Artificial intelligence, labour and society

Edited by
Aída Ponce Del Castillo

etui.

Artificial intelligence, labour and society

Edited by
Aída Ponce Del Castillo

European Trade Union Institute (ETUI)

Cite this book: Ponce del Castillo A. (ed.) (2024) Artificial intelligence, labour and society, ETUI.

Print: ETUI Printshop, Brussels

D/2024/10.574/10
ISBN: 978-2-87452-707-4 (print version)
ISBN: 978-2-87452-708-1 (electronic version)

The ETUI is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the ETUI. Neither the European Union nor the ETUI can be held responsible for them.

Contents

Foreword

This book represents a stage in the ongoing journey of my research into emerging technologies. It is the result of a project on artificial intelligence that began in 2017. At that time, the ETUI started organising training courses for trade unionists. In parallel, I launched 'AI Talks', a series of monthly online conversations with leading academic experts from around the world, invited to share their research, or their latest work on AI, with a wider audience.

Both the training courses and the AI Talks have given me the opportunity to engage and collaborate with knowledgeable and talented scholars and researchers. Their expert insights have been collected and then shaped into the chapters in this book, which are based on meticulous research.

The book focuses on the profound implications of AI for both the labour market and society at large, taking a multidisciplinary perspective and incorporating a diversity of geographical and cultural points of view. As AI gradually transforms the very fabric of our societies, understanding its dynamics is essential for both the general public and the labour movement. This book aims to bring knowledge to both audiences, although it is probably the latter that will benefit most from the insights. Whatever discipline we come from, I believe that it is by being well informed that we will best be able to navigate the AI future and increase our agency.

Human beings are not simple entities bound to a particular and well-defined context. Throughout our lifecycle, we interact with other people, cultures, technologies and the environment. We move seamlessly from one sphere to another, constantly oscillating between overlapping roles: citizen, worker, consumer, etc. The interdisciplinary nature of this book aims to enrich our understanding of AI by highlighting the multiple implications it has in the different contexts in which we find ourselves. The underlying aim is to encourage readers to move beyond singular perspectives and adopt a broader approach to AI.

The structure of this book is defined by themes that resonate with vast human dimensions – areas where AI raises broad, sometimes existential questions that also engage with the lifecycle of AI. To understand the present and future of AI, it is essential to recognise and situate it properly, to reflect on how it interacts with the planet and to consider the resources – both human and natural – that are needed to produce and deploy it. Unsurprisingly, the world of work is a central theme, addressed by many of the contributing authors within these pages. The governance choices made today will

affect the future of millions of citizens and workers around the world, and these will determine our ability to control AI.

AI is a powerful technology that raises major questions about what it means for people to live and to thrive. It forces all stakeholders – international organisations, national governments, civil society, trade unions, academics, activists, businesses, etc. – to review their strategies and their courses of action. This is a long and uncertain journey. The ambition of this book is to make that journey a little more informed and a little more inspired.

Aída Ponce Del Castillo
January 2024

Acknowledgement

The editor would like to express her gratitude to the authors of this book for their contributions. She equally extends her thanks to those whose work in the editing and production process might be less visible, yet has been essential in supporting her throughout this process.

Part 1

AI in society

Chapter 1
AI: the value of precaution and the need for human control

Aída Ponce Del Castillo

'*Whose duty is it in today's complex societies to foresee or forestall the negative impacts of technology, and do we possess the necessary tools and instruments for forecasting and preventing harm?*'

Sheila Jasanoff, Pforzheimer Professor of Science and Technology Studies, Harvard Kennedy School

'*The aim is quite simple. Let's use it more. Let's have artificial intelligence everywhere where it makes a difference.*'

European Commission Executive Vice-President Margrethe Vestager, at the 2021 Data Science & Law Forum

1. Introduction

When new technologies emerge, two opposing governance approaches may arise. One favours stringent regulation to safeguard society from unanticipated hazards while the other prioritises the promotion of technology deployment by eschewing what is often seen as expensive and innovation-stifling regulation (Kaal 2016; Cortez 2019; Mandel 2020). Between these two extremes, various governance approaches can be developed. These will include governance through international human rights, or through hard law (which includes risk-based regulation) or soft law.

Whatever approach is chosen, the timing of governance interventions is crucial. If interventions and course corrections are made early, they are likely to be less expensive and easier to carry out. However, the full implications of emerging technology and the need for change might not yet be fully understood. Delaying intervention until it becomes necessary can result in more challenging, time-consuming and costly course corrections (Collingridge 1982).

Artificial intelligence (AI), similarly to other high-risk emerging technologies such as biotechnology, blockchain, synthetic biology, metaverse environments and nanotechnology, presents certain key attributes: radical novelty; relatively fast growth; coherence over time; a prominent impact on society and the economy; and uncertainty and ambiguity (Rotolo et al. 2015). The European Commission (EC), in its proposed regulation on machinery products, mentions additional attributes including data dependency, opacity, autonomy and connectivity, recognising that these can increase both the probability of harm and its impact as well as negatively affect the safety of any machinery that integrates AI (European Commission 2021).

Compared to other high-risk emerging technologies, AI also raises several unique concerns related to possible bias and discrimination, the preservation of democratic values, the need to render automated decision-making more explicit (Taeihagh et al. 2021), widening inequalities, the impact of bad data, the protection of data privacy and the prevention of mass surveillance (Zuboff 2019; Zhang et al. 2021).

Against this high-risk background of emerging, rapidly evolving, uncertain and high-impact technology, the option of shifting the point of initial governance to an earlier stage of technological development is both necessary and valid. The risk control approach presently employed for established and clearly defined technologies, employing risk assessment followed by risk management to maintain exposure levels below the acceptable, is inadequate for emerging high-risk technologies: the quantitative data required is limited or unavailable; and the potential consequences of using the technology cannot be comprehensively listed (Linkov et al. 2018). Establishing a proactive, collaborative and flexible form of precautionary governance that evolves in conjunction with the technology may improve our prospects of safeguarding society from AI's potential impacts (Mandel 2013, 2020).

Such an approach, which establishes an initial point of governance at an early stage, ought to incorporate two essential principles: firstly, the legal principle of precaution, which has demonstrated its effectiveness in managing the risks associated with emerging and unpredictable technologies; and secondly, the principle of human-in-control.

This is crucial if we acknowledge that AI is an extension of conventional automation and operates under the same general principle – 'If it is technically and economically feasible to automate a function, automate it' (Billings 1996: 9) – which triggers an obvious question: can automation exist without human control?

2. AI case studies: why human intervention is necessary

The risks posed by AI are not theoretical and academic researchers have identified many. Sometimes, however, reality speaks louder than words. The four case studies described below illustrate how AI and non-AI algorithms have delivered wrong outputs or 'recommendations' and had a significant impact on the lives of thousands of vulnerable individuals: people have lost their right to welfare benefits, families have been torn apart, students have been assigned wrong grades, etc.

Case 1 UK A-level grades algorithm: the teachers feeding the system

In 2020, the UK Education Ministry made a decision to cancel exams due to Covid-19 and sought alternative solutions to assign grades to students. The two solutions identified were: (a) rely entirely on teacher assessed grades; or (b) standardise the results (Kippin and Cairney 2022).

Tool: the Office of Qualifications and Examinations Regulation (Ofqual) chose option 2 and designed and implemented an algorithm to grade students (Office for Artificial Intelligence 2020). Teachers were asked to supply for each student and for every subject an estimated grade and a ranking, compared with every other student at the school within the same estimated grade. A-Level students were awarded grades generated by the algorithm which 'looked at the historical grade distribution of a school and then decided a students' grade on the basis of their ranking' (Kolkman 2020).

Consequences: tens of thousands of school pupils received grades lower than they had anticipated (Kolkman 2020). Ofqual's grading algorithm also affected many schoolteachers. Government officials argued that 'basing grades on teacher estimates alone would damage the credibility of this year's results compared to previous years' and the whole issue turned into a public scandal (Coughlan 2020).

Role of the human: for policy reasons, the algorithm was preferred to the teacher. As the report from Ofqual seems to suggest, teachers were interviewed about the grading process and requested to feed the algorithm. However, it is unclear if the teachers who were interviewed were informed about the intention to use an algorithm for the A-level grading process (Office for Artificial Intelligence 2020).

Case 2 The Dutch childcare benefits scandal (*toeslagenaffaire*): where were the supervisors?

In 2003, the Dutch authorities developed an automated welfare fraud detection system called Systeem Risico Indicatie (SyRI) (van Bekkum and Borgesius 2021). Used by the Dutch Tax and Customs Administration, the system used algorithms in which 'foreign sounding names' and 'dual nationality' were used as indicators of potential fraud (European Parliament 2022). The existence of this system was reported by the media in 2020 and an investigation by the Dutch Data Protection Authority found that the algorithms were discriminatory, among other things because they took into account variables such as someone having a second nationality.

Tool: a risk detection algorithm to process the social security documents of individuals applying for childcare benefits, alongside another machine learning tool, namely a risk-scoring algorithm to automate the selection of childcare allowance recipients. The system derived risk factors based on the analysis of historical data in order automatically to process the documentation and select welfare recipients for audits. Then, officials scrutinised those claims with the highest risk label (Hadwick and Lan 2021).

Consequences: the output of the algorithms became biased and unfair as the algorithm concluded that non-Dutch welfare recipients were more prone to fraud. Some 26,000 parents were mistakenly accused by the Dutch tax authorities of fraudulently claiming child allowance over several years from 2012, while 10,000 families were forced to repay tens of thousands of euros, in some cases leading to unemployment, bankruptcies and divorces (Henley 2021).

Role of the human: the decision to cut a family off from benefits payments should have gone through an extensive review process. The choices were left to algorithms. The key question here is the role of supervisors as 'supervisors are still responsible for their work, even if part of it is performed by a computer' (Ten Seldam and Brenninkmeijer 2021).

Case 3 The French *Foncier Innovant* system to identify swimming pool fraud: surveyors not consulted

In 2021, the State Public Finance Department developed an AI tool to detect undeclared outbuildings and swimming pools as part of a plan to update and measure all buildings on the national territory. The State's objective was to update available maps, improve their quality and facilitate the collection of taxes to be paid by homeowners (Direction générale des Finances publiques 2022).

Tool: 'Foncier innovant' was developed in partnership with consulting firm Capgemini and Google. The tool used data captured by the government platform Le Géoportail and other available geographical information to locate pools and other outbuildings.

Role of the human: during its development, the project did not involve land surveyors, and their expertise and technical skills were not taken into account. Surveyors are now concerned about the gradual automation of their expertise, especially about the calculation of cadastral plans and the setting of the boundaries of private properties. Their lack of involvement means that the quality of the plans is not guaranteed, there may be an adverse impact on citizens and that the reliability of the whole process is limited. The public service that implements the AI tool has noticed that quality of service is decreasing.

Case 4 Serbian law on social cards

In 2022, the Serbian government introduced a social card system to promote administrative efficiency in the welfare system. This affected the Centres for Social Work across the country. The system centralised a government database of individuals who are recipients of or who apply for social security benefits, consolidating personal data from multiple data registries in a database that can be accessed by a significant number of employees in the social protection sector.

Tool: the system profiled individuals and used an automated decision-making tool able immediately and legally to suspend or reduce social benefits and social assistance, without considering people's life circumstances or allowing individuals to provide contextual or additional information (Amnesty International 2023; Government of the Republic of Serbia 2021).

Consequences: the system collected sensitive data, aggregating it from other databases (including information about ex partners) and violated privacy laws. In addition, 34,686 individuals lost their social benefits because their recorded earnings put them above the minimum threshold for assistance.

The NGO 'A11 Initiative for Economic and Social Rights' challenged the law establishing the social cards system before Serbia's Constitutional Court. The judgment is in preparation at the time of writing (A11 Initiative for Economic and Social Rights 2023).

Role of the human: social workers were unable to amend the data in the system or to override decision-making. The system uses a traffic light system with three options for social workers to click on: urgent, check, inform; and a 3-day deadline to perform a verification check in urgent notifications.

3. The precautionary principle – a legitimate way to address the risks of AI

In the EU context, the governance of AI will principally rely on the legislative tool that is the AI Act. Should the AI Act fail to deliver the necessary protection – by design or because the authorities involved either lack the resources to enforce it, act without the necessary level of coordination or because uncertainty is too high – there will be a need for solid legal principles to come to the rescue. The precautionary principle, which emanates from international environmental law, is such a principle. It represents an early warning system that 'enables decision-makers to adopt precautionary measures when scientific evidence about an environmental or human health hazard is uncertain and the stakes are high' (European Parliament 2015).

Developed in the early 1980s and formally adopted in 1992 at the UN Rio de Janeiro Conference on Environment and Development and in the UN Convention on Biodiversity, it was defined in 2005 by the UNESCO World Commission on the Ethics of Scientific Knowledge and Technology in the following manner:

> When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm. Morally unacceptable harm refers to harm to humans or the environment that is threatening to human life or health, or serious and effectively irreversible, or inequitable to present or future generations, or imposed without adequate consideration of the human rights of those affected. (UNESCO 2005: 13)

In the EU, the principle was included in the Maastricht Treaty in 1992 and is there in Article 191 of the Treaty on the Functioning of the European Union. In practice, the precautionary principle has underpinned the EU's environmental policy and has been a core element of its risk and public health policies. In the area of chemicals policy, Article 1.3 of the Regulation on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) states that 'its provisions are underpinned by the precautionary principle'. In the field of EU food safety, Article 7 of the General Food Law Regulation states that 'when faced with these specific circumstances, decision makers or risk managers may take measures or other actions based on the precautionary principle'. The principle is also invoked in the fields of food safety and trade.

Further, as Mazur (2019) argues, some of its elements can be seen in the General Data Protection Regulation (GDPR): the right to be informed about the collection and use of personal data (Articles 13 and 14); and the right to assess the impact of such processing (Article 35).

The European Court of Justice has interpreted the precautionary principle in various cases, including the 1983 *Sandoz* case,[1] *Alpharma Inc*[2] and *Solvay Pharmaceuticals BV*.[3] On the latter, the Court expanded the scope of the principle from the protection of the environment to the protection of public health, concluding that 'in the domain of [human health], the existence of solid evidence which, while not resolving scientific uncertainty may reasonably raise doubts as to the safety of a substance justifies, in principle, [the refusal to include that substance…]. The precautionary principle is designed to prevent potential risks.'

While it is a well-established principle that can be legitimately invoked and applied, careful reflection should be given to how and when it should be applied to AI and how it interrelates with decision-making (Fisher et al. 2006; Stirling 2006; Donati 2021). Some criticise the principle for being paralysing and unscientific, and for promoting a culture of irrational fear; a reaction, as Aven (2023) argues, often rooted in a lack of clarity over the meaning and the scope of the principle. The EC, to address the controversies, has established guidelines for its application whereby it can be invoked only when three preliminary conditions are met: (a) identification of potentially adverse effects; (b) evaluation of the scientific data available; (c) the extent of scientific uncertainty (European Commission 2000).

If, as Hansson (2023) argues, precautionary actions are based on the current state of science, that potential dangers of limited plausibility are excluded and that the precautionary principle is not used to make judgments between competing top priorities, then it can be recognised as an essential principle that must be at the heart of technological development. With detailed procedures, standards and guidelines developed to enable its application in specific contexts (Aven 2023), the precautionary principle should formally be included in the governance of AI and other emerging technologies. It can sustain their development, give direction to innovation, help build a governance based on dialogue that involves relevant societal actors and contribute to ensuring that technological innovations are safe for society.

4. Human-in-control as a response to the risks of automation and AI

The human-in-control concept first appeared in the aviation sector with the introduction of automated aircraft control technology. Recognising that 'automation is able to limit

1. Case 174/82, *Sandoz BV* ECLI:EU:C:1983:213, para. 16. The CJEU recognised the idea underlying the precautionary principle by stating that 'in so far as there are uncertainties at the present state of scientific research it is for the Member States, in the absence of harmonization, to decide what degree of protection of the health and life of humans they intend to assure'. See also Guida (2021).
2. Case T-13/99 *Pfizer Animal Health SA v Council of the European Union* ECLI:EU:T:2002:209, para. 444: 'The institutions cannot be criticised for having chosen to withdraw provisionally the authorisation of virginiamycin as an additive in feedingstuffs, in order to prevent the risk from becoming a reality, and, at the same time, to continue with the research that was already under way. Such an approach, moreover, was consonant with the precautionary principle, by reason of which a public authority can be required to act even before any adverse effects have become apparent.'
3. Case T-392/02, *Solvay Pharmaceuticals BV v Council of the European Union* ECLI:EU:T:2003:277, para. 3.

the operator's authority' and that 'sometimes, it is not obvious for the operator to know that this has occurred', NASA designed principles for what it calls human-centred automation (Billings 1996).

The main axiom is that humans, in this instance the pilot and the air traffic controllers, bear the responsibility and remain in command: the first of their flights; the latter of air traffic more generally. The corollaries are that pilots and controllers: (a) must be actively involved; (b) must be adequately informed; (c) must be able to monitor the automation assisting them; (d) the automated systems must be predictable; (e) the automated systems must monitor the human operators; and (f) every intelligent system element must understand the intent of other intelligent system elements (Billings 1996).

An important additional remark is made: 'Though humans are far from perfect sensors, decision-makers and controllers, they possess three invaluable attributes. They are excellent detectors of signals in the midst of noise, they can reason effectively in the face of uncertainty, and they are capable of abstraction and conceptual organization' (Billings 1996).

Here, if we look at the world of work, a relevant parallel can be established with the role of workers' representatives who are present in the workplace and are reliable detectors of variables, weak signals, hazards and other factors that influence the work organisation and environment.

Human-in-control has also been adopted by the military, including in the use of AI-based military applications for planning, decision support and intelligence, in particular the Intelligence, Surveillance, Target Acquisition and Reconnaissance (ISTAR) capabilities developed for the armed forces.

The principle has been promoted both by international organisations and by national jurisdictions. The International Committee of the Red Cross has stressed the need for 'human control' of certain 'critical functions' of weapons systems, in particular their ability to 'select (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy) targets' without human intervention (Davison 2018).

