Learning Pipeline[1]

Lukas Hondrich and Anne Mollen

1. Introduction: the employee interests in algorithmic management

Especially with increasing work from home constellations during and since the start of the Coronavirus pandemic, discussions about workplace surveillance and algorithmic management have reached wider public attention – through academic research (Aloisi and De Stefano 2022; Jarrahi et al. 2021), news media reporting, mishaps by major algorithmic management software, and algorithmic management practices that may violate national legislations.[2] Concerns are that algorithmic management could potentially stifle human autonomy (Prunkl 2022), exacerbate power inequalities and reinforce historical biases and forms of discrimination (Barocas et al. 2017; Noble 2018), while evading established forms of oversight and worker participation (Degryse 2017; Cefaliello and Kullmann 2022).

Labour organisations and employee representatives were engaged with the impact of increasing automation in the workplace even before the pandemic. However, their engagement with automation and algorithms has, so far, operated on a quite abstract level of how to deal with automation in a workplace, even though some more concrete guidelines and results, such as collective bargaining agreements, are slowly emerging (AlgorithmWatch 2023).

Additionally, current proposals for regulating AI systems, and perhaps specifically AI systems in the workplace, are focused especially on risk mitigation strategies. The European Union's AI Act, for instance, follows a risk-based approach. It is worth recognising that AI systems in a work context, as used for example for recruitment, advertising vacancies, screening or filtering applications and evaluating candidates, as well as in promotion and termination matters, task allocation, and for monitoring and evaluating performance and the behaviour of employees, are being recognised as posing

[1] Editor‘s note: the authors prefer to capitalise both Artificial Intelligence as well as Machine Learning (and, in this context, also Pipeline) as a way of distancing themselves from the terminology to describe what would be more correctly labeled as "statistical pattern recognition" and as a form of preventing the anthropomorphizing terminology from normalizing, while preserving readability.

[2] See for instance the *New York Times* articles on workplace surveillance and algorithmic management https://www.nytimes.com/interactive/2022/08/14/business/worker-productivity-tracking.html (published 14 August 2022); the privacy violations by Microsoft 365 https://www.theguardian.com/technology/2020/dec/02/microsoft-apologises-productivity-score-critics-derided-workplace-surveillance (published 2 December 2020); and a publication by AlgorithmWatch pointing out that in Germany, without the individual consent of employees or a company-wide agreement, the use of People Analytics systems might be illegal https://algorithmwatch.org/de/auto-hr/positionspapier/ (published 27 February 2022).

possibly high risks for workers (European Commission 2021). But risk mitigation strategies cannot be considered an adequate response from an employee perspective. That is why the AI Act can only be understood as a baseline protection that will prevent the most dangerous systems – from a fundamental rights perspective – from entering the European market or only with safeguards in place.

In its current form the AI Act does not, for example, sufficiently address the opacity of algorithmic management systems. The AI Act will not enable employers, employees and their representatives to gain more knowledge on how an algorithmic management system executes its decision-making. Such knowledge would, however, be necessary for employee representatives to move beyond risk mitigation. Their ambition should be not only to limit the risks for employees but to shape algorithmic management systems actively in their interests.

Algorithmic management systems are software-based systems that are used to replace or support typical tasks in workforce management. They can entail descriptive, predictive and prescriptive elements, for instance visualising data about employees (descriptive), making assumptions (predictive) or taking decisions (prescriptive) about employees (Gießler 2021). As these systems can be used to evaluate employees’ work performance, allocate tasks, suggest promotions or even terminate contracts, it simply cannot suffice to establish safeguards against the major risks associated with algorithmic management. Due to their potentially wide-ranging implementation, employees must have a say in how algorithmic management systems take their automated decisions. With algorithmic management systems often remaining ‘black boxes’ that allow few insights – even for employers or people in HR departments – this question is not trivial.

This chapter proposes that employee representatives make use of the concept of the Machine Learning Pipeline as a tool to help them in establishing and implementing employee interests when it comes to individual algorithmic management systems.

2. Identifying the spaces for worker action along the Machine Learning Pipeline

Artificial Intelligence remains a nebulous and hard to define term; more precise ones include algorithmic or automated decision-making (ADM) systems.

