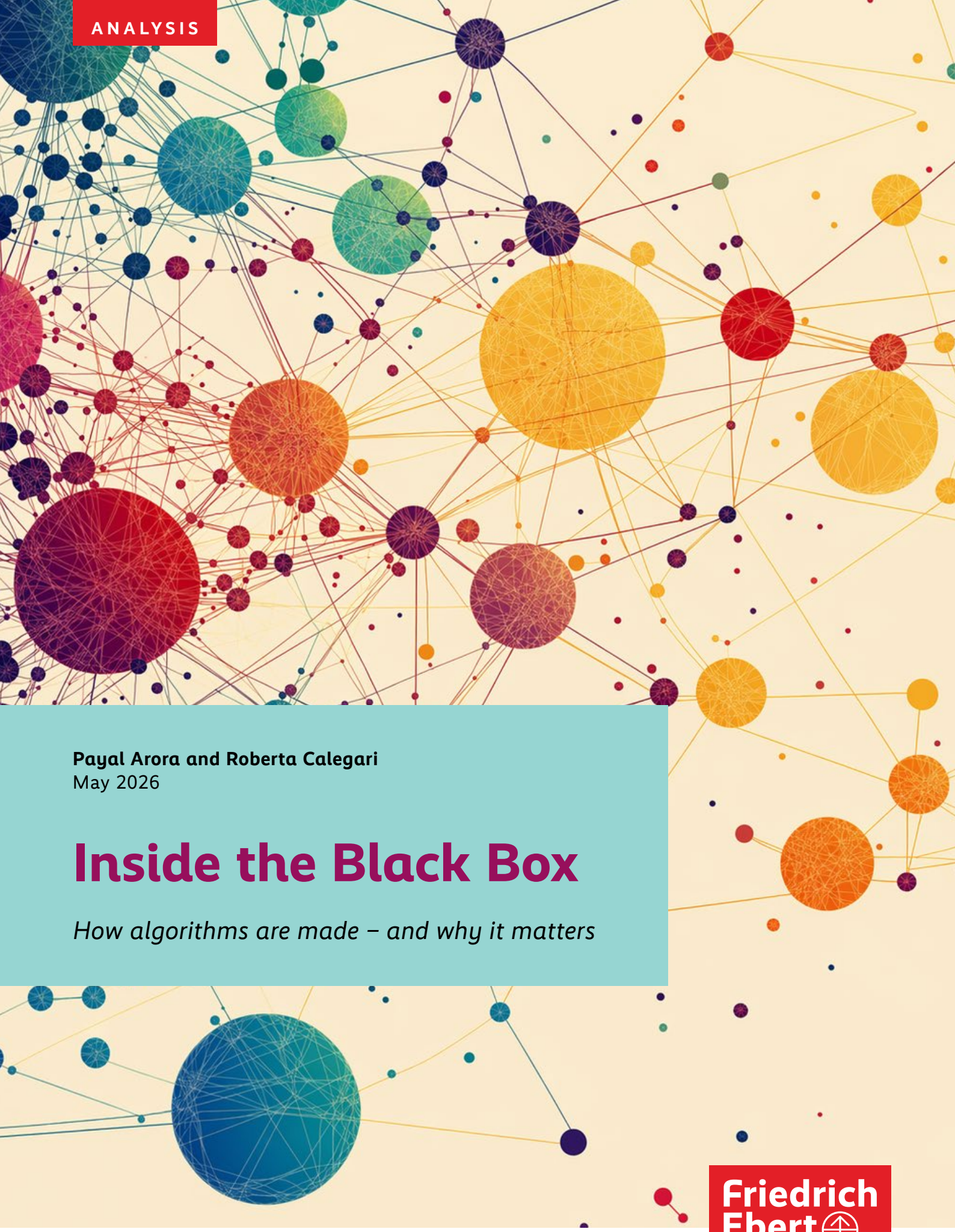
ANALYSIS



Payal Arora and Roberta Calegari
May 2026

Inside the Black Box

How algorithms are made – and why it matters

Friedrich Ebert Stiftung

Competence Centre on the Future of Work

Imprint

Published by
Friedrich-Ebert-Stiftung e.V.
Godesberger Allee 149
53175 Bonn, Germany
info@fes.de