In recent years, the principle has fully migrated to the AI sector and has been attributed with a range of different purposes by scholars (Davidovic 2023). Some qualify it as 'a key tool for assuring safety, dignity, and responsibility for AI and automated decision-systems' (Christen et al. 2023; Davidovic 2023). Other purposes relate to the need to ensure accuracy, safety and precision; to deliver accountability and responsibility; and to have sufficient understanding of the process to consent, dissent or to ensure trust in the institutions (Davidovic 2023).

As far as AI governance in the EU is concerned, no mention of human-in-control as such is made in the early EC communication on AI. Initially, the Commission referred to a 'human-centred' approach to AI and to the need to foster 'human-centric' digitalisation. The vision aims at ensuring people are at the centre and empowered, and that innovative business is fostered. It is with this vision in mind that, in December 2022, the European

Parliament, the Council and the Commission proclaimed a joint Declaration on Digital Rights and Principles for the Digital Decade, which they qualify as reflecting EU values and promoting a sustainable, human-centric vision for the digital transformation. Importantly, the Declaration should guide policymakers and put people at the centre of the digital transformation.

However, a shift of perspective has since taken place, from human-centric to human-in-control. Although not explicitly stated in the EC's digital agenda, human-in-control has gradually infused the speeches of high-level commissioners. When presenting the EC's digital strategy, Shaping Europe's Digital Future, President von der Leyen stated that 'Artificial intelligence must always comply with people's rights. This is why a person must always be in control of critical decisions' (President von der Leyen 2020). Later that same year, she also stated that 'we must ensure that our rights, privacy and protections are the same online as they are off it. That we can each have control over our own lives and over what happens to our personal information. That we can trust technology with what we say and do. That new tech does not come with new values' (von der Leyen 2020). In May 2023, in her speech at the 15th Congress of the ETUC, she stated that '[…] the answer to the challenges that AI raises is first of all a principle. This principle is called "human-in-control". That must be our underlying principle for everything' (von der Leyen 2023a). Then, in September 2023, in her speech at the Pulse Women Economic Network, she reaffirmed that 'the EU is promoting the "human-in-control" principle for sensitive applications of AI. Because the new digital world should not reproduce old inequalities, but open up new opportunities' (von der Leyen 2023b).

In 2017, in its opinion on 'Artificial Intelligence – the consequences of Artificial Intelligence on the (digital) single market, production, consumption, employment and society', the European Economic and Social Committee (EESC) had already called for humans to be in control of the technology, referring to a 'human-in-command' approach to AI 'including the precondition that the development of AI be responsible, safe and useful, where machines remain machines and people retain control over these machines at all times'. It called for 'transparent, comprehensible and monitorable AI systems, the operation of which is accountable, including retrospectively'. It also referred to the role of managers who should be involved 'so that they remain in control of these developments and are not the victims of them'. (European Economic and Social Committee 2017).

To operationalise human-in-control, such high-level recognition of its relevance and legitimacy in respect of AI governance is an essential prerequisite. As a concept, human-in-control goes beyond legislation and provides a supplementary level of protection, and it may well be needed if the obligations established under the AI Act and the standards attached to them become obsolete as the technology converges and becomes more intertwined in our lives.

However, it can only exist if humans are involved, not simply informed. In the world of work, while the EU legislation provides for information and consultation rights when

technology is introduced or modified in workplaces,4 the level of such information and consultation often have different interpretations. Here, one may cite the NASA principles as a source of useful inspiration as they establish that workers must play 'an active and necessary role apart from simply monitoring the course of the operation. That role may involve active control, or decision-making, or allocation of resources, or evaluation of alternatives, but it should not be passive, as it too often is today' (Billings 1996: 10).

In their contributions to this book, some of the authors address the various dimensions of human-in-control, including the need to consider the conditions of data production as well as the tools and equipment used to manufacture and market these systems, as Antonio A. Casilli argues. Benedetta Brevini reflects on the AI life cycle, the infrastructure and the often scarce resources it uses: should there be a discussion about who controls the essential infrastructures that power AI?

Real control also raises the question of adequate enforcement. Mario Guglielmetti insists on the correct allocation of competences between the sectoral competent authorities and the supervisory authorities. In their enforcement duties, these need to cooperate effectively and exchange relevant information, including at cross-border level.

As AI systems and other convergent technologies have the ability to influence humanity as a whole, in all its dimensions (physical and mental wellbeing, dignity and other fundamental rights), ensuring that humans are in control implies a multi-dimensional approach. Here, as Helga Nowotny has described, a movement devoted to digital humanism has appeared and is attempting to integrate a human-centred approach in the design, production and deployment of AI throughout their systemic interlinkages (Nowotny 2021). This seeks to identify specific intervention points and to be attentive to actual practices in various domains, as well as become part of the education system. The values on which digital humanism is based will be crucial for shaping the future of work and of liberal democratic societies.

The deployment of AI in the workplace could therefore support the improvement of working conditions and the prevention of occupational risks, but it can also exacerbate the deterioration of working conditions if the tools provided reinforce strategic and organisational orientations that endanger workers' physical and mental health. In this context, increasing AI literacy is indispensable. As the degree to which humans can exercise meaningful control is essential (Cavalcante Siebert et al. 2023), workers can increase their level of control if they become critical agents, able to understand the role of AI at work and its impact on their occupation, and anticipate how it may transform their careers, skills and roles. Passively using AI systems does not benefit them: a certain distance is needed for workers to see AI's overall influence (Ponce del Castillo 2020, 2023). They would need to be able to distinguish situations in which AI systems are effective or not (Brynjolfsson et al. 2023) and be sure whether they can use the technology reliably. This implies a shared understanding of technological advance and its consequences for work, as María Luz Rodríguez Fernández argues in these pages.

4. See Annex I Point 1(a) of the EU European Works Council Directive 2009/38.

Explicit discussions about the responsibilities and liabilities of each actor, including mechanisms for overruling the AI system 'through intervening and correcting behavior, setting new goals, or delegating sub-tasks' (Cavalcante Siebert et al. 2023) must also be had. This further entails an understanding of the costs behind the production of AI, in particular the natural resources used in its production, deployment and maintenance. Workers and trade unions need to develop this new skill which can help them navigate volatile and fast-moving technological developments.

As German Bender explains in his chapter, trade unions and employers need to be able to bargain – or codetermine – AI systems, regulating their use conditions and their possible known and unknown effects. This would extend the possibility of worker participation to areas usually beyond their reach because they are reserved to corporate actors, including 'black box' algorithms. The right to meaningful participation should, in the view of Rodríguez Fernández, allow worker representatives to 'see inside' and utilise what they see to guarantee that the decisions do not cause bias or differentiated treatment without justification.

Looking beyond the world of work, respect for intellectual and emotional autonomy is another essential dimension, as Frank Pasquale argues. Genuine control also requires an inclusive and participatory governance, involving a wide range of stakeholders including from marginalised and disadvantaged groups (see Ulnicane, this volume).

Finally, Helga Nowotny contends that exercising genuine control entails a reassessment of profit allocation. The inequitable distribution of the productivity gains stemming from AI systems raises the question of whether those responsible for generating profit for the technology sector should also benefit from it. This encompasses the workers involved in the development and application of AI, as well as those whose data is utilised to improve or generate the technology (Tubaro et al. 2020). It may therefore be necessary to reconsider how workers can participate not just in AI extraction and production, but also in the ensuing economic rewards for the data they provide (Brynjolfsson et al. 2023), possibly in the form of wage increases. This volume does not address this matter directly, but it does approach it indirectly.

5. Conclusion

As illustrated by the four case studies presented here, the risks posed by AI are serious and are having an impact on the lives of thousands of people globally. To address those risks better and to prevent harm, the point of initial governance must be moved to an earlier stage of technological development. This implies giving the legal precaution principle and the concept of human-in-control a central role in the governance approach. Such a proactive, anticipative and collaborative form of precautionary governance can improve our prospects of safeguarding society from AI's potential impacts. With the increasing autonomy of AI systems and the recent development of generative AI, the risk of creating AI systems that pursue undesirable goals becomes real and the need for control and effective human intervention even greater (Bengio et al. 2023).

The need for human control is one of the perspectives discussed in this book, which presents the insights of prominent scholars and specialists in the field hailing from Europe and other parts of the globe. When the project was initiated, the intention was to assess the challenges that AI poses, pinpoint the crucial aspects of a potential response and highlight some possible fundamental constituents of an all-encompassing framework for AI governance. We modestly hope the book achieves some of these objectives.

References

A11 Initiative for Economic and Social Rights (2023) Support grows for A 11 constitutional challenge to the social cards law. https://www.a11initiative.org/en/support-grows-for-a-11-constitutional-challenge-to-the-social-cards-law/

Amnesty International (2023) Serbia submission for European Union enlargement package/opinion, 2023. https://www.amnesty.org/en/wp-content/uploads/2023/04/EUR7066882023ENGLISH.pdf

Aven T. (2023) A risk and safety science perspective on the precautionary principle, Safety Science, 165, 106211. https://doi.org/10.1016/j.ssci.2023.106211

Billings C.E. (1996) Human-centered aviation automation: Principles and guidelines, NASA Technical Memorandum 110381, National Aeronautics and Space Administration.

Bengio Y. et al. (2023) Managing AI risks in an era of rapid progress. https://managing-ai-risks.com/

Brynjolfsson E., Li D. and Raymond L.R. (2023) Generative AI at work, Working Paper 31161, National Bureau of Economic Research. https://doi.org/10.3386/w31161

Cavalcante Siebert L. et al. (2023) Meaningful human control: Actionable properties for AI system development, AI and Ethics, 3 (1), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Christen M., Burri T., Kandul S. and Vörös P. (2023) Who is controlling whom? Reframing 'meaningful human control' of AI systems in security, Ethics and Information Technology, 25 (1), 10. https://doi.org/10.1007/s10676-023-09686-x

Collingridge D. (1982) The social control of technology, Palgrave Macmillan.

Cortez N. (2019) Digital health and regulatory experimentation at the FDA, Yale Journal of Law and Technology, 21 (4), 4–26.

Coughlan S. (2020) Scottish school pupils have results upgraded, BBC News, 8 November 2020. https://www.bbc.co.uk/news/uk-scotland-53740588.amp

Davidovic J. (2023) On the purpose of meaningful human control of AI, Frontiers in big data, 5. https://doi.org/10.3389/fdata.2022.1017677

Davison N. (2018) A legal perspective: Autonomous weapon systems under international humanitarian law. https://www.icrc.org/en/download/file/65762/autonomous_weapon_systems_under_international_humanitarian_law.pdf

Direction générale des Finances publiques (2022) L'intelligence artificielle au service de la lutte contre la fraude : bilan de l'expérimentation « foncier innovant ». https://www.impots.gouv.fr/sites/default/files/media/2_actu/home/2022/dp_foncier_innovant.pdf

Donati A. (2021) The precautionary principle under European Union law, Hitotsubashi Journal of Law and Politics, 49, 43–60. https://doi.org/10.15057/hjlp.2020003

European Commission (2000) Communication from the Commission on the precautionary principle, COM(2000) 1 final, 2.2.2000. https://op.europa.eu/en/publication-detail/-/publication/21676661-a79f-4153-b984-aeb28f07c80a/language-en

European Commission (2021) Proposal for a regulation of the European Parliament and of the Council on machinery products, COM(2021) 202 final, 21.4.2021. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2021:202:FIN

European Economic and Social Committee (2017) Opinion of the European Economic and Social Committee on: Artificial intelligence – The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society, Official Journal of the European Union, C 288, 31.8.2017. https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence-consequences-artificial-intelligence-digital-single-market-production-consumption-employment-and

European Parliament (2015) The precautionary principle: Definitions, applications and governance, European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA(2015)573876_EN.pdf

European Parliament (2022) The Dutch childcare benefit scandal, institutional racism and algorithms. https://www.europarl.europa.eu/doceo/document/O-9-2022-000028_EN.html

Fisher E.C., Jones J.S. and von Schomberg J.R. (2006) Implementing the precautionary principle: Perspectives and prospects, in Fisher E.C., Jones J.S. and von Schomberg R. (eds.) Implementing the precautionary principle: Perspectives and prospects, Edward Elgar, 1–18.

Government of the Republic of Serbia (2021) Government passes social card bill. https://www.srbija.gov.rs/vest/en/166629/government-passes-social-card-bill.php

Guida A. (2021) The precautionary principle and genetically modified organisms: A bone of contention between European institutions and Member States, Journal of Law and the Biosciences, 8 (1). https://doi.org/10.1093/jlb/lsab012

Hadwick D. and Lan S. (2021) Lessons to be learned from the Dutch childcare allowance scandal: A comparative review of algorithmic governance by tax administrations in the Netherlands, France and Germany, World Tax Journal, 13 (4), 609–645.

Hansson S.O. (2020) How extreme is the precautionary principle?, Nanoethics, 14 (3), 245–257. https://doi.org/10.1007/s11569-020-00373-5

Henley J. (2021) Dutch government faces collapse over child benefits scandal, The Guardian, 14 January 2021. https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal

Kaal W.A. (2016) Dynamic regulation for innovation, in Fenwick M., Kaal W.A., Kono T. and Vermeulen E.P.M. (eds.) Perspectives in law, business and innovation, Springer, 16–22.

Kippin S. and Cairney P. (2022) The Covid-19 exams fiasco across the UK: Four nations and two windows of opportunity, British Politics, 17 (1), 1–23. https://doi.org/10.1057/s41293-021-00162-y

Kolkman D. (2020) F** k the algorithm?: What the world can learn from the UK's A-level grading fiasco, Impact of Social Sciences Blog, 26 August 2020. https://eprints.lse.ac.uk/106366/1/impactofsocialsciences_2020_08_26_fk_the_algorithm_what_the_world_can.pdf

Linkov I. et al. (2018) Comparative, collaborative, and integrative risk governance for emerging technologies, Environment Systems and Decisions, 38 (2), 170–176. https://doi.org/10.1007/s10669-018-9686-5

Mandel G.N. (2013) Emerging technology governance, in Marchant G.E., Abbott K.W. and Brown J.E. (eds.) Innovative governance models for emerging technologies, Edward Elgar, 44–62.

Mandel G.N. (2020) Regulating emerging technologies, in Marchant G.E. and Wallach W. (eds.) Emerging technologies: Ethics, law and governance, Routledge.

Mazur J. (2019) Automated decision-making and the precautionary principle in EU law, TalTech Journal of European Studies, 9 (4), 3–18. https://doi.org/10.1515/bjes-2019-0035

Mieg H.A. (ed.) (2022) The responsibility of science, Springer. https://doi.org/10.1007/978-3-030-91597-1

Nowotny H. (2021) In AI we trust: Power, illusion and control of predictive algorithms, John Wiley & Sons.

Office for Artificial Intelligence (2020) A guide to using artificial intelligence in the public sector. https://www.gov.uk/government/publications/a-guide-to-using-artificial-intelligence-in-the-public-sector

Ponce Del Castillo A. (2020) Labour in the age of AI: Why regulation is needed to protect workers, Foresight Brief 08, ETUI. https://www.etui.org/publications/foresight-briefs/labour-in-the-age-of-ai-why-regulation-is-needed-to-protect-workers

Ponce Del Castillo A. (2023) AI: Discovering the many faces of a faceless technology: A hands-on tool to help map AI, strengthen critical thinking and support anyone involved in negotiating the deployment of AI systems, ETUI. https://www.etui.org/publications/ai-discovering-many-faces-faceless-technology-0

Pouget H. and Laux J. (2023) A letter to the EU's future AI office, Carnegie Endowment for International Peace, 3 October 2023. https://carnegieendowment.org/2023/10/03/letter-to-eu-s-future-ai-office-pub-90683

Robbins S. (2023) The many meanings of meaningful human control, AI and Ethics, 1–12. https://doi.org/10.1007/s43681-023-00320-6

Rotolo D., Hicks D. and Martin B.R. (2015) What is an emerging technology?, Research policy, 44 (10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006

Schwarz E. (2018) The (im)possibility of meaningful human control for lethal autonomous weapon systems, Humanitarian law and policy, 29 August 2018. https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/

Stirling A. (2006) Precaution, foresight and sustainability: Reflection and reflexivity in the governance of science and technology, in Voss J. and Kemp R. (eds.) Reflexive governance for sustainable development, Edward Elgar, 225–272.

Taeihagh A., Ramesh M. and Howlett M. (2021) Assessing the regulatory challenges of emerging disruptive technologies, Regulation and Governance, 15 (4), 1009-1019. https://doi.org/10.1111/rego.12392

Ten Seldam B. and Brenninkmeijer A. (2021) The Dutch benefits scandal: A cautionary tale for algorithmic enforcement, EU Law Enforcement, 30 April 2021. https://eulawenforcement.com/?p=7941

Tegtmeier P., Weber C., Sommer S., Tisch A. and Wischniewski S. (2022) Criteria and guidelines for human-centered work design in a digitally transformed world of work: Findings from a formal consensus process, International Journal of Environmental Research and Public Health, 19 (23), 15506. https://doi.org/10.3390/ijerph192315506

Tubaro P., Casilli A.A. and Coville M. (2020) The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence, Big Data and Society, 7 (1). https://doi.org/10.1177/2053951720919776

UNESCO (2005) The precautionary principle, World Commission on the Ethics of Scientific Knowledge and Technology. https://unesdoc.unesco.org/ark:/48223/pf0000139578

van Bekkum M. and Borgesius F.Z. (2021) Digital welfare fraud detection and the Dutch SyRI judgment, European Journal of Social Security, 23 (4), 323–340. https://doi.org/10.1177/13882627211031257

von der Leyen (2020a) Press remarks by President von der Leyen on the Commission's new strategy: Shaping Europe's digital future, 19 February 2020. https://ec.europa.eu/commission/presscorner/detail/nl/speech_20_294

von der Leyen (2020b) Shaping Europe's digital future: Op-ed by Ursula von der Leyen, President of the European Commission, 19 February 2020. https://ec.europa.eu/commission/presscorner/detail/es/ac_20_260

von der Leyen (2023a) Speech by President von der Leyen at the 15th Congress of the European Trade Union Confederation, 25 May 2023. https://ec.europa.eu/commission/presscorner/detail/en/speech_23_2926

von der Leyen (2023b) Speech by President von der Leyen at the Pulse Women Economic Network, via video message, 5 September 2023. https://ec.europa.eu/commission/presscorner/detail/en/speech_23_4351

Zhang B. Anderljung M., Kahn L., Dreksler N., Horowitz M.C. and Dafoe A. (2021) Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers, Journal of Artificial Intelligence Research, 71, 591-666. https://doi.org/10.48550/arXiv.2105.02117

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

All links were checked on 19.01.2023.