On a more technical level it makes sense to differentiate between rule-based algorithms, in which decision rules are explicitly stated and are thus readable by humans, and data-driven Machine Learning methods, in which rules are represented in complex mathematical functions, making them highly expressive but usually non-readable by humans.

These algorithms are labelled ‘data-driven’ as they learn directly from datasets and are thus especially susceptible to the unbeknown proliferating biases that may be present within them. To this group the lately popularised ‘deep learning’ and ‘neural network’ algorithms belong.

When these models are applied to unseen data, as is their purpose, they pose risks for the people affected. This is because they frequently lack the robustness of training data; that is, the datasets they have been trained on diverge from the inference data (i.e. the data they are applied to). At the same time, their complexity, resulting opaqueness and illegibility make it hard to foresee in which cases they will fail (Rudin 2019).

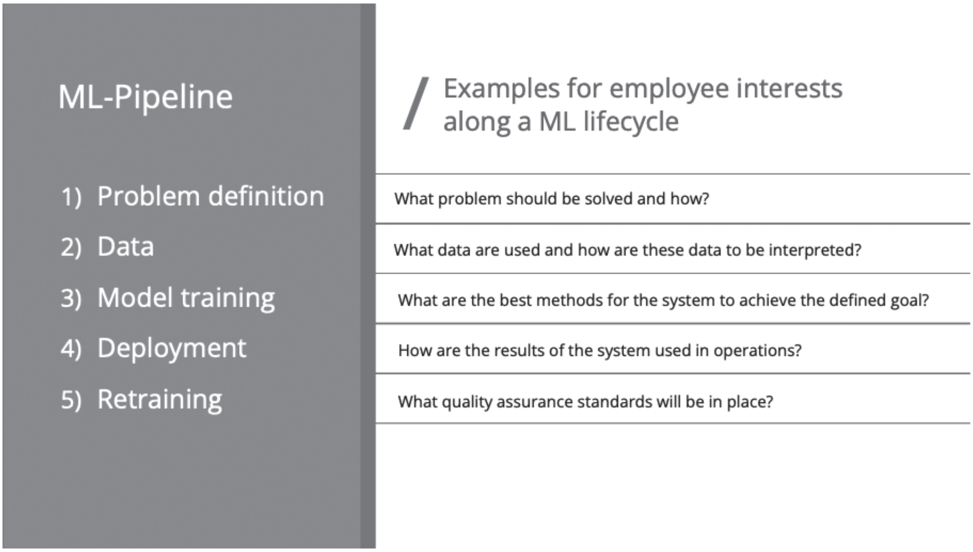
In the same way as choice of algorithm might affect transparency and robustness, other risks can be attributed to respective steps in the planning process or to specific technical components (Suresh and Guttag 2021). Because of the data-driven learning process, Machine Learning systems are, for instance, also discussed for their opacity and for the risk that not even their own developers know how they operate and make their decisions. While it might be true that the developers behind Machine Learning models might not know how exactly their models produce individual outcomes, it is important to note that – given the resources – explanations can be obtained, safeguards implemented and the interests of the people affected accommodated.

Narrowing down where exactly the risks in an ADM system are rooted can help developers, the people affected or any other stakeholder address them – either by shaping the technical components or by introducing specific organisational safeguards. Mapping these risks to the technical level of the Machine Learning Pipeline and addressing them there is thus a course of action worth exploring.

The concept of the Machine Learning Pipeline allows a dissection, along the lifecycle of a Machine Learning model, of how employee interests can potentially be integrated when an ADM system for algorithmic management purposes is developed and implemented in a work context. These reflections demonstrate what role employee representatives can have, at a conceptual level, in relation to ADM systems in the workplace. They sketch an ideal scenario which, until now, may well not be easy to implement in practice due to a lack of transparency and experience, as well as the regulatory frameworks in place to strengthen employee interests being inadequate. But they also give hands-on suggestions on what employee representatives should be considering when it comes to collective agreements, co-determination processes or regulatory proposals on ADM systems in the workplace. They are therefore a sound starting point for future discussions that will focus on the feasibility and practical implementation of integrating employee interests into ADM systems in the workplace.