Issuing Department
Competence Centre on the Future of Work
Cours Saint Michel 30a, 1040 Brussels, Belgium

For more information about the Competence Centre
on the Future of Work, please consult:
https://futureofwork.fes.de

Responsibility for Content and Editing
Dr. Inga Sabanova
inga.sabanova@fes.de

Design/Layout
pertext | corporate publishing
www.pertext.de

May 2026

ISBN 978-3-98628-879-2

Further publications of the Friedrich-Ebert-Stiftung can be found here:
↗ www.fes.de/publikationen

Friedrich Ebert Stiftung

Payal Arora and Roberta Calegari
May 2026

Inside the Black Box

How algorithms are made – and why it matters

Contents

Summary

From warehouses to recruitment platforms, algorithms are quietly becoming the new managers of modern work. Increasingly, decisions about hiring, scheduling, productivity and even dismissal are shaped by automated systems that analyse vast amounts of data. These technologies promise efficiency and innovation, but they also raise important questions about fairness, transparency and worker rights. Algorithms are often presented as neutral tools, but they reflect the choices, assumptions and data used to build them. Past data is likely to reflect inequalities – such as gender or racial bias in hiring – and algorithms may reproduce and even amplify such patterns. At the same time, many systems operate as »black boxes«, which makes it difficult for workers, unions or regulators to understand how decisions, which are likely to affect people's livelihoods, are really made.

This paper takes readers inside the process of how algorithms are designed, trained and deployed in workplaces. It highlights the key stages at which bias, exclusion or unfair treatment may enter the system, from definition of the problem to selection of data and testing of models. It also examines the growing importance of algorithmic audits and stronger governance to ensure accountability. The message is clear: algorithms are not destiny. With stronger oversight, worker participation and ethical design, these technologies can support more transparent, equitable and humane workplaces rather than reinforce existing inequalities.

Introduction – the algorithmic turn

Imagine standing on the floor of a massive warehouse, around which hundreds of workers move swiftly under the gaze of screens and scanners. Orders flash across handheld devices, dictating every motion, pick, scan, deliver, repeat. However, there is no visible supervisor shouting instructions. The authority sits elsewhere, embedded in lines of code.

To capture this reality, it is useful to speak not only of »AI« but more broadly of *algorithmic systems*. This includes traditional rule-based software (for example, scheduling tools or scoring systems), as well as machine-learning and foundation models. Algorithms operate across nearly all sectors, from payroll systems and shift schedulers to recruitment platforms, credit scoring tools and content recommendation engines. Not all algorithmic systems are »artificial intelligence«, properly speaking; many rely rather on rule-based software that follows predefined instructions. AI systems are a subset of algorithmic systems, built on algorithms but enhanced with learning techniques that detect patterns and make probabilistic predictions. In short, not every algorithm is AI, but every AI system relies on algorithms. For workers and unions, the key issue is not the label, but the impact of these systems on decisions about work, pay and opportunities.

While AI-driven digital management and monitoring systems promise greater efficiency, flexibility and responsiveness for the workforce, the devil lies in the details of how these systems are designed and governed. Without careful attention to context, consent and accountability, such technologies risk amplifying workplace inequalities and eroding trust (Capasso et al. 2024). Anchoring their deployment in an *ethics of care* (Arora, Raman and König 2023) – one that values relationality, transparency and worker well-being over mere productivity – can ensure that digital oversight empowers rather than disciplines, and that technological progress aligns with the broader vision of humane and inclusive work.

This paper takes readers behind the curtain to trace the technical and human process of algorithm creation. It explains, step by step, how algorithms are designed, trained and deployed, but also how choices at each stage can lead to exclusion or inequality. It explores how audits are conducted, where accountability falters and what can be done to ensure that these systems serve people rather than the other way around.