Cite this chapter: Ponce del Castillo A. (2024) AI: the value of precaution and the need for human control, in Ponce del Castillo A. (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 2
Navigating AI beyond hypes, horrors and hopes: historical and contemporary perspectives

Vassilis Galanos and James K. Stewart

1. Introduction

Since around 2010 the term 'artificial intelligence' (AI) as a label has been used in an increasing variety of domains, from workplace to entertainment and from healthcare to trading and the military. In 2023, the commercialisation of 'generative AI' made it a constant on newspaper frontpages and in corporate boardrooms. This is the latest resurgence of a term coined in the 1950s (McCarthy et al. 2006) for application to contemporary data-intensive computational systems. The term's perennial and more recent popularity seems to reflect the way it is both sufficiently ambiguous for many people to adjust it to their needs and agendas; and sufficiently precise to act as a robust springboard for human speculation.

This chapter introduces some of the different ways that the concept 'AI' has evolved and is currently used in a way that can be useful to non-AI practitioners, workers, researchers and members of the general audience, either out of curiosity or from a social demand that expects them to incorporate new computer technologies in their everyday routines.

In the following sections, we show that AI is used to label five different phenomena: (a) a set of scientific fields exploring the nature of human cognition and the potential for machines that demonstrate similar characteristics; (b) computational methods for creating such tools or improving the scientific field; (c) a range of practical applications of technology combining computational systems, data and physical machines within particular sociotechnical contexts; (d) a rhetorical device to shape markets and policy agendas; and (e) a concept to explore the human condition and a society that is increasingly ordered by its reliance on machines. This 'interpretative flexibility' (Pinch and Bijker 1984; Bakker et al. 2011) of the term allows it to be used simultaneously about research and debates over machines that may replicate, replace or surpass human cognitive abilities and as a catch-all term to label recent advances in computational technology. The term is used almost arbitrarily to capture and promote expectations, hopes, concerns and ulterior motives in order to problematise or promote new computer technology and uses. At certain moments in the past and present, other words have been used to describe exactly, or almost, the same issues: automation, electronic brains, computing, robotics, machine learning, big data, informatisation, machine intelligence (or machine vision/translation), algorithms and smart technologies, among others – before being (temporarily) relabelled AI. The use of the term AI to frame the issues today reflects all its future-oriented baggage, invoking hopes and fears but also obscuring other mundane issues with and uses of the same technology. AI thus offers a very intriguing case study in this respect due to the vast number of governments,

corporations, academic researchers, vendors and users who have an interest in using, working on and with, regulating, and having fun with AI.

What this chapter wishes to stress is the slippery, problematic use of the term AI. To quote Apple Macintosh co-developer Alan Kay, 'technology is anything that isn't around when you're born' (Kay 1996, in Frand 2000). New technologies have always fascinated and abhorred: the reader can think of seventeenth and eighteenth century automata, clocks, electricity, aether, steam power or telegraphy. An imagined future AI, or today's latest computer breakthroughs that imitate or replace elements of human expression and communication, perhaps triggers this response more than other technologies. The 'AI effect' has to do with the observation, since the 1960s, that every time AI-as-science accomplishes a practical achievement, the latter is not considered to be AI anymore, but discarded as 'mere' computational automation and not genuine 'intelligence' (McCorduck 1979). The rest of AI is left to the future, with the following two sarcastic definitions being applicable: 'AI can be defined as the attempt to get real machines to behave like the ones in the movies' (Beale, quoted in Sloman 2003: n.p.); and 'AI is whatever hasn't been done yet' (Tesler, misquoted in Hofstadter 1979: 601[1]). The chapter thus aims to offer advice that, in different settings, technologies elsewhere understood as AI might not be presented as such due to their intended or unintended banalisation – and at the same time, that technologies elsewhere not understood as AI will be presented as AI in order to meet specific expectations.

The chapter is divided into two main parts, based on prior historical research on AI (Galanos 2023) and a long-standing career in the field of the sociology of computing (Stewart 1998; Collier and Stewart 2022). The first is a historical presentation of how AI has been understood over the last seven decades; and the second a disambiguation of what contemporary debates on AI are about, including the way that today's real technologies are obscured by this long-term technical and social exploration of AI's imaginaries. It aims to help distinguish between these five different uses of the term 'AI' and how debates and developments in fields thus labelled shape our understanding of its contemporary relevance. An increasing number of AI scholars and commentators have been warning about the use of 'AI' for the misleading associations it produces and reproduces, having to do mostly with anthropomorphism and zoomorphism; that is, the idea that machines can think like, or surpass, humans (Salles et al. 2020). This chapter justifies such warnings by situating AI in its historical perspective while additionally explaining some of its key functions.

2. Does AI even exist? The implications of using this terminology, from artificial general intelligence to AI-phobia

Galanos (2018) argued: 'AI does not exist'. Most AI researchers will agree that there is no agreed definition of AI chiefly because there is no agreed definition of intelligence – we cannot define (and create) an artificial version of something that is not understood.

1. Hofstadter misquoted personal communication with his associate Larry Tesler, a well-known computer scientist who, in the future, went on to correct Hofstadter: 'What I actually said was: "Intelligence is whatever machines haven't done yet"' (Tesler 2024).

Some AI researchers will argue that it is precisely by being able to create artificial models of something as complex as intelligence that humans will be able to understand what the latter is. But the confusion does not end here as the division between artificial and natural is also ill-defined. From a naturalist's perspective, everything is natural, including creations by humans, as both belong to nature. From a constructivist perspective, everything that humans talk about is part of human experience and interpretation; thus, even natural phenomena are described according to human-made languages and human perception limitations. Therefore, any understanding of intelligence (human, machine or otherwise) cannot be understood but in artificial terms. Nevertheless, this debate does not prevent us using these words to describe intellectual applications, usually susceptible to different degrees of automation. The choice of words, however, has played an important role in AI becoming a pole of attraction, adopted as such in various contexts (Stewart 1998).

The fascination with creating a mirror image of the human has existed since ancient religious myth-making (in Greek, Abrahamic, Nordic and East Asian folklore) and reflects a fundamental reflection on the human condition. The creation of tools that replace or augment human faculties, such as wheels, clothing, cosmetics, hammers or writing, can be thought of as 'intelligence' – an ability to solve tasks with sufficient flexibility and resilience to circumstances: a narrow understanding of intelligent 'acts' might allow for a broad definition of AI as anything that mimics or augments that human faculty. However, an understanding of intelligence as the totality of these skills might demand a definition of AI as only that machine which is capable of self-awareness, sentience, intentionality, consciousness and instinct, and which completes tasks as humans do and knows or feels why. For some, this is the 'holy grail' of AI, sometimes referred to as 'artificial general intelligence' (AGI; Goertzel et al. 2010). That scientific articles and indeed an entire branch of AI research (practical and philosophical) is devoted to the development of AGI seems to encourage the mass media to generate hype about every recent achievement in computing, thus allowing researchers or commentators with or without expertise in AI or cognitive science to speculate whether this is an instance of sentience (Selwyn and Gallo Cordoba 2022). From there it is easy to bridge to philosophical speculation on the human condition: what happens when a machine achieves a status of intelligence that exceeds human levels, something that is defined according to different scholars' terminological flavour as 'ultraintelligence', 'superintelligence' or 'singularity' (Good 1965; Eden et al. 2012; Hoffmann 2022). These linguistic choices, especially popular between the late 1990s and early 2010s (in the case of singularity) and between 2010 and 2020 (in the case of superintelligence) are now further assisted by the visual overlap between the acronyms AGI (i.e. human-level or superhuman-level machine intelligence) and GAI (generative artificial intelligence, i.e. AI applications that are generating statistically new rearrangements of existing data based on prompts and models; more on these terms below[2]).

2. Anecdotally, one of the authors encountered on Twitter a job advertisement to research the impact of AGI at work, shared by an AI researcher who abstains from AGI debates. It was proven after relevant commentary that the post's author meant GAI, however, the recent hype about both terms generating an unconscious typographical slip of some magnitude.

While we do not subscribe to such hypotheses in practice, it is clear that they reflect a perennial human need to reflect on who we are and our role in the universe, as containers of a nearly religious hope for immortality (once we manage to create immortal machines that think like us, we will be able to transfer our sentience into machines), but also as scaremongering suggestions about a world where machines will replace humanity as part of an evolutionary survival of the fittest. The latter suggestions do not only create an AI-phobia on behalf of users and regulators, but also an AI-phobia-phobia on behalf of AI researchers who might choose to rebrand their research so as to avoid these potential fears. This has been the case historically with AI being rebranded in order to attract funding either because of excessive fear about AI or due to AI researchers' inability to deliver grandiose promises (the 'AI winter'). It seems that, today, the success of mundane applications of AI is balancing out past losses and potential fears; however, the latter are keeping the term sufficiently hyped. The purpose of this explication is to heighten awareness about the importance of the intended or unintended rhetorical effects of linguistic choices – it matters how we describe AI. John McCarthy, the person credited for coining the phrase AI, admitted during a 1973 television panel that: 'I invented it because we had to do something when we were trying to get money for a summer study' (BBC TV 1973).

3. Foundations – from AI as science to AI as technology

3.1 AI as scientific endeavour

Despite Alan M. Turing's 1950 term, 'computing machinery and intelligence', we find the first recorded use of 'Artificial intelligence' in August 1955 when John McCarthy, a young Assistant Professor in Mathematics at Dartmouth College, with the assistance of Marvin Minsky, Claude Shannon and Nathanael Rochester, submitted a funding proposal to the Rockefeller Foundation aimed at synthesising and clarifying the simultaneous advances in 'electronic brains', 'thinking machines' or, more formally, 'automata studies' and 'complex information processing', as a unified field termed 'artificial intelligence'. The proposal did not provide a specific definition although, after the term was proposed, it read: 'The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it' (McCarthy et al. 2006: 14). Broadly this is what is referred to as 'symbolic' AI, where data representing things in the world is manipulated according to pre-established rules (this approach has been widely referred to since the mid 1980s as 'good old' fashioned AI' or GOFAI (Haugeland 1985) to disambiguate from the approach which flourished later and, indeed, since 2010.

Two years later a second approach to electronic computers was championed by one of Minsky's childhood friends, Bronx High School of Science schoolmates and eventual peers, Frank Rosenblatt: that of the perceptron model in which a perceptron is an 'artificial neuron'. Although not named AI, the method outlined is essentially the technical basis of the chiefly inductive, data-driven pattern recognition techniques which make up the vast majority of contemporary AI in the form of machine learning and deep neural networks. In Rosenblatt's explanation: 'The proposed system depends

on probabilistic rather than deterministic principles for its operation, and gains its reliability from the properties of statistical measurements obtained from large populations of elements. A system which operates according to these principles will be called a perceptron' (Rosenblatt 1957: 2). That meant that a machine could operate based on several examples (hence, machine learning) instead of following very rigid sets of rules. It also meant that a childhood friendship and eventual rivalry would tacitly dominate the massive contemporary landscape of AI.

Conceptually, the two approaches signify that the assistance of the machine replication of intelligent processes can shed light on the nature of human intelligence, one side of which is rule-based, hierarchical, reasoning through sequences of deductive syllogisms; and the other that of experiential learning based on trial-and-error according to interpretation of givens ('data' in the Latin sense of the word). We, and our machines, either know the world by learning to follow instructions or by looking at many examples.

A description of AI by Minsky from that time is often employed as its definition: 'artificial intelligence, the science of making machines do things that would require intelligence if done by men [sic]' (Minsky 1968: v). Minsky promoted this field as a scientific rather than a technological or philosophical one ('the science of…'), suiting researchers who wished to belong within an established scientific environment. Secondly, one cannot but notice that 'intelligence' in 'AI' referred to a technical rendition of human intelligence as a prototype (and probably male). Most practical AI researchers have employed very narrow definitions of intelligence in order to evaluate their work – definitions which inherited the scientific racism and classism of IQ tests (Aylett and Vargas 2021: 110-11). Such definitions of intelligence have narrowed the research scope as to exclude intelligence derived from non-human animals, plants or emergent intelligent behaviour between collectives through interaction.

For nearly two decades, the scientific community prioritised the cognitive aspect of AI-as-science and the symbolic approach over applied robots, practical applications and the 'perceptron' model. However in the 1970s successful applications of industrial robotics, as well as the flourishing virtual environments of early videogames (themselves applications of advanced, for the time, digital problem-solving and symbolic AI), fuelled new futuristic speculation and re-merged AI and robotics with numerous writers and scientists offering new definitions of the scope of the field. Computer scientist and philosopher Aaron Sloman, for example, aimed at extending the scope of AI to encompass more grandiose goals as a field of theoretical inquiry consisting of three key domains: (a) 'theoretical analysis of possible effective explanations of intelligent behaviour'; (b) 'explaining human abilities'; and (c) 'construction of intelligent artefacts' (Sloman 1978: 17). Cognitive scientist and philosopher Douglas Hofstadter saw AI as an opportunity to test the limits of and the bridges between antithetical concepts such as inflexible computation and flexible creativity:

> The seemingly unbreathable gulf between the formal and the informal, the animate and the inanimate, the flexible and the inflexible. This is what Artificial Intelligence (AI) research is all about. And the strange flavor of AI work is that people try to put

> together long sets of rules in strict formalisms which tell inflexible machines how to be flexible. (Hofstadter 1979: 26)

By the 1970s, the political and economic environment had become very much in favour of narrow practical applications, then sparking in the following decade a domino effect of a global AI race between Japan, the USA and Europe, akin to contemporary US-China AI relations, and a resurgence of public interest in AI after a period of dormancy. During this time, computer research focused more on human-computer interaction rather than AI (Grudin 2009). AI experts thus aimed to reclaim the field and broaden its scope:

> Artificial intelligence is a subject that, due to the massive, often quite unintelligible, publicity that it gets, is nearly completely misunderstood by people outside the field. Even AI's practitioners are somewhat confused with respect to what AI is really about [...] Is AI mathematics? [...] Is AI software engineering? [...] Is AI linguistics? [...] Is AI psychology? [...] AI should, in principle, be a contribution to a great many fields of study. AI has already made contributions to psychology, linguistics, and philosophy as well as other fields. In reality, AI is, potentially, the algorithmic study of processes in every field of inquiry. [...] In some sense, all subjects are really AI. (Schank, in Partridge and Wilks 1990: 3-4, 13)

Nevertheless, computer scientist Alan Bundy attempted to preserve distinctions between 'different kinds of AI' which 'correspond to different motivations for doing AI' – these are: 'applied AI, where we use existing AI for commercial techniques, military or industrial applications'; the modelling of 'human or animal intelligence using AI techniques [...] called cognitive science, or computational psychology'; and 'basic AI' exploring 'computational techniques which have the potential for simulating intelligent behaviour' (Bundy, in Partridge and Wilks 1990: 216).

By 1995, Stuart Russell and Peter Norvig's *Artificial Intelligence: A Modern Approach*, which has become a standard textbook, created a quadripartite taxonomy of types of AI research: '1. Systems that think like humans. 2. Systems that act like humans. 3. Systems that think rationally. 4. Systems that act rationally' (adapted from Russell and Norvig 1995: 5). By the mid-1990s, there was agreement within AI communities that AI is a set of different methodologies focusing on the simulation or programming of intelligence.

3.2 Robotics, embedded AI and heralding a new era for AI

Since the 1990s, most debates around computational technologies have focused on the internet and its network metaphors and the liberating potential of millions of citizens connecting, communicating and generating their own content. Following the simple virtual entities of video games – computer generated characters that act as human opponents or assistants in virtual worlds of play – the idea of 'artificial agents', programmed to participate automatically in this online world, also took off. 'Bots' (stemming from 'robot') are online software agents, bringing to the fore a new distinction between embodied and disembodied 'AI' and the potential for intelligent

artificial agents to operate much more freely in their own 'native' context, the internet. Rodney Brooks, pioneer of the Artificial Life (or ALife) movement in AI and robotics, made a distinction between situatedness and embodiment in artificial systems in which variations of degree between the two exist across the two extremes: 'Under these definitions an airline reservation system is situated but it is not embodied. A robot that mindlessly goes through the same spray-painting pattern minute after minute is embodied but not situated' (Brooks 2002: 51-52). Brooks justified his expansion of AI's concept as to include a broader variety of types of intelligence: 'Judging by the projects chosen in the early days of AI, intelligence was thought to be best characterized as the things that highly educated male scientists found challenging. [...] The things that children of four or five years could do effortlessly [...] were not thought of as activities requiring intelligence' (Brooks 2002: 36). *A Beginner's Guide to AI* by specialist Blay Whitby crystallised that AI is 'the study of intelligent behaviour (in humans, animals, and machines) and the attempt to find ways in which such behaviour could be engineered in any type of artefact' (Whitby 2003: 1), thus offering a symmetrical approach to all three categories' types of intelligence but also emphasising the importance of intelligence being 'contained' within an artefact.

Computer scientist Nils J. Nilsson offered what appears to be the most comprehensive definition, appearing on the first page of *The Quest for Artificial Intelligence: A History of Ideas and Achievements*:

> For me, artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight *in its environment*. According to that definition, lots of things – humans, animals, and some machines – are intelligent. Machines, such as 'smart cameras,' and many animals are at the primitive end of the extended continuum along which entities with various degrees of intelligence are arrayed. [...] For these reasons, I take a rather generous view of what constitutes AI. (Nilsson 2010: xiii)

Nilsson's historical work helped the observer to look through the rear-view mirror of AI and understand all those achievements that, if it were not for the 'AI effect', would be part of AI, albeit that they are now part of our mundane interactions with computers and smartphones, among others: the heuristic search for shortcuts; videogames; search engines and search databases; machine translation; computer vision and object detection and recognition (including facial); robotics; speech recognition and processing; natural language processing; semantic networks; probabilistic reasoning; machine learning; recommender systems; GPS; and ubiquitous computing (such as the internet of things).