The Machine Learning Pipeline describes a common lifecycle of a Machine Learning-based ADM system (for a discussion on bias in this respect, see Schelter and Stoyanovich 2020; Suresh and Guttag 2021). It usually differentiates five consecutive steps (see Figure 1).

Figure 1 **Overview of the stages of the Machine Learning Pipeline and possible questions for employees to address**



ML-Pipeline	Examples for employee interests along a ML lifecycle
1) Problem definition	What problem should be solved and how?
2) Data	What data are used and how are these data to be interpreted?
3) Model training	What are the best methods for the system to achieve the defined goal?
4) Deployment	How are the results of the system used in operations?
5) Retraining	What quality assurance standards will be in place?

Each of these steps is essential in defining what purpose a Machine Learning model should be serving (its objectives); how it will reach its results (its methods); if the data used is suitable for the defined objectives; and what safeguards and monitoring procedures are implemented. Many decisions are taken during these steps, with each one having a possibly decisive influence on how the overall ADM system will make its decisions and exercise influence on the people affected.

It thus becomes very clear that employee representatives should be involved in this process in order to fulfil their mandate. The following subsections show what role employee representatives can take regarding the five steps of the Machine Learning Pipeline.

2.1 Problem definition

Even though the problem definition phase is not necessarily considered part of the technical process of an ADM system, it is essential for employees to be involved. It is here that the objective and the purpose of an ADM system is defined. Moreover, it is during this stage that the question of how the ADM system is integrated into the organisational context – for instance if it is supposed to support human decision-making or might work in completely autonomous ways – is discussed and decided.

The introduction of ADM systems in an organisation cannot, in most cases, be considered an isolated incident, but it is mostly accompanied by organisational restructuring and, as a part of that, long-term power shifts (Degryse 2017). When an organisation introduces an ADM system for automating parts of the internal and external hiring process, valuable knowledge on hiring procedures might, for instance, become lost to employees after being centralised within the ADM system and among the people working with it. The introduction of an ADM system for an internal hiring process might then have negative consequences for the negotiating power of employees. Also, employees can have an important oversight function regarding the area of application that an ADM system was originally defined for and the areas in which it might subsequently become used. It is not unusual for ADM systems to be designed as data-driven projects for predicting an outcome (for which correlation might be sufficient), but at later stages become used as predictive models (for which a causal model was required). That is why it is essential that employee representatives should be able to influence and be heard regarding these fundamental questions.

2.2 Data

An ADM system based on a Machine Learning model is trained on data. When developing and planning an ADM system it is thus essential to define what datasets can best reach the objectives defined in the previous problem definition phase and how such datasets can be generated (Holstein et al. 2019). The decisions taken on data selection, data collection and the related privacy protection questions are highly relevant for employees – especially with wide-reaching workplace surveillance practices already in place (Christl 2021) and the presence of many existing biased datasets that could potentially have discriminatory effects. Employee representatives should ensure that data selection and collection evolve with employee interests in mind. This task cannot be achieved without more transparency instruments in place. Data sheets or data cards (Gebru et al. 2021), that ideally provide encompassing documentation of the datasets used, could potentially be very helpful to employee representatives in assessing the suitability and quality of the data used.

Further, what key constructs are going to be used for a system's automated decision-making will be an important decision to take. Employee representatives need to be involved in operationalising the relevant criteria driving a system's decision-making. Considering biased datasets, it might for instance be important for employee representatives to use fairness metrics in order to establish safeguards against discrimination. But the involvement of the people affected is also essential in the light of the need to interpret the collected data adequately. If the number of keystrokes by employees is being used as a criterion to evaluate performance, or the quantity of messages sent between co-workers during a day are considered relevant aspects for assessing productivity, the context on which such data is founded needs to be provided. If workers are sitting opposite each other, a lack of messages sent between them needs to be evaluated differently. Also, periods without keystrokes might point towards off-screen tasks which might be absolutely fine in a given working constellation. The people affected have the relevant domain knowledge to provide similar and possibly much deeper context knowledge on the data being collected and the key criteria that should be used for a system's decision-making.