Part I: the making of an algorithm

At its core, an algorithm is simply a set of instructions, a recipe for solving a problem. But as any cook knows, the outcome depends on what ingredients you use, who wrote the recipe and what they intended it to taste like. Developing an algorithm starts with clearly defining the problem and the desired outcome. The task is broken down into logical steps, specifying inputs, outputs and constraints. These steps are translated into code, tested with representative data and refined to handle errors and improve performance. After validation and optimisation, the algorithm is deployed. This sequence – definition, design, coding, testing and optimisation – forms the practical »recipe« that turns an idea into a functioning system.

In the digital world, the recipe begins with a **problem statement**. A company might say: *We want to predict which job applicants will perform best, or we want to optimise delivery times.* From there, data scientists gather information to train a model, such as past employee records, CVs, performance metrics and customer ratings. The system then learns patterns, identifying what kinds of people have done well in the past.

The trouble starts here: if the past reflects inequality, the model will faithfully reproduce it. This is known as **historical bias**. For instance, when Amazon developed a hiring algorithm in 2018 (Pathak 2025), it trained the system on ten years of past resumes, most of which

came from men. The algorithm »learned« that being male correlated with success and began penalising resumes containing the word »women«. Amazon had to scrap the project.

Fast forward to 2025 and we find the gender bias in AI-driven recruitment platforms remains a persistent problem, as algorithms trained on historical hiring data often replicate discriminatory patterns, prioritising male-coded language in résumés (Bhatia et al. 2024), ranking women lower for technical roles, or filtering out candidates based on gaps in employment that in fact reflect caregiving responsibilities rather than competence. The European Institute for Gender Equality (EIGE) report (2021) on platform-work and AI in the labour market finds gendered patterns of risk in which women are more likely to be found in less secure, lower-paid or platform-type work.

To make this more tangible, think of training an algorithm as teaching a child. If you show the child only one kind of book – say, fairy tales in which only princes become heroes – the child will »learn« that only men can be brave. That's how bias takes root: not through malicious intent, but through skewed learning.

Next come **data cleaning and labelling**, the stage at which information is tidied and annotated so the algorithm can make sense of it. In facial recognition systems, for example, developers label thousands of faces with identities. If the dataset includes mainly white faces, as was the case in early commercial systems, the model will struggle to recognise darker skin tones. This is why, as Joy Buolamwini's MIT research (2024) showed, commercial facial recognition systems misidentified Black women up to 35 per cent of the time, while error rates for white men were below 1 per cent.

A report from the EU Agency for Fundamental Rights (2023) reveals that people of African descent still face widespread discrimination across Europe, and warns that the absence of reliable race and ethnicity data is surreptitiously baking bias into the very algorithms and data-cleaning systems shaping everyday decisions. What is not measured, or is measured poorly, tends to be misrepresented or ignored by the model.

Finally, the algorithm is **trained**, fed data until it can make reliable predictions. Developers test it using new data, refine it and deploy it in real-world systems. The output of the training is the model that will make, or heavily influence, concrete decisions about jobs, pay, promotion prospects or access to social services. A model is a mathematical representation of patterns in data, and the choice of which model to use (for example, a decision tree, regression model or neural network) depends on the type of problem being solved, the nature and size of the available data, and the trade-offs between accuracy, interpretability and computational complexity.

Each stage involves choices: what to optimise for (accuracy, speed, profit), what to ignore (outliers, minority data), and when to stop training. These seemingly technical decisions are, in fact, profoundly ethical and political. They determine whose errors are tolerated, whose interests are prioritised, and whose experience is treated as statistical noise. For workers, this is the first critical moment at which social conflict and inequality can become silently translated into code.

To better understand where these risks emerge, it is useful to look more closely at how AI and algorithmic systems are built. In practice, their development can be understood in terms of a **four-stage process**: problem definition, data preparation, model training, and validation and deployment.

(i) Definition – Framing the Problem

Every algorithm begins with a question: *What are we trying to predict or optimise?* For instance, Uber's dispatch and pricing algorithms optimise for efficiency, in the sense of getting a passenger to a driver as quickly as possible. But they do not necessarily optimise for equity, safety or driver well-being. That framing shapes everything that follows.