Nilsson's book was released in 2010, just one year after a major success in the field of neural networks was accomplished. The perceptron or connectionist AI method originally proposed by Rosenblatt and, until then, chiefly underestimated by key AI scientists (and supporters of GOFAI) due to the lack of sufficient data to offer a basis to provide sufficient training examples, resurfaced due to the vast amounts of data becoming available online from user-generated content in the time of a highly unregulated internet where privacy concerns were overshadowed by the liberating ideology of oversharing. A series of techniques that, in the 1960s and today, would

be called AI were writing very successful stories in the late 2000s but without then employing the term 'AI'.

3.3 The technology turn: AI as socioeconomic future

By the 2000s, microprocessors were becoming sufficiently powerful finally to allow the significant practical implementation of the 'perceptron approach' of brain-inspired neural networks. These neural network approaches to machine learning enable probabilistic and statistical software models to be created directly from sets of empirical data describing a real world problem. An application of a computational method advanced by Geoffrey Hinton and his colleagues in 1985 (Rumelhart et al. 1985) called the 'back-propagation algorithm' (more on these terms in section 4.2), enabled his team in the late 2000s and early 2010s (Krizhevsky et al. 2012) to extract patterns and build computational models that would recognise images without predefined rules, using a large database of annotated images collected from the internet called ImageNet. The assembling of collections of 'training data' was facilitated by the explosion of content being produced and shared as part of the 'Web 2.0' boom, the post-2000 period when user-generated content was considered to be a revolutionising, bottom-up approach for internet governance. The main instigator of ImageNet competition was computer scientist Fei Fei Li (Li et al. 2009; Deng et al. 2009) as an attempt to classify image databases semantically. However, nowhere in the first two publications are the terms 'AI', 'machine learning' or 'neural networks' to be found. Nevertheless, this technique quickly started to facilitate increasingly better speech and visual object recognition as well as improving internet search and recommender systems.

While Web 2.0 discourses were focusing around well-defined 'information', this turn to 'data' as of 2010 was very much a feature of its time – businesses, science, libraries and governments were starting to generate and accumulate vast banks of data, generated in 'real time', but with few means or understanding of what could be done with it. While much of the data was locked away in organisational systems, an increasing amount was available seemingly 'for free' thanks in part to the ideology and practicalities of the open data movement and broader discourses about the internet as a realm of free communication and information (Wyatt 2021). While information scientists in the 2000s were concerned about the environmental and cognitive danger of information flux, data scientists working on machine learning approaches that would benefit from access to large datasets saw the opportunity not only to take their own work forward, as in the ImageNet case, but to turn this into a new business opportunity (Mayer-Schönberger and Cukier 2013). From 2010 companies such as IBM championed 'Big Data' as a promise that future computing would unlock value for business and government, opening a new paradigm in computing and data.

The data and computing systems infrastructures built for the new internet economy and society of the 2000-2010s soon had to rely largely on economic concentration in the hands of 'Big Tech' – the major computer companies like Microsoft, Baidu, Apple, Facebook/Meta, Tencent and others, whose business raison d'être is the exploitation of computers and the search for novel profitable techniques and uses that could be delivered

at vast scale via platform computing. To make smart computer-based services more attractive and easier to use, computers needed to be brought closer to humans – to be able to operate in the human world of vision, music, speech, language, the body, human emotional behaviour, on roads and in factories, all of which required new techniques. The continued growth in the speed of microprocessors and the breakthroughs in computing outlined in the previous section, when applied to Big Data, started to make this practical and the label 'AI' re-emerged as a marketing and investment-orienting concept at the service of corporations, capitalism and governments seeking new solutions and positive messages about the future after the 2008 financial crisis (Mazzucato 2020).

It was during this period that concerns about technological singularity and artificial general intelligence resurfaced on the part of science commentators such as Stephen Hawking and Elon Musk, credible as scientists and technologists but whose limited knowledge of AI stirred a debate in the mid 2010s around the potential need to regulate AI and robotics as autonomous agents (Galanos 2019). After a series of failed policy attempts to understand AI without its complex historical and labyrinthine components, novel definitions were offered on behalf of policy bodies in the light of the term being appropriated for corporate interests, thus inviting regulation. Since 2020, the landscape appears to have culminated in considering AI as a technical field based on data-driven algorithmic applications that may influence various domains, with AI's philosophical history facilitating the excitement and public rhetoric but being outside what regulation oversees and what commercial applications practically do.

In order to develop law, regulation and policy, definitions are required – what counts as 'AI' for the purposes of control or support. This has not proved to be straightforward; and most policy-level definitions are found to be unsatisfactory due to their overly deterministic undertone of AI as a transformative or hazardous force, mostly referring to Big Data applications, rather than being rooted in a rules-based approach. After plenty of negotiation around AI's definition, the European Commission's appointed advisory group for an AI strategy, the High-Level Expert Group on Artificial Intelligence (HLEGAI), offered what appears to be the first distinction between AI as a field of technical application and as a scientific discipline, with an emphasis nonetheless on machine learning:

> Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. […]
>
> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of

> all other techniques into cyber-physical systems). (High-Level Expert Group on Artificial Intelligence 2019: 8)

AI's purposes either for the comprehension of intelligence or for the building of human-level intelligence have been lost from the definitional landscape.

This group's discussions evolved into the first comprehensive legal framework for regulating AI industry, business, research, placement and usage according to a scale of potential risk. The proposed 'single future-proof definition of AI' (European Commission 2021: 3) was developed in response to consulted stakeholders who 'mostly requested a narrow, clear and precise definition for AI' (European Commission 2021: 8) and moved away from HLEGAI's broad and highly encompassing definition. For the Commission's 2021 understanding of AI, an '"artificial intelligence system" (AI system) means software that is developed with one or more [...] techniques and approaches [...] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with' (European Commission 2021: 39). Robotics and other types of hardware have vanished, while AI as a scientific field is not part of any of the discussions within the 100-page document. Elsewhere, AI is presented as 'a fast evolving family of technologies that can bring a wide array of economic and societal benefits across the entire spectrum of industries and social activities' (European Commission 2021: 1; similar variations on pages 18, 34).

History shows, however, that technology and science are modified by the different social groups who adopt it or have an interest in it. The complexity of the factors involved in the terminology, paired to the high adoption rates of the term, also call for modified versions to accommodate the positionality of the relevant social groups that have an interest in AI but no direct access to shaping it at policy level while nonetheless having to represent their interests agonistically within AI datasets. Such versions of AI are inspired by movements of social justice and intersectional appreciations of what is of social importance, and are sensitive to datasets' representations of various intersecting social categories. Three examples reflect how different these social groups are:

(a) 'Crip AI': inspired by mobility engineering design, being an approach to AI that is 'understanding and working with friction', resisting 'the automation of accessibility' and valuing instead 'the care and expertise provided by humans in their interface with machines (Hickman, in Manouach and Engelhardt 2022)

(b) 'Indigenous AI': lacking a specific definition, admittedly because 'a single "Indigenous perspective" does not exist', this approach emphasises the importance of involving underrepresented indigenous modes of thinking, being and learning in AI techniques, applications and methodologies, specific to places and contexts, and in relation to non-human species (Kite, in Manouach and Engelhardt 2022)

(c) 'Tactical AI': counteracting AI practices that 'serve administrative and managerial agendas [...] often in service of capital, and often in order to curb divergent elements within a given system', tactical AI 'manifests in oppositional, diversionary, and critical

approaches that resist strategic implementations of AI technologies', sometimes involving 'hacking or reverse engineering of mainstream AI' (Zeilinger, in Manouach and Engelhardt 2022).

Some may further argue that such definitions stem or are encompassed within the broader design framework of 'human-centered AI' (Shneiderman 2022); however, such phraseology does not take into account the anti-speciesist critique of AI and the inclusion of non-human animals in the debate (Singer and Tse 2022).

Our key argument, therefore, is that AI's long historical negotiation about earning respect as a scientific field has involved optimistic and grandiose will as a means of expanding its scope to a multitude of domains. The unsuccessful nature of attempts to realise such grandiose promises as well as rivalries within the discipline have led to periods in which AI research continued, but disguised under different names, online applications being one. Recent successes have been associated with just a small portion of the broader AI field and have led the majority policy, corporate, military and research understandings of the term to perceive it as a purely technical application. It is the rest of the much wider portion of AI research, however, dealing with the 'big questions' concerned with understanding the essence of intelligence and the possibility of constructing artificial beings, that fuels much of the hype and, at the same time, obscures plenty of the inner technical and social mechanisms that lie underneath contemporary AI applications. These are examined below.

4. Contemporary AI

Contemporary use of the term 'AI', that some of the policy definitions above try to capture, refers mostly to a set of technologies and applications based on 'neural networks' – layers of interconnected simulated artificial neurons. These software and hardware techniques are embedded in consumer products and improve commercial operations, support scientific research and advanced videogames, automatically translate conversations, drive experimental cars and are leading to the imminent transformation of warfare. The 'generative' language, code and image tools advanced in the mid-2010s and commercialised by 2022-3 have attracted the immediate attention of the press, inventors and investors, and provided a very public vehicle in which to communicate and explore many of the well-known benefits and controversies of data-driven computational modelling. Despite any attempt at a wide or narrow definition, contemporary AI is not one technology, a single technical breakthrough, but a combination of many technical developments and their incorporation into existing systems and processes. The machine learning, or computational, approach to finding patterns in data and building analytical and actionable models has been established for decades (Mitchell 1980) and been incorporated into many industries. As outlined above, what stands out as novel in current AI are the increases in the speed and storage capacity of microprocessors and memory, and the scaling of computational techniques that enable the exploitation of the vast amounts of data that are generated and collected by the informatised economy.

In this final section we discuss some of the key ideas that are associated with contemporary AI; namely the algorithms, statistics and models in machine learning applications of AI. These are the necessary technical components to make AI function. A typical conversation when talking about AI is a dismissal of the term via the use of one of the following terms, for example: 'it is not AI, it's just an algorithm!' or 'it is not AI, it's just statistics!'. Such responses have credence but, linguistically, they also refer to AI as a potentially human- or superhuman-level entity. What this chapter stresses in this respect is that all these 'just's are the main operational material of AI.

4.1 Algorithms and statistical models

During the 2010s the concept of 'algorithm' came to dominate much commercial and critical debate – algorithmic decision-making could be faster, more accurate and predictable than older processes relying on human decision-making, but also carry the biases of social injustice (O'Neil 2016). 'Algorithm', however, is a pre-modern concept describing mathematical formulas or functions deriving its name from the Persian astronomer and mathematician, Muhammad ibn Musa al-Khwarizmi (Howard 2022; Barany 2023), and the concept has been deployed consistently across various domains in computer science. A typical explanation of an algorithm would not differ much from the following, widely cited, one from 1967:

> An algorithm is a system of symbols connected according to pre-established rules. […] [T]he algorithmic system becomes a calculating machine, as conversely every calculating machine is materialization of an algorithm. Suitable data being fed in, the machine runs according to the pre-established rules, and eventually a result drops out which was unforeseeable to the individual mind with its limited capacities. This is the essence of mathematical reasoning, prediction in science and control of nature in technology. (von Bertalanffy 1967)

In debating the meaning of 'suitable' as implying the power to deem data suitable, the algorithm as a concept took on a life of its own, the crystallisation of value and power – the choice of algorithm could make or break a tech company. Like AI, algorithms became objects of deep public mystification, as imagined entities that 'can read' a customer's mind (Natale 2019). They became the focus of critical concern over alienation, manipulation, discrimination and technology out of control in the form of automated decision-making or the promotion of 'fake news' and harmful ideas (Pasquale 2020).

Despite the protestations of tech companies and developers that algorithms are 'neutral' (Kenneth and Rubinstein 2023), there is a whole range of real world outcomes demonstrating the weakness of this argument. Poor practices in data collection and model training, cooperative interests and trade-offs, weak governance and basic misunderstandings about how algorithm development involves choices (Raji et al. 2020) have undermined the idea of the algorithm as neutral and simply a tool. Even relatively simple algorithms have been found to create or entrench social biases when deployed within systems applied to people. When deployed at huge scale, the emergent effects of algorithms such as recommender systems produce unintended emergent

effects. Simple algorithmic decision trees based on negotiated rules are just the tip of the iceberg of algorithmic systems. The entire thrust of modern computing has been to use pattern-finding algorithms to generate complicated models that simplify even more complex systems, from natural language to consumer behaviour. As Lepore showed, today's algorithmic philosophy of 'if then' is nevertheless just an extension of a long trajectory of belief in this methodology, reflecting an ideology of a society that can be controlled on the basis of a faith in algorithmic instructions (Lepore 2020). The development of algorithms based in machine learning, often called 'pre-trained models', especially those using deep neural nets where simple rules are substituted by probabilistic statistical models that are impenetrable 'black boxes' due to the multitude of algorithmic processes for pattern matching, may allow computers to address many previously intractable tasks. However, they reinforce the many complex undesired biases recorded in the 'training data' that are fed into the system, thus necessarily introducing new undesired biases.

Models are ways in which humans understand (or attempt to understand) the world, from simple heuristics, stereotypes and rules of thumb to scientifically tested theory and laws. Models enable us to find patterns of order in complex systems, but are necessarily a simplification and, at that, a simplification based on the goals of those making and commissioning them (Tseng, in Manouach and Engelhardt 2022). The development of models in various fields often arises from a combination of research, practical application, negotiation and expediency. These models, through the computer's iterative processes, determine numerical weights for each variable or dimension. They can then be utilised to classify existing data, identify errors and biases or generate new data, such as predicting future events or determining the most probable word in a given text. Models typically serve as decision-making tools, predictors or reflective frameworks, providing simplified guidelines based on established principles or empirical observations. Machine learning thus employs basic computer algorithms or instructions to create or enhance mathematical models that capture relationships between data points representing real-world phenomena of interest. While deductive modelling is feasible when well-established 'first principles' exist, many scenarios require a blend of principles and empirical data that are only partially established.

It is precisely within this idealistic understanding of the model that problems in the public perception of AI models arise: a belief in models seems to imply belief in prior tested principles – hence humans are likely to believe an algorithmic recommendation based on the perceived robustness of the first principles constructing the model behind it. Such a view, however, overlooks the many data sources that are left outside the construction of the model as well as the ulterior motives for selecting one type of model instead of another. This has led several contemporary computer scientists to encourage machine learning practitioners to interrogate their models and data sources, accompanying the release of a machine learning model with 'model cards' documenting the model's history, purpose, data sources and possible limitations or unwanted biases (Mitchell et al. 2019; Gebru et al. 2021). Achieving this ideal balance between tentative usefulness and accuracy of results and responsible documentation may involve the definition of acceptable simplifications, the establishment of accuracy benchmarks and the development of limits for appropriate usage. The determination of what is deemed

acceptable and accurate enough becomes a social, economic or political consideration, influenced by broader societal factors.

4.2 Machine learning techniques: a simplified introduction

Although this is a vast, sophisticated and developing field, there are three main techniques for machine learning that are useful to grasp: supervised learning, unsupervised learning and a combination of the two known as reinforcement learning. The following descriptions are meant to be simplified guides for the interested reader and stem from the authors' experience (for greater technical detail, see Nilsson (2010: 443-449) and Goodfellow et al. (2014)).

'Supervised learning', a common approach in machine learning, involves labelling data objects (e.g. images or sounds) according to predetermined categories (e.g. pictures of cats and dogs) typically assigned by human annotators – in the vision of the Web 2.0 era, internet users were annotators of their own content, for example by adding hashtags and descriptions to pictures uploaded online. Engineers employ various techniques and their own expertise to train computational models using these labelled data samples until the model accurately assigns labels to unseen instances of, in this example, cats or dogs. The inclusion of labels guides the adjustment of weights in the model's equations until it satisfies the developer's criteria. While these systems often appear simplistic, they heavily rely on extensive preparatory work involving labour-intensive human efforts that sometimes goes unacknowledged regarding its value. Despite their potential for producing remarkable outcomes, these techniques necessitate significant expert involvement to identify the most suitable form of input data for a given problem. For instance, such techniques may be employed to train models either to detect cancer cells or to recognise facial features.

However, it is important to note that the patterns identified by the computational model to represent the input data are not predetermined or necessarily comprehensible to the developers. They may encompass factors like fur patterns, ear shapes or even extraneous details in the background that are unrelated to the intended subject. Consequently, these models may occasionally make errors and can be intentionally deceived, either during training or in operational use, as certain colours or shapes may 'confuse' the model's recognition abilities. It is noteworthy that, despite the existence of mathematical and computational frameworks for describing these issues, as previously mentioned, there is a propensity to anthropomorphise the explanations in everyday language of how these systems function.

The alternative approach to machine learning involves the utilisation of 'unlabelled' training data, which encompasses datasets representing the world without prior categorisation based on human-defined interests. This approach is referred to as 'unsupervised learning' and became particularly popular in the first half of the 2010s after a series of surprisingly convincing success stories. In unsupervised learning, the computer is tasked with identifying patterns within the data without explicit guidance on which ones are of *a priori* significance. It is assumed that relationships exist between

the features represented in the data, such as the connections between words in text or the interplay of behaviours and demographic characteristics within a population. The computer's objective is to uncover these underlying relationships.

The patterns derived from the multitude of relationships encoded within the data inform the setting of weights within the models' equations. Some such patterns may be of interest to the developers, potentially representing causal links within the system under investigation. However, many will consist of incidental correlations lacking substantive importance (as per the old dictum in statistics, correlation does not necessarily imply causation). As with supervised learning, the determination of what qualifies as important is a decision to be made by developers; that is, those who commission the work and, ultimately, society at large. Unsupervised learning methods can be applied to diverse tasks such as optimising truck routes or the efficient storage of movies on internet servers. While unsupervised learning requires substantial computational resources, it alleviates the considerable costs associated with human labelling efforts. In one sense, the relationships present in real-world phenomena are implicitly 'self-labelled' within the data, albeit in a probabilistic manner. By employing techniques like 'attention', unsupervised learning models automatically establish probabilities regarding missing words in a vast collection of texts or the likely colour of an adjacent pixel within an image. These probabilities are determined based on the discernible patterns observed in the original training data.