2.3 Model training

In the model training phase, a Machine Learning model extracts rules, statistical patterns and links between data points in the training data (Barocas et al. 2017). The outcome is a mathematical function, the Machine-learned model, on the basis of which the system will generate output. This mathematical function can be more or less opaque and more or less difficult to understand (Rudin 2019). At this stage, employee representatives need to establish safeguards that guarantee there are no harmful biases in the Machine-learned model. Further, they need to ensure that the model training procedure leads to a model that bases its decision-making on patterns in the data that align with employee interests.

One common concern in this regard is that a model might establish patterns that are both useful and harmful, with these not always being easy to separate (Zhang et al. 2018; Zhao and Gordon 2022). Another concern relates to the complexity and opacity of a model – where decisions possibly have to be taken between the better performance of a system or a higher level of transparency. Here, employee representatives can advocate methods that potentially provide greater insights into the systems.

2.4 Deployment

During the deployment phase employee representatives need to make sure that the Machine Learning model does not develop any unwanted tendencies in its decision-making (Suresh and Guttag 2021). In this phase, the objectives defined in the problem definition phase are put into practice for the first time. At this point, the model has learned purely based on training data but, in the deployment phase, the system will be integrated into a software environment that is likely already to exist, and will process real world data and thereby generate outputs. Employee representatives need to be alerted to Machine Learning models typically experiencing a drop in performance when being confronted with real world data; special scrutiny by the people affected by, but also the people working with, these systems is thus essential.

In addition, sufficient feedback loops and mechanisms should be implemented so that feedback can actually have an impact on the systems in question. Next to a focus on the technical system, organisational structures again become more relevant. Equally, it will be essential to monitor how output by the ADM system will be integrated into organisational decision-making processes. Again, power imbalances and bias might manifest themselves, for instance if HR staff act only selectively on the decisions taken by an ADM system. Establishing clear guidelines on how to use the output generated by an ADM system might be helpful in that regard.

2.5 Retraining

A Machine Learning model needs to be maintained. Retraining as a form of maintenance should ensure that the model continues to serve the objectives originally designed for the system. This is necessary because the model might potentially encounter unexpected data and because it is being integrated into a complex sociotechnical system that might develop unanticipated dynamics – for instance, slight shifts between the data that the system encounters in the real world and the data for which it was optimised in the training phase. That is why models are often retrained with more current data (Huyen 2022).

Here, employee representatives again need to exercise oversight because retraining introduces a number of new challenges. One example is that retraining might lead to the ADM system producing self-fulfilling prophecies; that is, emergent bias (Stoyanovich 2020; Barocas et al. 2017). The reason is that the new data on which the model is trained has been produced by the system itself. Thus, the patterns along which the model makes its decisions has influenced the data on which it is being retrained. This effect can accumulate and could, for instance, lead to certain groups of employees being preferentially treated in job matching decisions.

Employee oversight thus does not stop with the deployment of ADM systems but should continue once systems become established in their organisational contexts. Due to the complexity of the oversight tasks, employee representatives will also need external support by Machine Learning experts when it comes to executing oversight on a technical level. Next to the oversight and control mechanisms for employee representatives there should also be redress mechanisms established for people who are being affected by Machine Learning models that may deteriorate.

3. Capacity building for employee representatives

So far, the discourse around ADM systems in general, but perhaps specifically regarding ADM systems being used in a work context, has focused on the risks associated with them and how these may be mitigated. Indeed, there are many risks associated with ADM systems especially in a work context where there is already a power imbalance between employees and employers, and where decisions by ADM systems can have a huge influence on people's livelihoods and wellbeing. Exactly because of the immense impact that ADM systems may potentially have on employees when it comes to a person's recruitment, their salary, their everyday working conditions etc., it cannot be considered sufficient that employees and their representatives mitigate such risks. Instead, they should be shaping these systems according to their interests. The ambition should also be for employees to profit from the potential benefits of these systems.

In this regard, this chapter presents the concept of the Machine Learning Pipeline as an attempt to demystify ADM systems and Artificial Intelligence. Often the technology is being perceived as having almost unprecedented magical capabilities (Campolo and Crawford 2020). This narrative is not only inaccurate; it builds up barriers for stakeholders to perceive Artificial Intelligence as something they can potentially co-create and co-govern. Of course, coming up with a Machine Learning Pipeline for a to-be-developed ADM system is equally, from an employee perspective, not an easy task. But it is something that can be achieved: with support from Machine Learning experts on the outside; but also by training employee representatives to be aware of the potential pitfalls and to be able to ask the right questions about Machine Learning models.