The same is true in workplaces. A retailer might introduce a scheduling tool to minimise labour costs per hour, or an HR department might deploy a scoring model to rank candidates by predicted performance. If the objective is defined purely in terms of speed or cost, the system will not »see« stability of income, work–life balance or non-discrimination as goals worth protecting. Those values have to be deliberately built into the definition of the problem.

(ii) Data preparation – feeding the system

Algorithms are data-hungry. They consume massive datasets, including photos, resumes, transaction records. But data is never neutral. Who collected it? When? With what tools? Under what conditions? And whose experiences were left out altogether? In small towns and villages, for example, limited infrastructure and unequal access to digital technologies mean that entire populations may be underrepresented or remain invisible in datasets built on city-centric sources. When such data is used for training, the resulting models inherit and amplify these blind spots.

If training data for a welfare risk-scoring system or an employability algorithm over-represent certain neighbourhoods, age groups or migrant communities in »risk« categories, those associations can become self-reinforcing. Choices made during data cleaning and labelling – what is treated as an error, what is dropped as an »out-

lier«, which proxies (for example, postcode or school) are allowed – directly shape who will later be flagged as a problem, a risk or a »low potential« worker.

(iii) Modelling – training the algorithm

Developers use mathematical techniques to detect patterns in the data and apply learning algorithms to create a trained model. Some algorithms are simple (such as decision trees or logistic regression), others extremely complex (such as deep neural networks). The process involves balancing multiple objectives: accuracy, transparency and interpretability, computational cost and fairness. Each choice – for example, prioritising predictive precision over interpretability, or excluding sensitive attributes but keeping highly correlated proxies – influences how the system will behave when deployed. A highly accurate but opaque model may be attractive from a technical standpoint, but very difficult for workers, unions or inspectors to scrutinise, even when it has a direct impact on hiring, promotion or dismissal.

(iv) Validation and deployment – testing the real world

Before launching, teams test algorithms on new data. But testing environments are often controlled and so may fail to capture the messiness of reality. Once deployed, models may encounter strikes, pandemics, new management strategies, changes in legislation and shifting worker behaviour. As data patterns change over time – a phenomenon known as **»concept drift«** (Tsymbal 2004) – models may become less accurate or more biased. An algorithm trained on pre-pandemic traffic and commuting data, for instance, may struggle to adapt to post-pandemic mobility habits and hybrid work patterns.

A metaphor may be helpful here. Designing an AI system is like building a self-driving car. You can perfect it on closed tracks, but once it's deployed on real roads – complete with rain, potholes, pedestrians and so on – it will encounter things it was never trained for. In the workplace, this means that even a system that looked acceptable in a pilot can, over time, create new forms of pressure, discrimination or exclusion. That is why validation and deployment should not be seen as the end of the process, but as the beginning of a continuous cycle of monitoring, revision and, where necessary, contestation by workers and their representatives.

Part II: the teams behind the machines

The myth of the algorithm is that it is neutral; a pure expression of logic, untouched by human bias. In reality, every algorithm carries the fingerprints of its makers. Behind each system stands a team of engineers, data scientists and product managers, often drawn from similar backgrounds: young, male and educated in elite institutions in Silicon Valley, Beijing, Bangalore, Berlin, London or Paris. Their worldviews, cultural contexts and incentives shape the design choices they make, influence how data is selected, what objectives are optimised and which trade-offs are accepted.

As researcher Ruha Benjamin (2019) puts it, »the default settings of technology reflect the default settings of society«. If the design team does not include people from the populations affected – such as women, racial minorities or low-income groups – the system is likely to miss or misinterpret their realities. Diversity in design is not only ethical, but an essential safeguard against blind spots in how data and outcomes are defined.