Both these approaches have drawbacks – one depends on expensive labelling, the other can find many interesting, but also many potentially uninteresting or confusing, patterns. Combining elements of both draws on their strengths but does not, however, eliminate the common problems. Supervised learning relies on the explicit labels provided by humans to guide the model's construction, conforming to predetermined classifications. In a similar vein, 'reinforcement learning' draws on pre-trained unsupervised models to achieve the same objective. To determine which outputs from the models are deemed valuable or socially desirable, techniques are required to assess and suppress undesirable outputs. Often, this necessitates the involvement of human evaluators, reintroducing the costly human element. However, this integration of (expensive) human judgment enables computers to tackle complex tasks that would otherwise pose considerable challenges, such as generating human language.

The most recent AI systems that analyse and generate language, called Large Language Models (LLMs), depend on these reinforcement techniques. Since the models are trained on trillions of input texts, images or other data sources, representing almost every possible language combination, then the outputs are limitless. LLMs produce very convincing language, but the economic value would seem to come from having that language communicate 'knowledge' or 'truth', to which they are not well suited. The challenge for developers is to limit the socially or organisationally unacceptable outputs, using as yet rather ad hoc techniques and running the gauntlet of criticism over what is 'acceptable' speech, being generated by an LLM. Neither does this description deal with the issue of the data that might be used to train the model in the first place. We know that there are vast sources of information, but how is this turned into forms that can be used to train a large model both effectively and efficiently, and in a way

that might avoid some of the difficulties of subsequent reinforcement retraining? These questions present themselves as ones faced by AI specialists and social scientists alike.

5. Conclusions

AI has produced many contemporary debates. Several need to be flagged before this chapter closes: the huge amount of work that still needs to be done to understand and control the newest AI systems to generate value and not undermine fundamental human values; the concentration of power in vast datacentres in the hands of a few companies; the energy costs of these datacentres in a time of climate crisis; the blatant disregard for the ownership and privacy of data being assembled to train large pre-trained models, and the rights of those who have worked to produce it; the massive global scale of deployment of AI via cloud computing, seamlessly and invisibly integrated into millions of apps and devices; the potential for harm to our society and politics, and the criminal and military uses to which it might be put; and the struggle by governments and firms to understand how it can be simultaneously regulated and exploited.

These controversies are technical, ethical, economic and cultural – some claim 'big AI' will become smaller, open source and less constrained by the need for hugely expensive datacentres, but will unleash dangerous technology into the hands of anyone with access to moderate computer power. Others struggle over how to translate the way the technology operates into forms that can be subject to the normal means of social and political control – among others the negotiation of contracts, limits of use, amenability to democratic debate, legal responsibility and liability, and testing and certifying.

For the public, and politicians influenced by the public, we return to the less practical and more philosophical reasons for why we explore AI – to learn what it means to be intelligent and to reflect on the human condition. The apparently magical demonstrations of automatic text and image generation, or the potential for a machine autonomously to select, target and kill a human being plunge us back into existential questions that are millennia old. For one, AI is not 'intelligent' (as a squirrel or as a human), 'conscious', 'self-aware' or 'sentient', but rather a combination of circuits, modelling techniques and machinery. Admittedly, many scientists will use such terms in their work out of metaphorical convention – but one must proceed with care when using these words across diverse contexts. In terms of problem solving, AI cannot tell the difference between spatiotemporal contexts (Sloman 2018) and thus does not 'decide' unless humans with a will to believe its suggestions let it recommend something. Furthermore, AI does not 'perceive' or 'learn' like humans do and its rationale echoes an ideal 'view from nowhere' (Baumgartner et al. 2023). Neither can AI build complex systems out of simple ones (no ideologies, no political systems, no embodied brains or genuine humour). Not all solutions are precisely applicable to all problems (a commercial recommender algorithm is not the same as a jurisdiction decision support tool) and, therefore, it is rather implausible for AI systems to perform higher knowledge abstractions, as several people tend to imagine, based on the – typical in AI – first step fallacy: one will not reach the moon as easily as one climbed a tree (Dreyfus 2012).

While there is little chance of superintelligences or sentient machines emerging within anyone's lifetime, these thought experiments about distant future concerns are distracting given the real and urgent problems being experienced today. Both of these frighten people and both are being used rhetorically to muddy the waters of the political and commercial games of the most powerful global interests. We therefore need to remember constantly the many ways that 'AI' is used as a term and to consider carefully what level we are at and what are the salient issues that need to be addressed. There is no one 'AI' and no inevitable path of development and control – it can be deployed for economically efficient purposes, future medicinal ones, or to gain military power or for social justice – but all this requires us to assemble actions and narratives at philosophical, scientific, technical and rhetorical levels.

References

Aylett R. and Vargas P.A. (2021) Living with robots: What every anxious human needs to know, MIT Press.

Bakker S., van Lente H. and Meeus M. (2011) Arenas of expectations for hydrogen technologies, Technological Forecasting and Social Change, 78 (1), 152–162. https://doi.org/10.1016/j.techfore.2010.09.001

Barany M.J. (2023) "Some call it Arsmetrike, and some Awgryme": Misprision and precision in algorithmic thinking and learning in 1543 and beyond, in Ames M.G. and Mazzotti M. (eds.) Algorithmic modernity, Oxford University Press, 31–44.

Baumgartner R. et al. (2023) Fair and equitable AI in biomedical research and healthcare: Social science perspectives, Artificial Intelligence in Medicine, 144, 102658. https://doi.org/10.1016/j.artmed.2023.102658

BBC TV (1973) The general purpose robot is a mirage lighthill controversy debate at the royal institution with Professor Sir James Lighthill, Professor Donald Michie, Professor Richard Gregory and Professor John McCarthy, Television video archive. https://www.aiai.ed.ac.uk/events/lighthill1973/

Brooks R.A. (2002) Robot: The future of flesh and machines, Penguin Books.

Collier B. and Stewart J. (2022) Privacy worlds: Exploring values and design in the development of the Tor anonymity network, Science, Technology, and Human Values, 47 (5), 910–936. https://doi.org/10.1177/01622439211039019

Deng J., Dong W., Socher R., Li L.J., Li K. and Li F. (2009) Imagenet: A large-scale hierarchical image database, paper presented at 2009 IEEE conference on computer vision and pattern recognition, Miami, 20-25.06.2009. https://www.computer.org/csdl/proceedings/cvpr/2009/12OmNy4r3R2

Dreyfus H.L. (2012) A history of first step fallacies, Minds and Machines, 22, 87–99. https://doi.org/10.1007/s11023-012-9276-0

Eden A.H., Steinhart E., Pearce D. and Moor J.H. (2012) Singularity hypotheses: An overview, in Eden A.H., Moor J.H., Søraker J.H. and Steinhart E. (eds.) Singularity hypotheses: A scientific and philosophical assessment, Springer, 1–12.

European Commission (2021) 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, 21.4.2021. https://digital-

strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Galanos V. (2018) Artificial intelligence does not exist: Lessons from shared cognition and the opposition to the nature/nurture divide, in Kreps D., Ess C., Leenen L. and Kimppa K. (eds.) This changes everything – ICT and climate change: What can we do?, HCC13 2018, IFIP Advances in Information and Communication Technology, vol. 537, Springer, 359-373. https://doi.org/10.1007/978-3-319-99605-9_27

Galanos V. (2019) Exploring expanding expertise: Artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018, Technology Analysis and Strategic Management, 31 (4), 421–432. https://doi.org/10.1080/09537325.2018.1518521

Galanos V. (2023) Expectations and expertise in artificial intelligence: Specialist views and historical perspectives on conceptualisation, promise, and funding, Doctoral thesis, The University of Edinburgh. http://dx.doi.org/10.7488/era/3188

Gebru T., Morgenstern J., Vecchione B., Vaughan J.W., Wallach H., Daumé H. and Crawford K. (2021) Datasheets for datasets, Communications of the ACM, 64 (12), 86–92. https://doi.org/10.48550/arXiv.1803.09010

Goertzel B. et al. (2010) OpenCogBot: Achieving generally intelligent virtual agent control and humanoid robotics via cognitive synergy, Proceedings of ICAI, 10.

Good I.J. (1965) Speculations concerning the first ultraintelligent machine, Advances in Computers, 6, 31–88. https://doi.org/10.1016/S0065-2458(08)60418-0

Goodfellow I.J., Shlens J. and Szegedy C. (2014) Explaining and harnessing adversarial examples. https://doi.org/10.48550/arXiv.1412.6572

Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair, S., Courville A. and Bengio Y. (2020) Generative adversarial networks, Communications of the ACM, 63 (11), 139-144. https://doi.org/10.1145/3422622

Grudin J. (2009) AI and HCI: Two fields divided by a common focus, AI Magazine, 30 (4), 48–57. https://doi.org/10.1609/aimag.v30i4.2271

Frand J.L. (2000) The information-age mindset: Changes in students and implications for higher education, Educause Review, 35, 14–25.

Haugeland J. (1985) Artificial Intelligence: The very idea, MIT Press.

High-Level Expert Group on Artificial Intelligence (HLEGAI) (2019) A definition of AI: Main capabilities and scientific disciplines, European Commission. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

Hoffmann C.H. (2022) A philosophical view on singularity and strong AI, AI and Society, 38, 1697–1714. https://doi.org/10.1007/s00146-021-01327-5

Hofstadter D.R. (1979) Gödel, Escher, Bach: An eternal golden braid, Basic Books.

Howard J. (2022) Algorithms and the future of work, American Journal of Industrial Medicine, 65 (12), 943–952. https://doi.org/10.1002/ajim.23429

Kenneth T. and Rubinstein I. (2023) Gonzalez v. Google: The case for protecting 'targeted recommendations', Duke Law Journal Online, 72. http://dx.doi.org/10.2139/ssrn.4337584

Krizhevsky A., Sutskever I. and Hinton G.E. (2012) ImageNet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, 25, 1097–1105.

Lepore J. (2020) If then: How one data company invented the future, John Murray.

Li F., Deng J. and Li K. (2009) ImageNet: Constructing a large-scale image database, Journal of vision, 9 (8), 1037. https://doi.org/10.1167/9.8.1037

Manouach I. and Engelhardt A. (eds.) (2022) Chimeras: Inventory of synthetic cognition, Onassis Foundation.
Mayer-Schönberger V. and Cukier K. (2013) Big data: A revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt.
Mazzucato M. (2020) Capitalism's triple crisis, Project Syndicate, 30 March 2020. https://www.project-syndicate.org/commentary/covid19-crises-of-capitalism-new-state-role-by-mariana-mazzucato-2020-03
McCarthy J., Minsky M.L., Rochester N. and Shannon C.E. (2006) A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955, AI Magazine (27) 4, 12. https://doi.org/10.1609/aimag.v27i4.1904
McCorduck P. (1979) Machines who think: A personal inquiry into the history and prospects of artificial intelligence, W.H. Freeman.
Minsky M. (1968) Semantic information processing, MIT Press.
Mitchell T. (1980) The need for biases in learning generalizations, Rutgers University. https://www.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf
Mitchell M., Wu S., Zaldivar A., Barnes P., Vasserman L., Hutchinson B., Spitzer E., Raji D. and Gebru T. (2019) Model cards for model reporting, Proceedings of the conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596
Natale S. (2019) Amazon can read your mind: A media archaeology of the algorithmic imaginary, in Natale S. and Pasulka D.W. (eds.) Believing in bits: Digital media and the supernatural, Oxford University Press, 19–36. https://doi.org/10.1093/oso/9780190949983.003.0002
Nilsson N.J. (2010) The quest for artificial intelligence: A history of ideas and achievements, Cambridge University Press.
O'Neil C. (2016) Weapons of math destruction: How big data increases inequality and threatens democracy, Crown.
Partridge D. and Wilks Y. (1990) The foundations of artificial intelligence: A sourcebook, Cambridge University Press.
Pasquale F. (2020) New laws of robotics: Defending human expertise in the age of AI, Harvard University Press.
Pinch T.J. and Bijker W.E. (1984) The social construction of facts and artefacts, or, how the sociology of science and the sociology of technology might benefit each other, Social Studies of Science, 14 (3), 399–441. https://www.jstor.org/stable/285355
Raji I.D., Smart A., White R.N., Mitchell M., Gebru T., Hutchinson B., Smith-Loud J., Theron D. and Barnes P. (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing, Proceedings of the 2020 conference on fairness, accountability, and transparency, 33–44. https://doi.org/10.48550/arXiv.2001.00973
Rosenblatt F. (1957) The perceptron–A perceiving and recognizing automaton, Report 85-460-1, Cornell Aeronautical Laboratory.
Rumelhart D.E., Hinton G.E. and Williams R.J. (1985) Learning internal representations by error propagation, in Rumelhart D.E. and McClelland J.L. (eds.) Parallel distributed processing, Volume 1: Explorations in the microstructure of cognition, MIT Press, 318–362. https://doi.org/10.7551/mitpress/5236.003.0012
Salles A., Evers K. and Farisco M. (2020) Anthropomorphism in AI, AJOB neuroscience, 11 (2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

Selwyn N. and Gallo Cordoba B. (2022) Australian public understandings of artificial intelligence, AI and Society, 37 (4), 1645–1662. https://doi.org/10.1007/s00146-021-01268-z
Shneiderman B. (2022) Human-centered AI, Oxford University Press.
Singer P. and Tse Y.F. (2022) AI ethics: The case for including animals, AI and Ethics, 3, 539–551. https://doi.org/10.1007/s43681-022-00187-z
Sloman A. (1978) The computer revolution in philosophy: Philosophy, science and models of mind, Harvester Press.
Sloman A. (2003) What is artificial intelligence?, University of Birmingham, School of Computer Science, 29 April 2003. http://www.cs.bham.ac.uk/~axs/misc/aiforschools.html
Sloman A. (2018) Huge, but unnoticed, gaps between current AI and natural intelligence, in Müller V. (ed.) Philosophy and theory of artificial intelligence, Springer, 92–105. https://doi.org/10.1007/978-3-319-96448-5_11
Stewart J.K. (1998) Interactive television at home: Television meets the Internet, in Jensen J.F. and Toscan C. (eds.) Interactive television: TV of the future or the future of TV?, Aalborg University Press.
Tesler L. (2024) CV: Summary. https://www.nomodes.com/larry-tesler-consulting/adages-and-coinages
Turing A.M. (1950) Computing machinery and intelligence, Mind, 59 (236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
von Bertalanffy L. (1967) Robots, men and minds: Psychology in the modern world, George Braziller.
Whitby B. (2003) Artificial intelligence: A beginner's guide, Oneworld.
Wyatt S. (2021) Metaphors in critical Internet and digital media studies, New Media and Society, 23 (2), 406–416. https://doi.org/10.1177/1461444820929324

All links were checked on 22.01.2024.

Cite this chapter: Galanos V. and Stewart J.K. (2024) Navigating AI beyond hypes, horrors and hopes: historical and contemporary perspectives, in Ponce del Castillo A. (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 3
In Humans We Trust: rules, algorithms and judgment

Hamid R. Ekbia

1. Introduction

Modern humanity is facing numerous challenges in all arenas of life – from the economy, politics, culture and the environment to law, healthcare and governance. Cracks are appearing on the seemingly consistent surface of social reality, revealing its irremediable fragility and evoking the image of an ice sheet during a thaw. Many of these have escalated to the level of crisis in the sense that we experience a constant presence of negativity in our lives. Humanity has accomplished much, but despite – or, depending on your outlook, perhaps because of – this, it has left many problems untouched. Not only that, every step of the way it has added new predicaments for human beings, for other species and for our globe not least through the reckless hubris emanating from its marvellous prowess in scientific and technical discovery. Individuals and communities face this situation in different ways, listening to experts and commentators for guidance. It is in this environment that AI has emerged as either the *cause célèbre* of these crises or as the solution to them, depending on whose views, whether of AI detractors or AI prophets, you take seriously. Both of these miss the mark by large margins because AI does not cause these issues, nor can it cure them on its own.

The detractors, those who see in AI the root cause of our predicaments, recommend withdrawal from technology. Their idea is that we can choose to disconnect from technology at will with little or no social cost or, even more naively, that taking leave from technology can help us deal with the growing number of challenges that we face as individuals and communities by falling back on our 'good old ways'. Little do the proponents of such ideas understand that withdrawal is an unrealistic option and that taking nostalgic refuge in a bygone era will not save the day for us. There was no such thing as an 'ideal' past where things were presumably in apple pie order, nor could the old ways work for our current issues, even to the extent that they did work in the past.

The prophets and devout advocates of AI, on the other hand, see in it an invisible hand, all too ready to be extended outward, with a purported magic ability to make things work. You have too many road casualties? Self-driving cars will reduce or eliminate them. Mysterious diseases? AI diagnostics can reveal them. Too many applications for jobs, admissions, benefits or parole? AI can screen them. Too many items on your daily agenda? AI assistants can sort them. Too long a history of legal cases for litigation? Computerised clerks can filter them. Fraudulent actors? AI detection systems can reveal their hand. Looking for voters to support your political campaign? AI tools can convince them. Yearning for romance? AI can find the perfect partner, even giving you the bots and robots that would play that role themselves – satisfaction guaranteed! Perceived

in this way, AI appears to be much more effective than its modern predecessor, the miraculous market, to take care of our issues. If that invisible hand could solve just one problem – coordinating buyers and sellers of commodities – AI can solve a whole slew of social, economic, political, professional and personal problems: essentially, anything under the sun that would lend itself to 'pattern recognition' and the logic of computing.

A case in point is the management and allocation of social benefits to different individuals and social groups. Local and national governments increasingly rely on so-called 'smart algorithms' to do this for them. A recent report by the Electronic Privacy Information Center, for instance, revealed the extensive use of algorithmic decision-making by government agencies in Washington DC, adding up to 29 automated systems used by 20 agencies. At the same time, the report highlights numerous loopholes and mistakes present in these systems – for instance, a 93% error rate in an unemployment fraud detection system that led to 40,000 false cases and 1.1 million false flags (Johnson 2022). Washington DC is just one example among many; software systems are currently used for all kinds of purposes by government agencies on different levels in the name of efficiency and objectivity.