References

AlgorithmWatch (2023) Algorithmic transparency and accountability in the world of work: A mapping study into the activities of trade unions. https://www.ituc-csi.org/IMG/pdf/2023_aw_ituc_report_final.pdf

Aloisi A. and De Stefano V. (2022) Your boss is an algorithm: Artificial intelligence, platform work and labour, Bloomsbury.

Barocas S., Hardt M. and Narayanan A. (2017) Feedback and feedback loops, in Fairness and machine learning: Limitations and opportunities, 13–15. https://fairmlbook.org/pdf/fairmlbook.pdf

Campolo A. and Crawford K. (2020) Enchanted determinism: Power without responsibility in artificial intelligence, Engaging Science, Technology, and Society (6), 1–19. https://doi.org/10.17351/ests2020.277

Cefaliello A. and Kullmann M. (2022) Offering false security: How the draft artificial intelligence act undermines fundamental workers' rights, European Labour Law Journal, 13 (4), 542–562. https://doi.org/10.1177/20319525221114474

Christl W. (2021) Digitale Überwachung und Kontrolle am Arbeitsplatz. https://crackedlabs.org/daten-arbeitsplatz

Degryse C. (2017) Shaping the world of work in the digital economy, Foresight Brief 01, ETUI. https://www.etui.org/publications/foresight-briefs/shaping-the-world-of-work-in-the-digital-economy

European Commission (2021) Annexes to the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial intelligence act) and amending certain Union legislative acts, COM(2021) 206 final, 24.4.2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Gebru et al. (2021) Datasheets for datasets, Communications of the ACM, 64 (12), 86–92. https://doi.org/10.1145/3458723

Gießler S. (2021) Was ist automatisiertes Personalmanagement, AlgorithmWatch. https://algorithmwatch.org/de/wp-content/uploads/2021/05/Was-ist-automatisiertes-Personalmanagement-Giesler-AlgorithmWatch-2021.pdf

Holstein K., Wortman Vaughan, J., Daumé H., Dudík M. and Wallach H. (2019) Improving fairness in machine learning systems: What do industry practitioners need?, International Conference on Human Factors in Computing Systems, Paper 600, 1–16. https://doi.org/10.1145/3290605.3300830

Huyen C. (2022) Designing machine learning systems, O'Reilly Media.

Jarrahi M.H., Newlands G., Lee M.K., Wolf C.T., Kinder E. and Sutherland W. (2021) Algorithmic management in a work context, Big Data and Society. https://doi.org/10.1177/20539517211020332

Noble S.U. (2018) Algorithms of oppression: How search engines reinforce racism, NYU Press.

Prunkl C. (2022) Human autonomy in the age of artificial intelligence, Nature Machine Intelligence, (4), 99–101. https://doi.org/10.1038/s42256-022-00449-9

Rudin C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, (1), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schelter S. and Stoyanovich J. (2020) Taming technical bias in machine learning pipelines, IEEE Data Engineering Bulletin. https://ssc.io/pdf/taming-technical-bias.pdf

Stoyanovich J., Howe B. and Jagadish H.V. (2020) Responsible data management, Proceedings of the VLDB Endowment, 13 (12), 3474–3488. https://doi.org/10.14778/3415478.3415570

Suresh J. and Guttag J. (2021) A framework for understanding sources of harm throughout the machine learning life cycle, paper presented at EAAMO'21: Equity and Access in Algorithms, Mechanisms, Optimization, New York, October 2021. https://doi.org/10.1145/3465416.3483305

Zhang B.H., Lemoine B. and Mitchell M. (2018) Mitigating unwanted biases with adversarial Learning, ACM Conference on AI, Ethics, and Society, 335–340. https://doi.org/10.1145/3278721.3278779

Zhao H. and Gordon G.J. (2022) Inherent tradeoffs in learning fair representations, Journal of Machine Learning Research, 1–26. https://doi.org/10.48550/arXiv.1906.0838

All links were checked on 29.01.2024.

Cite this chapter: Hondrich L. and Mollen A. (2024) Implementing employee interest along the machine learning pipeline, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.