Take the case of Google's health AI-based retinal screening system (Talby 2020), which aimed to predict diabetic retinopathy. The model worked well in lab conditions but failed in clinics across Thailand. Why? Because the training process assumed ideal conditions – perfect lighting, high-resolution images and stable internet connections – conditions that didn't exist in rural hospitals. The engineers had optimised the model for a world that resembled their offices, not the places it would actually serve.

Similar issues arise in European workplaces when »off-the-shelf« AI tools developed elsewhere are imported into very different legal, cultural and organisational environments. A scheduling or performance-scoring system designed for a US warehouse, for example, may embody assumptions about working time, unionisation or productivity targets that collide with European labour standards and collective agreements.

Ultimately, algorithms do not mirror only data, but also the decisions, assumptions and priorities of those who build them. Design teams, like societies, mirror power structures. A lack of diversity is not just a moral issue, but a technical flaw. Homogeneous teams make homogeneous models, and unless workers and their representatives have a seat at the table when these systems are specified, procured and adapted, their interests risk being systematically left out of the design.

Part III: designing for purpose or profit

It is also important to point out that not all algorithmic harm is accidental. Sometimes, AI and algorithmic systems are **intentionally designed** to privilege certain interests over others.

Consider social media platforms. Their recommendation algorithms are not neutral content sorters; they are engineered to maximise engagement, because engagement drives ad revenue (Bhadani 2021). The underlying

algorithms are trained to predict what will keep each user active for the longest time, continually adapting to individual behaviour. That's why outrage, misinformation and sensationalism spread faster; they keep users scrolling (Geers et al. 2024). In this sense, the system is not malfunctioning, but doing precisely what it was designed to do, even if the broader social consequences are toxic.

In the world of work, algorithmic management systems can be tuned to extract maximum productivity. Gig platforms such as Deliveroo or Uber use real-time data and optimisation algorithms to allocate jobs, monitor performance and adjust pricing, all automatically (Zhu et al. 2024). These mechanisms decide who gets the next task, how long they have to complete it and whether their account will remain active. For many workers, the model effectively functions as both supervisor and evaluator, a digital boss that enforces rules without negotiation. Drivers rarely know how their »rating« is calculated or how fares are set, because the decision logic is embedded deep within proprietary systems. The algorithm acts as both boss and judge.

Similar logics appear in more traditional workplaces. Warehouse workers can find their performance broken down into seconds per task, with dashboards that rank them against colleagues and trigger automated warnings when they fall behind (Claburn 2025). Call-centre agents may be scored on call length, script adherence and sales conversion in real time, with little scope to challenge how these metrics are defined. In such cases, the optimisation goal – maximising output per unit of time – has been hard-wired into the system, while concerns such as health and safety, autonomy or the right to organise are treated, at best, as secondary constraints.

These systems can also be **strategically opaque** (Langer and König 2023). When an algorithm decides who gets a shift or bonus, or who is first in line for redundancy, the company may claim that the system is too complex to explain. This obscures accountability and makes contestation nearly impossible even where EU law formally grants workers' rights to information and explanation.

In other cases, exclusion is built in by design. For example, predictive policing tools (Ziosi and Pruss 2024) such as *PredPol* rely on historical arrest data, datasets that overrepresent minority communities because of existing systemic racial bias in past policing. The algorithm then directs more patrols to those areas, reinforcing the same bias it inherited. The result is a feedback loop: biased data produces biased predictions, which in turn generate more biased data.

Employment and welfare systems can create similar loops (Considine et al. 2022). A risk-scoring model that classifies certain neighbourhoods or demographic groups as »high risk« for fraud or »low employability« will channel more inspections or fewer resources towards them, making it harder for individuals in those groups to improve their situation and confirming, over time, the system's original assumptions.

In each case, the algorithm does not act alone; it operationalises and amplifies the intentions, incentives, inequalities and power structures of the humans who built and deployed it. Recognising this is crucial for trade unions and regulators: changing outcomes is not just a matter of »fixing the code«, but of questioning the objectives, business models and governance arrangements that those systems are designed to serve.

Part IV: auditing the black box

As awareness grows, so does demand for **algorithmic audits**, processes that evaluate whether systems are fair, transparent and accountable.