Government agencies are not alone in their penchant for efficiency. Similar other reasons are invoked for the increasing application of AI and algorithms in other domains of social life. Prominent among these are accuracy, reliability, neutrality and transparency, which are often played, implicitly or explicitly, against such human fallibilities as social and cultural bias, cognitive limitations, emotional caprice and ethical evasiveness. Road accidents, for instance, kill 1.35 million people around the globe every year (CDC 2023) because drivers are distracted, exhausted, unskilled, under the influence or they simply make mistakes; medical errors due to misdiagnosis, drug interactions or sheer neglect annihilate another 2.6 million people every year in low and middle-income countries around the globe (WHO 2019), more than 250,000 in the US alone (Anderson and Abrahamson 2017); even judges, who are expected to be 'objective' in their views, are shown to be subject to 'judicial temperament', ruling differently at different times of the day because of their mood which might, in turn, depend on their physiological state (Danziger et al. 2011). Human beings are indeed fallible, biased, lousy, clumsy, capricious and evasive – and hence they are unreliable.

It is against the backdrop of such observations that advocates of AI propose technical alternatives as solutions to the increasing predicaments of modern societies. Given the numerous sources of human fallibility, they ask, why should we trust human beings with our social, personal and professional issues and not AI-enabled technologies that are not only devoid of bias and emotion but also more exact and efficient? In asking this question, such proponents are comparing, directly or indirectly, the reliability of rule-based algorithms with the unruliness of human judgment. Rules, in other words, play a central role in this comparison. But rules come in different varieties. In the examples mentioned earlier, the rules of traffic are different from the rules of medical practice, and both are yet more different from the rules of judicial decision-making. This brings up an important question: what are 'rules' and how have they come to play such a central role in human affairs?

2. Varieties of rules

Rules come in many varieties from rules of games, grammars and poetry to the etiquettes of behaviour in a school, a bus or an office, and from government regulations, laws of science and statutes of war to cooking recipes, design guidelines and computer algorithms. Behind this diversity of forms and contexts, there is a long history with discernible continuities and discontinuities.

In her book, the science historian Lorraine Daston (2022) identifies three broad clusters or kinds of rules that have operated throughout history across cultures: tools of measurement and calculation; models and paradigms; and laws. The first cluster includes rules followed over the centuries in the arts and crafts such as painting, cooking and baking, and even the medieval practices of artillery and fortification. The second cluster spans the traditional rules of conduct according to exemplary behaviours of specific individuals all the way to the more recent notion of 'paradigm' in the historical development of modern science (Kuhn 1962). This was, for instance, how rules were understood in the monastic Rule of St. Benedict, where the authority to apply the codes of conduct was vested in the abbot who was considered a rule, or a 'model', to be emulated by others. The third cluster covers the laws and regulations that govern social behaviour based on sanctions, for example the tax laws of ancient cities and medieval fiefdoms or the traffic laws of modern societies.

The historical development of rules can be understood as the interplay between these clusters on three oppositional dimensions which Daston describes as thin-thick, flexible-rigid and general-specific. Thin rules are concise descriptions and 'commands' that can be followed to the letter, without the need for intelligent interpretation and adaptation to context. They are, as such, 'exceptionless and infallible codes of conduct' (Chirimuuta 2023). Thick rules, on the other hand, are context-dependent and require extra capacity to see when and how best to apply them. But then that same capacity should be governed by a higher set of rules that would determine what type of context we are dealing with and this, in turn, would demand yet another rule which itself begs a rule of higher order … ad infinitum. This chain of rules leads to an endless regress and a paradox that has been known to philosophers for a very long time. In the eighteenth century, Immanuel Kant gave voice to this in the following manner:

> If the understanding in general is explained as the faculty of rules [Regeln], then the power of judgment is the faculty of subsuming under rules, i.e., of determining whether something stands under a given rule (*casus datae legis*) or not. General logic contains no precepts at all for the power of judgment, and moreover cannot contain them. For since it abstracts from all content of cognition, nothing remains to it but the business of analytically dividing the mere form of cognition into concepts, judgments, and inferences, and thereby achieving formal rules for all use of the understanding. Now if it wanted to show generally how one ought to subsume under these rules, i.e., distinguish whether something stands under them or not, this could not happen except once again through a rule. But just because this is a rule, it would demand another instruction for the power of judgment, and so it becomes clear that although the understanding is certainly capable of being

> instructed and equipped through rules, the power of judgment is a special talent that cannot be taught but only practiced. (Guyer and Wood 1998)

Kant then talks about a physician, a judge or a statesman who can have many fine pathological, juridical or political rules in their head but who can easily stumble in their application. He describes this lack of power as that which is properly called stupidity, reminding us that such a failing is not to be helped. What, then, distinguishes between a competent physician, judge or statesman and someone who is a mere stockpile or kitbag of rules? Or, by the same token, what distinguishes between a skilled craftsperson such as a master chef and a newbie who has never touched a spatula, between a master and a novice chess player, and between a good and a clumsy musician? Or, even more fundamentally, what is the difference between the conduct of an ordinary human being who adjusts to the demands of situations as they arise, with or without explicit knowledge of the 'rules', and someone who behaves 'mechanically', clueless 'like a robot' – to use modern parlance – in describing a rigid adherence to rules?

On the surface, these are different questions that speak to various domains of human behaviour – that is, expertise, skill and common sense. Deep down, however, they are related questions having to do with how some humans are better than others at applying, adjusting and following rules, not just having a knowledge of them. Ludwig Wittgenstein made an important distinction between following a rule and merely acting according to a rule. Rules, he argued, are not just causes that determine our behaviour in the same way that the law of gravity determines the behaviour of a planet around a star. Rather, they have a normative force because they provide measures of correct and incorrect behaviour. They are not simply dispositions that can drive behaviour like a mechanism; they have to be practised through experience. That is why we can invoke rules to explain and justify ourselves: 'To obey a rule, to make a report, to give an order, to play a game of chess, are customs (uses, institutions)' (Wittgenstein 1953: § 199).

This is how rules-as-models have been understood for an exceptionally long stretch of human history. The model as exemplar for following the rule could be a person, a work of art or simply a well-chosen example in grammar or algebra (Daston 2022: 8). Putting a high premium on experience and practice, such rules brought the head and the hand together, creating a balance between strict discipline and adherence to rules, on the one hand, and, on the other, the creativity and flexibility called for in applying them. A cooking recipe, for instance, leaves room for human experience in judging when the batter is thick, the flour is cooked enough or the cheese spread evenly on the crust.

To talk about these aspects of rules, Daston (2022: 36-38) draws on the notion of 'discretion' as a form of judgment 'which embraces not only knowing when to temper the rigor of rules but also matters of taste, prudence, and insight into how the world works, including the human psyche'. Discretion, as such, 'combines intellectual and moral cognition… but [it] goes beyond cognition… [It] is a matter of the will as well as the mind… Cognitive discretion without executive discretion is impotent; executive discretion without cognitive discretion is arbitrary'.

This way of thinking about rules-as-models, involving the moral capacity to exercise discretion, has been gradually edged out in modern times in favour of the more calculative notion of rules-as-algorithms. This shift started with, among other things, the introduction throughout the nineteenth century of mechanical machines that could perform some of the tasks of human beings and which culminated in the digital automation techniques of the late twentieth century. But it was not only the meaning of 'rules' that underwent a shift. In the process, the meaning of a whole slew of other related terms also underwent significant change. The concept of 'algorithm' was one such notion.

3. Varieties of algorithms

The term 'algorithm' goes back to the ninth century CE when the Persian mathematician, astronomer and geographer, Musa Al-Kharizmi, wrote a treatise on algebra, Indian numerals and astronomical tables. This text, which is considered the founding document of modern algebra, was translated into Latin in the twelfth century, and the earliest surviving manuscript in the Latin text starts with the words 'Dixit Algorizmi' ('Thus spoke Algorizmi'). That's how Al-Kharizmi's name was transformed into the notion of algorithm.

Early on, algorithms had to do with the solution of specific problems, not abstract generalities – for instance, calculating the length of the lunar month or the square and cubic roots of integers. All such calculations were carried out by people, of course, using pencil and paper (or something equivalent). Of special significance here is that the head and the hand were both involved in doing the calculations. In the industrial era, too, almost all the way to the mid-twentieth century, algorithms were applied by astronomers, insurance companies, census bureaus and weapons projects on an industrial scale, and this was done by humans and machines working in tandem.

It was only in the second half of the twentieth century when algorithms became 'automated' and understood as similar to recipes and procedures that can be followed through failsafe computation. That is how Donald Knuth, who wrote the bible of computer science, described them except that he added five features for computer algorithms: 'finiteness, definiteness, input, output, effectiveness' (Knuth 1973). Without getting into the details of these terms, essentially an algorithm is understood as an effective method that can be expressed and executed in a finite time and space. By adding these requirements, Knuth highlighted that the medium of computation matters. For, in principle, the medium on which an algorithm is implemented should not be relevant – one can carry out the same algorithm on paper, on a mechanical device or a digital computer, even with a set of wooden tokens – except that, for most practical and interesting purposes, the finite-time-and-space constraint would exclude all but the digital alternative (and any other that would surpass their speed and efficiency – for

example, quantum computing).[1] In short, only algorithms that can produce a result in finite time and space are practically useful.

This unique advantage of the digital medium enabled the automation of algorithms in ways that were inconceivable earlier, launching a new era that has brought us to the current moment of algorithmic decision-making which seeks to replace not only human judgment in the sense of discretion but also the laws and regulations that govern our social relations. 'By driving the exercise of discretion underground', Daston argues, 'rules-as-algorithms blow up the bridges that connected universals to particulars in rules-as-models' (2022: 21). The question facing us is what bridges are blown up now that rules-as-algorithms are driving underground in addition the exercise of social norms and laws?

To address this question, we need to step back and take a closer look at the development of algorithms in recent decades.

3.1 Expert systems

The automation of algorithms in early computing is best exemplified in the development of 'expert systems' in AI. By the late 1960s and early 1970s, a decline of interest in AI research on the part of funding agencies had led practitioners to look for practical problems to solve.

As knowledge was conceived to be the key to such endeavour, a new class of artefacts called expert systems then appeared on the scene. Mycin, for instance, was one of the first expert systems for the diagnosis of infectious blood diseases. It used 'production systems' – that is, a set of 'if then' rules, like the following, along with the mechanisms for deciding when and how to apply the rules separated from the set of rules themselves.

IF
the site of the culture is blood, AND the gram strain is positive, AND
the portal of entry is gastrointestinal tract, AND
the abdomen is the locus of infection, OR
the pelvis is the locus of infection

THEN
there is strongly suggestive evidence that Enterobacteriaceae is the
class of organisms which therapy should cover.

1. Quantum computing uses the principles of quantum mechanics to perform calculations. Specifically, it uses specialised hardware to manipulate 'qubits' — the equivalent of a 'bit' in digital computing. Like a bit, a qubit can be in one of two states but, unlike bits, qubits can also exist in superpositions of those states. Furthermore, they can also be entangled with other qubits. Roughly speaking, entanglement is the result of non-local correlations among the parts of a quantum system. What this means is that we cannot fully understand the state of the system by dividing it up into parts and studying the separate states of the parts. Information can be encoded in non-local correlations among the parts of the system. Much of the art of designing quantum algorithms involves finding ways to make efficient use of these non-local correlations. Superposition and entanglement are the sources of power of quantum computing because, unlike the classical case, a qubit is equivalent to a vector in a two-dimensional space of real numbers (Ekbia 2008: 70).

This kind of linear reasoning borrows strong elements from logic, which relies on the explicit expression of facts and a formal modelling of the rules. The origins of this approach go back to rationalism and developments in mathematical logic in the last three hundred years (Ekbia 2008). Systems such as Mycin represent what is commonly referred to as 'symbolic AI'; symbolic because they use the formalism of symbolic logic (predicate calculus) to represent the world as a set of objects, predicates (attributes, properties) and relations – that is, as a set of symbols. The first proposition (sentence) in the above chain of reasoning, for instance, will be represented in the following way:

Box 1 The use of symbols in logic

Let P(x) be 'x is the site of the culture' and Q(x) be 'x is blood'. Then the proposition can be represented as: $\exists x\ (P(x) \land Q(x))$

This can be read as 'there exists an x such that x is the site of the culture and x is blood.'

Similarly, to represent the proposition 'the gram strain is positive' in predicate calculus, we can use a predicate symbol and a variable to represent the subject of the proposition. Using the predicate symbol 'P(x)' to represent 'x has a positive gram strain' gives 'x' as a variable that can take on different values depending on the context.

Therefore, the proposition 'the gram strain is positive' can be represented in predicate calculus as: P(x)

where 'x' refers to the subject being discussed such as a bacterium, a sample of biological material or a patient.

The whole production rule (chain of reasoning), for instance, might be represented as in Box 2.

Box 2 Symbolic representations of early AI

$(P \land Q \land R \land (S \lor T)) \rightarrow U$

where:*

P: the site of the culture is blood
Q: the gram strain is positive
R: the portal of entry is the gastrointestinal tract
S: the abdomen is the locus of infection
T: the pelvis is the locus of infection
U: therapy should cover Enterobacteriaceae
$\land$: AND
$\lor$: OR
$\rightarrow$: implication

* The symbol $\rightarrow$ in logic is called 'implication' or 'conditional'. It is used to denote a logical relationship between two statements, where the first statement (the antecedent) implies the truth of the second statement (the consequent). The implication symbol is often read as 'if then' or 'implies'.
Source: author's own elaboration.

Expert systems with such attributes – deep and linear inferences drawn from a small amount of information captured in a few variables – had some early success, bestowing on AI the respect it was longing for not only in academia but also in business, where billions of dollars were invested in expert systems for manufacturing, financial services, machinery diagnosis and so on. But this success was limited because the competence of such systems was restricted to very narrow domains. Two of the most notorious examples of such limitations come from a medical diagnosis program that, given the data for the case of a 1969 Chevrolet having reddish-brown spots on its body diagnosed it as suffering from measles; and a car loan authorisation program that approved a loan to a teenager who had claimed to have worked at the same job for twenty years (Ekbia 2008: 96-97). The problem, of course, was that the builders of the system had failed to include certain facts in the knowledge base: that the given Chevrolet is not a human being; and that one cannot have work experience that is longer than one's age. Another program, given the case of a patient with a kidney infection, prescribed boiling the kidney in water – a good remedy against infection, but a terrible failure to understand the basics of what was going on (Ekbia 2008: 96-97).

To state the obvious, these programs seemed to lack an important feature of intelligence: despite apparent sophistication and expertise in specialised areas like medicine, they demonstrated a clear lack of understanding of very basic facts that a human being takes for granted. They either had to be given the minutest details, as in the first two cases, or else they seemed to be missing fundamental knowledge (e.g. about living organisms), as in the third.

3.2 Algorithms in machine learning

AI researchers and practitioners tried to overcome these issues in various ways but, by and large, they failed until big data and machine learning came to their 'rescue', although with limitations of their own. Machine learning algorithms build on earlier techniques in AI and computing, such as neural networks, while breaking away from many aspects of symbolic AI.

Smith (2019: 47) describes machine learning as a suite of statistical techniques for the statistical classification and prediction of patterns, based on large quantities of data, using an interconnected fabric of processors that are arranged in multiple layers. What makes machine learning distinct from earlier techniques, such as earlier AIs, Smith adds, is its reliance on (a) shallow (few step) inference; (b) by a massively parallel process; using (c) massive amounts of information; and (d) involving a very large number of (e) weakly correlated variables.

In some rough sense, like their neural networks predecessors, machine learning systems are brain-like structures made up of heavily interconnected nodes ('neurons') that run in parallel, finding patterns in the vast amount of data that is fed into them from various sources (sensors, cellphones, social media, location tracking devices, etc.). Broadly speaking, such patterns are statistical correlations that are implicitly captured in connections between the nodes. Because they are implicit, it is hard, if not impossible,

for a human observer to recognise and understand what the correlational patterns mean. Unlike the explicit propositions of symbolic AI, which were legible to a human with a basic knowledge of symbolic logic, these patterns are encoded in a 'language' that is not accessible to human beings. This, along with the distributed (multiprocessor) character of current AI systems, is the source of the opacity of machine learning algorithms, making it hard for human beings to understand how they arrive at decisions. That is why you'd be doomed if your social benefits, as a resident of Washington DC, are cut off on the basis of the patterns that a machine learning system has identified in your data. Neither you as the beneficiary, nor the government workers who use them, have access to the 'logic' behind the system's decision.

4. Varieties of Judgment

This shift in techniques from symbolic AI to machine learning, from algorithms that reasoned explicitly to black-boxed systems that we don't understand, has further complicated our relationships with AI systems, putting more power in the hands of those who design and market them and chipping away from the rights and recourses that were available to the rest of us. To see how, we need to examine the principles on which such systems are built, three of which – about social reality, about human behaviour and about our relationship with modern technology – have particular resonance.

The first principle can be described as the principle of regularity. As discussed earlier, modern times have witnessed a gradual transition from rules-as-models to rules-as-algorithms, or from thick rules to thin rules. Computer algorithms are the thinnest of rules not because they are short or simple – which they are obviously not – but because they are built on the assumption of the regularity of the outside world and, more specifically, the uniformity of the social world. This uniformity manifests itself in various ways: in the way similar algorithms are applied to various domains of life, including online trading and online dating, traffic regulation and job, loan or university entrance applications; in the way people are pigeonholed into categories based on criteria deemed relevant to powerful players such as large corporations (and sometimes governments); and in the way statistical generalisations paper over meaningful differences among personal, cultural and economic backgrounds, assuming homogeneity within each category of people. Old statistical techniques failed to capture nuance and difference because they were coarse-grained. Machine learning, equipped with the immense computational power of current technology, has refined those techniques but it has not overcome these basic problems. When it comes to human differences, fine-grained statistics fare no better than their coarse-grained predecessors.