In theory, auditing an algorithm is like inspecting a financial statement: an independent party reviews inputs, processes and outputs. In practice, it's far trickier. Most large-scale models are proprietary and protected under trade secrecy or intellectual property laws. Companies often argue that disclosing their internal mechanisms would compromise competitiveness or security.

There are three main types of audit (Metaxa et al. 2021):

(i) **Code audits** – reviewing the source code and model parameters. These are rare because few companies grant access.

(ii) **Outcome audits** – analysing how decisions affect different groups (for example, gender, race, region);

(iii) **Process audits** – evaluating governance, documentation and design decisions.

There is growing recognition that meaningful auditing requires strong, enforceable rights to information and explanation (Kerckhofs 2025). Transparency cannot be reduced to a generic notice that »an algorithm is in use«: affected individuals need to understand what the system does in practice (for example, whether it allocates tasks, scores performance, monitors productivity or supports disciplinary decisions), which data it collects and processes, who can access those data (including external providers), and which key factors drive its decisions. Where adverse outcomes occur – such as dismissal, demotion, pay reduction, or loss of work opportunities – clear, reasoned explanations must be provided and made open to challenge. The same standard should apply in recruitment contexts, particularly where high-risk AI systems determine who is shortlisted or invited to interview.

When researchers conducted an outcome audit of Twitter's photo-cropping algorithm in 2021 (Li et al. 2023) they discovered that it consistently favoured white faces in previews. Twitter responded by retiring the algorithm altogether. That's an example of a successful outcome audit leading to reform. However, audits face real limits. Complex AI models such as GPT or Google's ranking systems are too vast, continuously updated and partly data-driven to be fully »opened up« or replicated. Even when companies release transparency reports, they often disclose only high-level summaries. The result is what scholars call **»transparency theatre«** (Cellard 2024), which merely gestures towards accountability without enabling independent verification. To avoid such »transparency theatre«, these rights must be coupled with **substantive review powers**. In the workplace, mechanisms should be put in place to ensure meaningful human review of decisions taken or supported by AI systems. Affected individuals should receive clear explanations and be able to request reconsideration by a competent human authority. Where violations occur, procedures must allow for rectification and, where appropriate, modification or suspension of the system to prevent recurring harm.

One emerging practice is **participatory auditing** (Costanza-Chock, Raji and Buolamwini 2022), in which affected communities – workers, unions, consumers – are involved in testing and reviewing systems. This expands the notion of expertise beyond data scientists, recognising that people impacted by automated decisions have contextual knowledge about harms and unintended effects. The AI Now Institute and European labour groups have proposed such frameworks, arguing that those most impacted must have a say in evaluation.

Beyond individual audits, there is also a question of **who** is empowered to look under the hood.

Given the increasing deployment of AI systems in the workplace, employers should be required to carry out regular **fundamental rights impact assessments** for all AI systems used there, not only narrow algorithmic management tools. These assessments should look not just at isolated decisions, but at **systemic patterns**: for example, whether an attendance-scoring system systematically penalises workers with care responsibilities, or whether a risk model consistently channels inspections towards particular groups. Where such assessments reveal unlawful, discriminatory or otherwise unfair impacts, employers should be obliged to modify or discontinue the system, not simply to document its behaviour. A further step is to clarify what **meaningful oversight** looks like in practice. Persons charged with monitoring AI systems at work should receive specific training, have the authority to override automated decisions, and enjoy legal protection against dismissal or other retaliation when they exercise this role. They should be able to flag high risks of discrimination or fundamental-rights violations and, where necessary, trigger the modification or suspension of the system. Without such protections, »human oversight« risks becoming a rubber stamp rather than a safeguard.

Conclusion – towards algorithmic solidarity and critical literacy

Algorithms are not destiny. They are human artefacts, conceived, built, tested and deployed through choices. Like all tools, they can liberate or constrain, democratise or dominate. The question is not whether algorithms will govern work, but how, for whom and under whose control.

The challenge is not to reject technology, but to **reshape** it. That means demystifying the black box, insisting on enforceable rights to information and explanation, and turning audits and fundamental rights impact assessments into levers of accountability rather than box-ticking exercises. It means asserting collective rights over data and digital infrastructures, so that AI systems in the workplace are aligned with labour standards, equality law and occupational health and safety, not just with productivity metrics.

In the end, understanding how algorithms are made is not just a technical exercise. It is a political act, one that reminds us that even in the age of automation, the most important variable in any equation remains human judgement and collective organisation.

As scholar Shoshana Zuboff (2019) warned, »surveillance capitalism claims private human experience as free raw material«. But as unions have always known, what is taken can also be reclaimed. The task ahead is to reclaim the algorithm, not as a boss, but as a contested, negotiated collaborator; to ensure that the next generation of systems reflects solidarity, not subordination; to insist that the digital future is, above all, **human**.

Bibliography

Arora, P., Raman, U. and König, R. (eds) (2023): *Feminist futures of work: reimagining labour in the digital economy* (1st ed.). Routledge. Available at: https://www.taylorfrancis.com/books/oa-edit/10.5117/9789463728386/feminist-futures-work-ren%C3%A9-k%C3%B6nig-payal-arora-usha-raman

Benjamin, R. (2019): *Race after technology: abolitionist tools for the new Jim code*. Polity Press.

Bhadani, S. (2021): Biases in recommendation system, in: *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 855–859.

Buolamwini, J. (2024): *Unmasking AI: My mission to protect what is human in a world of machines*. Random House.

Capasso, M., Arora, P., Sharma, D. and Tacconi, C. (2024): On the right to work in the age of artificial intelligence: ethical safeguards in algorithmic human resource management, in: *Business and Human Rights Journal* 9(3): 346–360.

Cellard, L. (2024): Theaters of algorithmic transparency and the politics of exemplarity, in: *Science, Technology, & Human Values*.

Claburn, T. (2025): Amazon accused of using algorithms to push warehouse workers to breaking point, in: *The Register* (18 March). Available at: https://www.theregister.com/2025/03/18/amazon_algorithmic_worker_management/

Considine, M., McGann, M., Ball, S. and Nguyen, P. (2022): Can robots understand welfare? Exploring machine bureaucracies in welfare-to-work, in: *Journal of Social Policy* 51(3): 519–534.

Costanza-Chock, S., Raji, I.D. and Buolamwini, J. (2022): Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem, in: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (June), pp. 1571–1583.

European Institute for Gender Equality (2021): *Artificial intelligence, platform work and gender equality*. Publications Office of the European Union. Available at: https://data.europa.eu/doi/10.2839/372863

European Union Agency for Fundamental Rights (2024): *Being Black in the EU: Experiences of people of African descent*. Publications Office of the European Union. Available at: https://fra.europa.eu/en/publication/2023/being-black-eu

Geers, M., Swire-Thompson, B., Lorenz-Spreen, P., Herzog, S.M., Kozyreva, A. and Hertwig, R. (2024): The online misinformation engagement framework, in: *Current Opinion in Psychology* 55, 101739.

Kerckhofs, P. (2025): Collective bargaining on artificial intelligence at work. European Foundation for the Improvement of Living and Working Conditions. 26 September. Available at: https://www.eurofound.europa.eu/en/publications/all/collective-bargaining-on-artificial-intelligence-at-work

Langer, M. and König, C.J. (2023): Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management, in: *Human Resource Management Review* 33(1), 100881.

Li, R., Kingsley, S., Fan, C., Sinha, P., Wai, N., Lee, J., ... Hong, J. (2023): Participation and division of labor in user-driven algorithm audits: How do everyday users work together to surface algorithmic harms?, in: *Proceedings of the CHI Conference on Human Factors in Computing Systems* (April), pp. 1–19.

Metaxa, D., Park, J.S., Robertson, R.E., Karahalios, K., Wilson, C., Hancock, J., and Sandvig, C. (2021): Auditing algorithms: understanding algorithmic systems from the outside in, in: *Foundations and Trends in Human–Computer Interaction* 14(4): 272–344.