The second principle is closely related to the first, having to do with the malleability of the human condition and of human behaviour. The origins of this go back to the behavioural psychology of the early twentieth century and its emphasis on conditioning. The cognitive revolution of the mid-century, of which AI is an intellectual heir, dislodged the behaviourist emphasis on the observability of the conditioning mechanism, pushing it to the 'subconscious' and the inner workings of the mind. But it maintained the earlier belief in the explainable character of human behaviour from an intentional standpoint

– that is, in terms of beliefs, desires and intentions. In and of itself, that belief does not incur harm to the social fabric. The issue is how current systems enabled by AI techniques seek to mould human behaviour in the image forged by corporations. Daston (2022: 5) captures this issue succinctly:

> An island of stability and predictability in a tumultuous world, no matter what the epoch or locale, is the arduous and always fragile achievement of political will, technological infrastructure, and internalized norms.

Our epoch, it seems, is increasingly resorting to technology to create not islands but continents of stability and predictability, and not on a local but on a global scale. It is not hard to see the fragility of these arrangements in a growingly tumultuous world.

Third, and perhaps most relevant to the present discussion, is the binary principle of comparing humans and machines, often in terms of the 'competitive advantage' of machines over human beings. This false binarism is behind many of the claims about the alleged superiority of AI systems in regard to accuracy, reliability, objectivity, etc. In reality, however, these systems are ultimately the product of the embedding social, economic and cultural environment in which they are developed, designed and deployed. Separating them from this environment under the rubric of 'autonomy' is as misguided as it is to think of human beings in isolation from the social environment. That environment includes, among other things, the very technologies that human societies have created throughout their history and continue to create at this particular juncture. From this perspective, there is no 'pure' human detached from technology and no technology detached from humanity. By the same token, to speak of putting humans 'back in the loop' sounds somewhat like a tautology because humans already are, and will continue to stay, in the loop in the foreseeable future. It is ultimately technologically-enabled humans who make decisions, who act on those decisions and who bear their consequences. Someday 'smart' machines may be able to go their own way, built on thick rules that can emerge from their deep embeddedness in a social environment. Until then, human judgment remains our last resort, as flawed and fallible as it might often be. In humans we should trust!

References

Anderson J.G. and Abrahamson K. (2017) Your health care may kill you: Medical errors, Studies in Health Technology and Informatics, 234, 13–17.

CDC (2023) Road traffic injuries and deaths – A global problem, Center for Disease Control and Prevention. https://www.cdc.gov/injury/features/global-road-safety/index.html#print

Chirimuuta M. (2023) Rules, judgment, and mechanization, Journal of Cross-Disciplinary Research in Computational Law. https://journalcrcl.org/crcl/article/download/22/12

Danziger S., Levay J. and Avnaim-Pesso L. (2011) Extraneous factors in judicial decisions, Proceedings of the National Academy of Sciences, 108 (17), 6889–6892. https://doi.org/10.1073/pnas.1018033108

Daston D. (2022) Rules: A short history of what we live by, Princeton University Press.

Ekbia H.R. (2008) Artificial dreams: The quest for non-biological intelligence, Cambridge University Press.
Guyer P. and Wood A.W. (eds.) (1998) Immanuel Kant, Critique of pure reason, Cambridge University Press.
Johnson K. (2022) Algorithms quietly run the city of DC — and maybe your hometown, Wired, 3 November 2022. https://www.wired.com/story/algorithms-quietly-run-the-city-of-dc-and-maybe-your-hometown/
Knuth D.E. (1973) The art of computer programming, Volume 1: Fundamental algorithms, 2nd ed., Addison-Wesley.
Kuhn T. (1962) The structure of scientific revolutions, University of Chicago Press.
Smith B.C. (2019) The promise of artificial intelligence: Reckoning and judgment, MIT Press.
WHO (2019) WHO calls for urgent action to reduce patient harm in healthcare, News, 13 September 2019. https://www.who.int/news/item/13-09-2019-who-calls-for-urgent-action-to-reduce-patient-harm-in-healthcare
Wittgenstein L. (1953) Philosophical investigations, Blackwell.

All links were checked on 23.01.2024.

Cite this chapter: Ekbia H.R. (2024) In Humans We Trust: rules, algorithms and judgment, in Ponce del Castillo A. (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 4
In AI We Trust: power, illusion and the control of predictive algorithms

Helga Nowotny

1. Introduction: what is AI? The societal context of digital technologies

Intense and accelerating involvement with digital technologies is having a profound impact on the ways in which we work and live, transforming our societies and economies. It challenges us to invent novel ways to use digital technologies to advance the common good instead of mainly increasing the concentration of economic power in the hands of the few. This entails unlocking the great potential of digital technologies to meet the impending risks of climate change, the next pandemics and other emergencies over the horizon. In the workplace, digital technologies continue the processes of automation that began a long time ago, complementing and replacing an ever larger range of tasks, skills and professional activities. Hence, the challenge before us is how to guarantee the development and deployment of a technology – epitomised by artificial intelligence – in a way that makes it sufficiently responsive to the human needs and rights that are being redefined in the process. This includes allocating accountability and responsibility within a complex association between the users of digital technologies and the designers, producers, owners and regulators of the latter.

As have other technologies previously, digital technologies raise questions about whether machines will eventually control, dominate or even fuse with humans, raising fears about dehumanisation, surveillance and totalitarian control. On one side, techno-utopians hail the disruption as bringing the solution to all problems, a 'liberating' force that would even lead to 'greater world harmony' (Negroponte 1998). On the other side are pessimists with their dystopian visions, heralding the end of humanity. Before entering this discussion, we should remind ourselves of the more nuanced relationship between technologies and the humans that design and use them. As David Nye, the technological historian, observed some time ago 'artefacts emerge as the expression of social forces, personal needs, technical limits, markets and political considerations' (Nye 2006). The digital gadgets, infrastructures, networks and machines that serve us now, and that we serve through our behaviour, are not predetermined once and for all. Rather, it is up to us to appropriate, modify and shape them through the choices we make, albeit within societal and technological constraints.

These are some glimpses of the societal context of digital technologies that must be kept in mind when answering the question: what is AI? Today, we have arrived at a crucial point in a long, unprecedented, evolutionary journey marked by the entanglement of multiple interactions between humans and digital machines. Its beginning dates back to the 1940s, but it was only around the first decade of the twenty-first century that a

convergence of three different strands unleashed the power of artificial intelligence that we are witnessing today: the enormous increase of computational power that enables sensors and computer chips of miniature size to be installed in almost every device and everywhere; the development of ever more sophisticated algorithms; and, last but not least, access to and the increasing availability of an enormous amount of data coming from many fields of application.

Therefore, the definition of AI varies according to where we find ourselves in this trajectory. In the beginning, a naïve definition predominated, built on the mathematical-formal approach that started with Alan Turing's insights into the possibilities of developing a mathematical code to run a machine (Turing 1936). This opened the gate to a world in which, as per the definition of the Turing test, AI would simulate human intelligence in machines programmed to think like humans and mimic their actions, whereby 'thinking' was largely equated with formal reasoning. Mathematical code needed hardware to operate electronically in a computing machine and, spurred by the war effort in the 1940s, the technological progress of computers and their performance quickly advanced. The term 'artificial intelligence' was coined at the Darthmouth Conference in 1956 and, even if it is not the most fortunate term, it is still with us. It invites different meanings and interpretations of what 'intelligence' is, especially when juxtaposing human intelligence with the very different 'intelligence' of a machine.

This ambitious but naive definition became replaced by a more realistic one as the formal logical approach failed to yield many of the hoped for practical applications. A decline of funding set in, followed by a period remembered as the 'AI winter'. The decisive turn came at the beginning of the twenty-first century when neural networks began to be deployed, capable of discovering patterns and statistical correlations in data with astonishing efficiency and accuracy. This led to the rise of procedures termed 'machine learning' and 'deep learning'. Algorithms are trained for pattern recognition with the help of large amounts of data capable of training themselves in an approach referred to as 'unsupervised learning'. Hence, AI is defined as any agent or system that perceives its environment and takes actions that maximise its chance of achieving its goals. This implies that a machine must be able reliably to recognise patterns in the environment in which it is expected to act, as with automated vehicles. The goal needs to be defined in precise ways, for instance when following and exploring the rules of games like chess or Go which have resulted in spectacular demonstrations of AI defeating the world's best players.

A third speculative definition of AI has pervaded the public discourse as part of the attempt to realise artificial general intelligence. This entails the still hypothetical ability of an intelligent agent to understand and learn any task that a human being can do. In this definition, an AI would be able to achieve a kind of superintelligence through recursive self-improvement, leading to a point called 'the singularity' by inventor Ray Kurzweil in which the AI will overtake human intelligence. Not surprisingly, this would fundamentally change what it means to be human. While some techno-utopians celebrate this as the ultimate feat of overcoming our humanity by reaching transhumanism, for others this poses an existential risk and the end of humanity as we know it.

2. The power of predictive algorithms: where it comes from and how it affects us

Wanting to know what the future holds has been an ardent wish in all civilisations we know, resulting in divination practices to be found everywhere. Ancient Chinese oracle bones show cracks on the shoulder blades of sheep or turtles that had been held over fire by divinatory experts in order to 'read' the future. Today, we resort to foresight reports and analysis of future trends. While the tools have changed, we are as keen as our ancestors to learn what to expect in the decades ahead and rely increasingly on predictive algorithms. They allow us to build simulation models that answer the question 'what if?' and to expand our imagination while engaging in future-making.

In this process, financial markets for example underwent an intense phase of computerisation and the rapid evolution of automated computer algorithms. Developed by humans, the actual decisions to buy and sell are made by the algorithms and executed through a comprehensive digital infrastructure that links individual trading firms to the various exchanges on which they trade (MacKenzie 2021). The accuracy of weather predictions has also increased dramatically, enabling the worldwide transportation and mobility networks we rely upon today. The trajectory of a hurricane can be followed in real time and its landfall predicted, providing more time for evacuations. Another rapidly expanding field with more dire consequences are automated weapons systems, including drones deployed for military purposes but which also have commercial applications.

Predictive algorithms work not only for governments, the military and business, but for all of us. We rely on them as individuals when wanting to know our future state of health and the risks carried through our genes or lifestyle. We use predictive algorithms for everyday decisions that facilitate our work and personal choices. We collude with the large digital corporations when we feed them our personal data, transforming ourselves into 'the product' that is then sold by them to advertisers who target us in return for the convenience of receiving their services. As Shoshana Zuboff has shown in impressive detail, we have become part of the surveillance capitalism that thrives on the widespread use of predictive algorithms (Zuboff 2019).

Decision-making based on predictive algorithms rapidly pervades not only the business world but public institutions like the police, judiciary, education and the health system. The line to tread between a desirable increase in efficiency and threats to privacy is thin and needs to be continuously re-negotiated within a firm regulatory framework. An example of the trade-offs often involved comes from Arbeitsmarktservice (AMS; the Austrian Public Employment Service) which decided to install an algorithm dividing employment seekers into three groups according to the 'objective' criteria of their profile's prediction of their chances of finding employment. This was followed by a public outcry as the establishment of the category of the least employable was deemed to be socially unjust. Despite assurances from the AMS that this group would also have their needs looked after, the algorithm had to be withdrawn. Although the criteria of 'transparency' had been fully met, the demand for social justice prevails, at least for now.

There are more risks that come with predictive algorithms. As I describe in my book *In AI We Trust. Power, Illusion and Control of Predictive Algorithms*, by transferring ever more agency to an algorithm we tend to believe what it predicts and forget that the algorithm is based on an extrapolation of data from the past. All predictions are based on probabilities as the future remains inherently uncertain. When human behaviour follows a belief in the predictions, self-fulfilling prophecies result and this may herald a return to a deterministic worldview. At the heart of our trust in AI lies a paradox: we leverage control over the future and uncertainty while, at the same time, the performativity of AI and the power it has to make us act in the ways it predicts reduce our agency over that same future (Nowotny 2021).

3. Keeping humans in the loop: towards a digital humanism

The digital devices that surround us and with which we continuously interact provide us with feedback and answers to our questions but also nudge us in pre-set directions. A myriad of sensors and digital tools are automating the ways in which business is conducted, cutting costs and increasing efficiency. While the automation of work is not new, nobody knows how fast new jobs will be created to replace the ones that are vanishing. There are benefits to be gained, but they are unequally distributed. Other serious downsides loom as AI reinforces existing power structures and their concentration into monopolies and oligopolies. Biases in society are transferred to the data and to the algorithms on which decisions rely. These may cause harm and support discriminatory practices, with injustice becoming ingrained, in the absence of institutionalised mechanisms to appeal to human judgment. We are currently witnessing the rampant abuse and misuse of AI in spreading 'fake news' and the threat to liberal democracies.

The unanimous response has been the call to develop an ethical or beneficial AI, expected to fulfil a series of criteria like transparency, explainability, responsibility, fairness and more. However, the problems in implementing these legitimate demands are considerable. First, no consensus exists on the ethical principles themselves, as shown in a study of the ethical guidelines issued by governments and leading corporations worldwide (Jobin et al. 2019). Second, attempts to insert ethical procedures into the design of algorithms, like making self-driving cars 'safe', encounter technical problems as no single interface exists to make an algorithm 'see' like a human driver. Third, the focus on the behaviour of entire ethical systems by auditing the outcome, for example whether human rights are being respected, requires implausible specificity (Danks 2022). Ethics, in line with such a conclusion, is necessary but not sufficient. It is not a checklist and its implementation remains open.

What could work instead? Undoubtedly, more regulation is necessary although it is difficult to achieve at international level with Europe poised between the US and China. A movement devoted to digital humanism is attempting to integrate a human-centred approach in the design, production and deployment of AI throughout their systemic interlinkages (Werthner et al. 2022). It seeks to identify specific points of intervention and to be attentive to actual practice in various domains, as well as becoming part of

the education system. The values on which digital humanism is based will be crucial for shaping the future of work and of liberal democratic societies.

The co-evolutionary journey between humans and digital machines has only begun. AI has considerably expanded human capabilities and opened new spaces of knowledge. It is a powerful technology created by humans and therefore it is social. We can use it to build the society we wish to live in by designing and using AI for the common good.

References

Danks D. (2022) AI ethics as translational ethics. https://www.youtube.com/watch?v=UBo5wVs2qgs

Jobin A., Ienca M. and Vayena E. (2019) The global landscape of AI ethics guidelines, Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

MacKenzie D. (2021) Trading at the speed of light: How ultrafast algorithms are transforming financial markets, Princeton University Press.

Negroponte N. (1998) One-room rural schools, Wired, 6 (9). https://web.media.mit.edu/~nicholas/Wired/WIRED6-09.html

Nowotny H. (2021) In AI we trust: Power, illusion and control of predictive algorithms, Polity Press.

Nye D.E. (2006) Technology matters: Questions to live with, MIT Press.

Turing A. (1936) On computable numbers, with an application to the Entscheidungsproblem, Proceedings of the London Mathematical Society, S2-42 (1), 230-265. https://doi.org/10.1112/plms/s2-42.1.230

TU Wien (2019) Vienna manifesto on digital humanism. https://dighum.ec.tuwien.ac.at/dighum-manifesto/

Werthner H., Prem E., Lee E.A. and Ghezzi C. (eds.) (2022) Perspectives on digital humanism, Springer.

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

All links were checked on 23.01.2024.

Cite this chapter: Nowotny H. (2024) In AI We Trust: power, illusion and the control of predictive algorithms, in Ponce del Castillo A. (ed.) Artificial intelligence, labour and society, ETUI.

Part 2

Global and environmental perspectives

Chapter 5
The politics of purpose: AI for a global race or societal challenges?

Inga Ulnicane

1. Introduction

Artificial intelligence (AI) today is at the centre of intense policymaking activity. Since 2016, national governments and international organisations have adopted AI strategies and action plans, launched funding and training programmes, established agencies, expert groups and consultations, and are developing regulatory frameworks (OECD 2021). Among these multiple activities, it is of particular importance not to forget the big picture and the key questions – what is the purpose of AI development and deployment? Why is taxpayers' money being invested in it? Why are elected officials supporting its development and use? Why is so much time and effort spent on discussing the details of future AI regulation? What is the overarching goal to which all these activities aim to contribute? Whose interests, values and norms does this overarching goal represent?

To address these questions, this chapter discusses how recent AI policy frames its purpose according to two well-known goals of technology policy; namely, economic competitiveness and societal challenges (Schiff 2023; Ulnicane 2022). It invites a reflection on whether the way these goals are formulated and presented in AI policy is helpful and representative of broader societal interests. First, the chapter introduces two stylised approaches to technology policy, focusing on economic competitiveness and societal challenges. Afterwards, it examines how these approaches show up in recent AI policy.

2. Economic competitiveness and societal challenges: the two main goals of technology policy

While traditionally the main purpose of technology policy has been to support economic competitiveness, recently it has also been recognised that an important goal is to tackle societal challenges.

Concerns about national economic competitiveness have for a long time been an important driver of national investments in technology. According to this thinking, technology is always good and therefore we need as much technology as possible and as fast as possible. National governments compare their technological development to that of other countries, worrying about falling behind. In the second half of the twentieth century, France was worried about the 'American challenge' and the European Community was concerned about the technology gap with the US, while the US and the UK were, in turn, worried about emerging Japanese technological superiority (Ulnicane

2022). One of the best-known historical examples of international technological rivalry is the 'space race' between the then superpowers, the US and the Soviet Union, competing for supremacy in conquering the Moon.

This discourse of international technological rivalry remains very popular in policymaking. For example, when in the early twenty-first century the EU launched the Lisbon Strategy and the European Research Area, it was largely motivated by concerns that the EU was lagging behind the US and Japan. Accordingly, the ambition for the EU was to become the most competitive knowledge-based economy in the world. Despite the limited successes of such initiatives, this thinking of technology development as a global race remains very popular.

The economic competitiveness discourse can help mobilise resources and draw attention to the importance of technology development. It can be used by interest groups trying to lobby for more funding or favourable policies for certain technologies. However, it has also been criticised as a 'dangerous obsession' that portrays international technological development as a zero-sum game in which one country wins but others lose (Krugman 1994). Moreover, this discourse might prioritise prestige technology projects over more urgent social needs.