Pathak, S. (2025): From data to discrimination: How AI reflects the gender biases we live with, in: *News18* (3 August). Available at: https://www.news18.com/explainers/from-data-to-discrimination-how-ai-reflects-the-gender-biases-we-live-with-9481004.html

Talby, D. (2020): Three insights from Google's »failed« field test to use AI for medical diagnosis, in: *Forbes* (9 June). Available at: https://shorturl.at/vrHe8

Tsymbal, A. (2004): *The problem of concept drift: Definitions and related work*. Technical report. Trinity College Dublin, Computer Science Department.

Vinod, B.K., Capasso, M., Arora, P., Castillo, C. and Saldivar, J. (2024): Proxy discrimination risks in hiring: A qualitative analysis of a set of real CVs, in: *SSRN* (3 December). Available at: http://dx.doi.org/10.2139/ssrn.5048771

Zhu, G., Huang, J., Lu, J., Luo, Y. and Zhu, T. (2024): Gig to the left, algorithms to the right: A case study of the dark sides in the gig economy, in: *Technological Forecasting and Social Change*, 199, 123018.

Ziosi, M. and Pruss, D. (2024): Evidence of what, for whom? The socially contested role of algorithmic bias in a predictive policing tool, in: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1596–1608.

Zuboff, S. (2019): *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* Public Affairs.

About the authors

The Inclusive AI Lab at Utrecht University is dedicated to incubating leaders and helping to build inclusive, responsible and ethical AI data, tools, services, policies and platforms, with a special focus on the Global South.

Dr. Payal Arora is Professor of Inclusive AI Cultures at Utrecht University and Founder of the Inclusive AI Lab. She is the author of award-winning books, including *The Next Billion Users* (Harvard Press) and *From Pessimism to Promise* (MIT Press). She is listed in the 100 Brilliant Women in AI Ethics 2025, nominated as one of the Eight Women with a New Vision for the Earth in the 2026 GLF Women Campaign. and won the 2025 Women in AI Benelux Award on Diversifying AI. *Forbes* called her the champion of the next billion champion and the »right kind of person to reform tech«.

Dr Roberta Calegari is a Professor of AI and AI Ethics at the Department of Computer Science and the Alma Mater Research Institute for Human-centred Artificial Intelligence, University of Bologna. Her work focuses on fairness, non-discrimination and the societal impact of AI. She coordinates the Horizon Europe project AEQUITAS (GA 101070363), delivering a framework within which to analyse and mitigate bias before deployment. She serves on the Editorial Board of ACM Computing Surveys, has authored 100+ peer-reviewed publications, and leads multiple trans-regional projects in collaboration with industry partners.

Inside the Black Box: How algorithms are made – and why it matters

From warehouses to recruitment platforms, algorithms are quietly becoming the new managers of modern work. Increasingly, decisions about hiring, scheduling, productivity and even dismissal are shaped by automated systems that analyse vast amounts of data. These technologies promise efficiency and innovation, but they also raise important questions about fairness, transparency and worker rights. Algorithms are often presented as neutral tools, but they reflect the choices, assumptions and data used to build them. Past data is likely to reflect inequalities – such as gender or racial bias in hiring – and algorithms may reproduce and even amplify such patterns. At the same time, many systems operate as »black boxes«, which makes it difficult for workers, unions or regulators to understand how decisions, which are likely to affect people's livelihoods, are really made.

This paper takes readers inside the process of how algorithms are designed, trained and deployed in workplaces. It highlights the key stages at which bias, exclusion or unfair treatment may enter the system, from definition of the problem to selection of data and testing of models. It also examines the growing importance of algorithmic audits and stronger governance to ensure accountability. The message is clear: algorithms are not destiny. With stronger oversight, worker participation and ethical design, these technologies can support more transparent, equitable and humane workplaces rather than reinforce existing inequalities.

Further information on the topic can be found here:
↗ **fes.de**

Friedrich Ebert Stiftung