In recent years, it has been recognised that an important goal of technology development and use is to contribute to tackling societal challenges in areas such as health, environment and energy (Diercks et al. 2019; Ulnicane 2016). This approach sees technology as contributing to the achievement of the United Nations Sustainable Development Goals, for example, 'no poverty', 'zero hunger' or 'gender equality'. In order to tackle these complex and uncertain societal challenges, broad-ranging collaborations are needed that involve representatives from diverse disciplines and sectors including science, civil society, government and the private sector. As these societal challenges can involve cross-border issues, international collaboration might be needed to address them. Thus, technology development internationally here is seen as a positive-sum game in which many can benefit.

While the discourse on the role of technology in tackling the major societal challenges of our times through boundary-spanning collaborations addresses important social issues, its realisation is far from straightforward. Development of technology is highly uncertain, the possibility of steering it should not be exaggerated, the societal challenges are complex and their solution cannot be guaranteed. Moreover, while involving diverse stakeholders and voices is of great importance, in practice it remains challenging to balance the existing power asymmetries, as the most resourceful and better organized interest groups tend to dominate. Moreover, this approach recognises that the social effects of technology are not always good and that it can also create harmful effects, for example, for the environment or health.

The main characteristics of these two goals of technology policy – namely, economic competitiveness and societal challenges – are summarised in Table 1 below. These two approaches often coexist. Next, this chapter examines these two approaches in recent AI policy.

Table 1 **Stylised technology policy frames**

	Economic competitiveness	Societal challenges
Purpose of technology development	To support national economic competitiveness	To tackle societal challenges and UN SDGs
Global technology development	Zero-sum game	Positive-sum game
Impact of technology	Always good	Can be good and bad

Source: author's own elaboration.

3. The purpose of AI development and use

In recent policy and media debates about AI, we can find discussions of both goals – boosting economic competitiveness as well as helping to tackle societal challenges.

3.1 Economic competitiveness, global race and leadership in AI

Policy often represents AI as a new basis for economic growth and a major opportunity for boosting productivity, efficiency and cost savings. AI development is depicted as taking place within fierce global competition. Sometimes, it is compared to a new 'space race' or cold war, highlighting rivalry between the two superpowers – the US and China, representing two different political and economic systems. Many countries from China and the US to Finland, Germany, South Korea and Singapore have declared their ambitions to be leaders and frontrunners in AI (OECD 2021; Ulnicane et al. 2022). Often, repeated statements about global AI leadership, such as the one by Russian President claiming that 'whoever leads in AI will rule the world', have strong neo-imperialist undertone.

Global leadership in AI is seen as crucial not only for the national economy, security and society but also as a way of promoting national values globally. This can be seen in the Executive Order of the US President 'Maintaining American Leadership in Artificial Intelligence', which opens with the following statement:

> Artificial Intelligence (AI) promises to drive growth of the United States economy, enhance our economic and national security, and improve our quality of life. The United States is the world leader in AI research and development (R&D) and deployment. Continued American leadership in AI is of paramount importance to maintaining the economic and national security of the United States and to shaping the global evolution of AI in a manner consistent with our Nation's values, policies, and priorities (Executive Order 13859).

The EU policy on AI sends mixed messages about its interest in being a global leader, from recognising that there is fierce global competition going on and that the EU is lagging behind, to statements that it wants to be a leader in its own way and based on its values or that it is not interested in winning or losing the race but pursuing its human-

centric approach (Ulnicane et al. 2022). In some EU documents, we see ambitions which are quite similar to those of the US, of being a leader and promoting its values globally. For example, in its 2021 communication 'Fostering a European approach to Artificial Intelligence', the European Commission highlights the EU's efforts to be a global leader in the promotion of Trustworthy AI and states that European coordination of AI investments and policies:

> … will enable the latest technologies to be developed and adopted through Europe's global competitiveness and leadership. Such coordination will allow Europe to seize benefits of AI for the economy, society and the environment and help to promote European values worldwide. (European Commission 2021: 8)

To establish or maintain their global leadership, countries have set out a range of national and international measures, including protectionist approaches to ensure their advantage in AI technologies. This again can be seen in the Executive Order of the US President, which states that:

> The United States must promote an international environment that supports American AI research and innovation and opens markets for American AI industries, while protecting our technological advantage in AI and protecting our critical AI technologies from acquisition by strategic competitors and adversarial nations.

Hand-in-hand with the global race and leadership discourse goes a fear of lagging behind and missing out on the opportunities offered by AI. This creates a sense of urgency to make an effort and do something so as not to be left in an inferior position. This discourse of a global AI race, with winners and losers, is reinforced by various rankings and indices that compare countries' performance according to a range of indicators. Often, it is also uncritically or strategically promoted by experts and stakeholders.

Thus, the popular discourse of an AI global race and leadership is a new version of the traditional approach to technology policy that presents technology as an important contributor to national economic competitiveness. As mentioned above, it has received some criticism, for example for depicting global technology development as a zero-sum game and drawing resources and attention to prestige projects instead of broader social needs. However, in today's AI policy this approach is largely unchallenged except some reservations which are expressed in EU policy. However, there is an urgent need to challenge this discourse and ask some critical questions: Is national global leadership in AI really worth the investment and effort? Is it what society needs? Is it drawing away attention and resources from more important issues of broader relevance? Whose interests does it serve? Is it a convenient tool for vested interests to use strategically, for example to argue for less regulation in the EU so that it can be more competitive globally?

3.2 Societal challenges and Sustainable Development Goals

AI policy includes highly optimistic statements about the potential social benefits of AI, outlining positive expectations towards its contribution to addressing social, environmental and health issues. According to the European Commission:

> The potential benefits of AI for our societies are manifold, from less pollution to fewer traffic deaths, from improved medical care and enhanced opportunities for persons with disabilities and older persons to better education and more ways to engage citizens in democratic processes, from swifter adjudications to a more effective fight against terrorism and crime, online and offline, as well as enhancing cybersecurity (European Commission 2021: 1).

Similarly, in *Ethics Guidelines for Trustworthy AI* we find a very hopeful approach to the potential of AI to contribute to achieving the UN's Sustainable Development Goals:

> AI systems can help to facilitate the achievement of the UN's Sustainable Development Goals, such as promoting gender balance and tackling climate change, rationalising our use of natural resources, enhancing our health, mobility and production processes, and supporting how we monitor progress against sustainability and social cohesion indicators (European Commission 2019: 4).

To realise these positive social benefits of AI, the policy documents call for inclusive and participatory governance involving a wide range of stakeholders including from marginalised and disadvantaged groups.

While the 'societal challenges' discourse offers a very hopeful and positive view on the potential of AI, it also needs to be critically examined and its shortcomings highlighted. It tends to present a rather one-sided picture of AI. For example, it emphasises the potential of AI to promote gender balance whereas in many cases AI has reinforced gender and racial bias. Similarly, potential of AI to tackle climate change is mentioned but high environmental costs of AI are ignored. This discourse presents AI as a simple technological fix (Johnston 2018) to complex and uncertain societal issues. It might over-promise the social benefits of AI, which could lead to a backlash. Moreover, previous experience demonstrates that multi-stakeholder forums for AI have been captured by the better organised and funded vested interests.

4. Concluding remarks

Although AI is presented as a novel technology, its recent policy draws on very traditional and well-known technology policy ideas that see the boosting of economic competitiveness as the main goal of technology development. It is often seen as a global race and as a zero-sum game. In addition, in AI policy we also find a more recent discourse highlighting the potential of technology to contribute to tackling societal challenges. In AI policy both discourses co-exist and it is often expected that technology can help to achieve both goals of economic competitiveness and societal challenges. A

reflection on the compatibility of both goals and the potential trade-offs between them is missing.

This contribution points out the problematic aspects of both of these goals. Chasing global leadership in AI might happen at the expense of much needed global collaboration or the tackling of broader societal issues. However, optimistic statements about the potential of AI to help tackling societal challenges often come across as quick technological fixes in response to societal problems that are complex and uncertain. This contribution is an invitation to think critically about these goals, how they are presented and promoted, and whose interests do they serve. It is also a call to think about potential alternatives and better ways to articulate what we expect from AI and other technologies.

References

Diercks G., Larsen H. and Steward F. (2019) Transformative innovation policy: Addressing variety in an emerging policy paradigm, Research Policy, 48 (4), 880–894. https://doi.org/10.1016/j.respol.2018.10.028

European Commission (2019) Ethics guidelines for trustworthy AI, High-level expert group on artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission (2021) Fostering a European approach to artificial intelligence, COM(2021) 205 final, 21.4.2021. https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence

Executive Office of the President (2019) Maintaining American leadership in artificial intelligence, Federal Register, 84 (31), 3967-3972. https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence

Johnston S.F. (2018) The technological fix as social cure-all: Origins and implications, IEEE Technology and Society Magazine, 37 (1), 47–54. https://doi.org/10.1109/MTS.2018.2795118

Krugman P. (1994) Competitiveness: A dangerous obsession, Foreign Affairs, 73 (2), 28–44. https://doi.org/10.2307/20045917

OECD (2021) State of implementation of the OECD AI principles: Insights from national AI policies, OECD Digital Economy Papers 311, OECD Publishing. https://doi.org/10.1787/1cd40c44-en

Schiff D. (2023) Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI policy, Review of Policy Research, 40 (5), 729–756. https://doi.org/10.1111/ropr.12535

Ulnicane I. (2016) 'Grand challenges' concept: A return of the 'big ideas' in science, technology and innovation policy?, International Journal of Foresight and Innovation Policy, 11 (1-3), 5–21. https://doi.org/10.1504/IJFIP.2016.078378

Ulnicane I. (2022) Emerging technology for economic competitiveness or societal challenges? Framing purpose in artificial intelligence policy, Global Public Policy and Governance, 2 (3), 326–345. https://doi.org/10.1007/s43508-022-00049-8

Ulnicane I., Knight W., Leach T., Stahl B.C. and Wanjiku W.-G. (2022) Governance of artificial intelligence: Emerging international trends and policy frames, in Tinnirello M. (ed.) The global politics of artificial intelligence, CRC Press, 29–55. https://doi.org/10.1201/9780429446726-2

All links were checked on 26.01.2024.

Cite this chapter: Ulnicane I. (2024) The politics of purpose: AI for a global race or societal challenges?, in Ponce del Castillo A. (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 6
An Eco-political economy of AI: environmental harms and what to do about them

Benedetta Brevini

1. Introduction

As the media continues to heighten awareness of the growing popularity of generative AI models, major 'Digital Lords' (Brevini 2020b) like Microsoft, OpenAI and Google have acknowledged that meeting the increasing demand for their AI tools comes at a substantial cost, including expensive semiconductors, massive energy use and an unprecedented impact on water consumption (George et al. 2023). In its most recent environmental report, Microsoft (2022) revealed a significant increase of 34% in its worldwide water consumption between 2021 and 2022, amounting to nearly 1.7 billion gallons. This uptake is closely linked to the company's AI research efforts and marks an increased liability compared with previous years.

As the obsession with AI uptake continues, COP 27, which took place in November 2022 in Egypt, reiterated once again that the planet is 'sending a distress signal' (UN 2022). The UN's *State of the Global Climate Report* for 2022 painted a 'chronicle of climate chaos', concluding that the previous eight years were on track to be the warmest on record (WMO 2022). The scientists writing the report estimated that global temperatures have now risen by 1.15 C since pre-industrial times, warning of the other wide-ranging impacts of climate change including the acceleration of sea level rise, unprecedented losses in glacier mass and record-breaking heatwaves. The Intergovernmental Panel on Climate Change (IPCC) has been sounding the alarm for years and it is now clear that, if we want to meet the Paris agreement target of keeping global warming below a 1.5 C threshold, we will need to cut emissions globally by 50% in the next decade (IPCC 2022).

While all this is unfolding, at the same time the world is attempting to recover from a global pandemic, the race towards green new deals has started: Europe is leading the way in developing strategies for a green recovery. Technological innovation and digital services are at the core of recovery with the potential to create millions of jobs and boost economies devastated by the pandemic. The European Commission proposed a major recovery plan for Europe on 26 May 2020, approved by the European Council on 21 July 2020. Alongside the recovery package, EU leaders agreed on a 1,074.3 billion euro long-term EU budget for 2021-27. Among other things, the budget will support investment in the digital and green transitions and in resilience.

The Communication by the European Commission entitled 'Strategic Foresight Report 2022 Twinning the green and digital transition in the new geopolitical context' (European Commission 2022) stressed once again the crucial role of the 'twin transition', green and

digital, both of which are at the top of the EU's political agenda. What is crucial about this Communication is that, for the first time, the European Commission is explicit that digital technologies will also bring additional environmental burdens with them. In particular, it explains that:

> Unless digital technologies are made more energy-efficient, their widespread use will increase energy consumption. Information and communications technology (ICT) are responsible for 5-9% of global electricity use and around 3% of greenhouse gas emissions. (…) However, studies show that ICT power consumption will continue to grow, driven by increasing use and production of consumer devices, demand from networks, data centres, and crypto assets (European Commission 2022).

It further acknowledges that 'further tensions will emerge in relation to electronic waste and environmental footprints of digital technologies' (European Commission 2022).

Despite growing attention on the environmental costs of ICT systems, artificial intelligence is principally heralded as the key technology to solve contemporary challenges, including the environmental crisis, with climate action being one of the UN's Sustainable Development Goals. Unfortunately, debates on green recovery plans and AI developments continue to avoid some crucial questions: how green is artificial intelligence? And how can we build AI applications that are truly sustainable?

This chapter address these questions by examining the set of environmental harms associated with AI technologies and offering solutions to this problem.

2. An Eco-political economy of AI: understanding AI's environmental costs

The book *Is AI Good for the Planet?* (Brevini 2021) addressed the question of the environmental harms of AI through an exploration of its extractive production and supply chain, thus unveiling the environmental costs of current data-driven communications systems and AI in particular.

In developing an Eco-political economy of AI, the book investigated artificial intelligence from resource, infrastructural and material points of view as 'a set of technologies, machines or infrastructures that demand and use huge amounts of energy to compute, analyse or categorise' (Brevini 2021: 94). Such a definition is key to changing our understanding of AI – which is more usually defined with a focus on its function and on its abilities to bring about desired radical change. Recent scholarship within communications studies, for example within human-machine communication, an emerging area of communications research, has defined AI as the study of the 'creation of meaning among humans and machines' (Guzman and Lewis 2019: 71). Furthermore, embracing the tradition of the critical political economy of communications allows us to view communications systems as assemblages of material devices and infrastructures (Brevini and Murdock 2017).

Is *AI Good for the Planet* argued that, if we want to develop an Eco-political economy of AI that helps us understand its environmental harms, it is imperative to initiate a new and comprehensive endeavour to define its parameters (Brevini 2021: 40). Here, the definition adopted by the white paper on artificial intelligence issued by the European Commission serves as a good starting point to regain an understanding of the materiality of AI, highlighting the connection between AI, data and algorithms: 'AI is a collection of technologies that combine data, algorithms and computing power. Advances in computing and the increasing availability of data are therefore key drivers of the current upsurge of AI' (European Commission 2020: 2).

3. From the 'sublime phase' of AI to its environmental costs

Technology has long been considered a fix-all solution to the inequalities of capitalism. As Vincent Mosco eloquently argued, 'one generation after another has renewed the belief that, whatever was said about earlier technologies, the latest one will fulfil a radical and revolutionary promise' (Mosco 2004: 117). Embedded in this neoliberal, techno-determinist discourse is a belief that digital technology can disrupt inequalities and power asymmetries, without the need to challenge the status quo. Following similar mythologies, the 'sublime phase' of AI offers its applications as solutions to the greatest challenges of the age: addressing chronic illness, repairing the economy, managing social services, anticipating cybersecurity threats and solving the climate crisis.

However, this portrayal of AI as the magic, sublime hand that will rescue society obfuscates the materiality of the infrastructures (Brevini 2020a, 2021) that are central to the environmental questions that have been so consistently, and so artfully, side-stepped (Brevini 2020a). Instead, we need to understand AI in its infrastructural context as depleting scarce resources throughout its production, consumption and disposal, increasing the amount of energy used and thus exacerbating the climate crisis. We need, instead, to develop an Eco-political economy of AI which entails studying its entire global supply chain in order to comprehend why it generates an array of environmental problems, most notably energy consumption and emissions, material toxicity and electronic waste.

4. An Eco-political economy of AI: understanding its global production/supply chain and its life cycle

To recognise the environmental harms of AI, the starting point of every discussion should be an analysis of its global supply chains, starting with the extractivism and neglect of social and environmental justice (NRDC 2022) in terms of the environmental costs that AI currently has and which lie in the production, transportation, training and disposal of the technologies on which it operates (Brevini 2021).

In order to produce the material resources needed for AI, we need to start with the extraction of rare metals and mineral resources which follow the logics of colonialism (NRDC 2022). In her work on digital developments with humanitarian structures,

Mirca Madianou has developed the notion of 'technocolonialism' in order to analyse how 'the convergence of digital developments with humanitarian structures and market forces reinvigorate and rework colonial legacies' (Madianou 2019: 2). The same colonial genealogies and inequalities characterise the global production/supply chains of artificial intelligence, as the extractive nature of technocolonialism resides in the minerals that need to be mined to make the hardware for AI applications. So, for example, the demand for mineral resources is growing exponentially: the European Communication has stressed that the demand for lithium in the EU, mainly for use in batteries, is projected to rise by 3,500% by 2050 (European Commission 2022).

Moving to the second section of the global production/supply chain, the production of AI models also shows high environmental costs. A study published by the College of Information and Computer Sciences at University of Massachusetts Amherst (Strubell et al. 2019) quantifies the energy consumed by running artificial intelligence programs. In the case examined by the study, a common AI linguistics training model can emit more than 284 tonnes of carbon dioxide equivalent. This is comparable to five times the lifetime emissions of the average American car. It is also comparable to roughly 100 return flights from London to New York (Brevini 2021). Moreover, more recent studies focusing on ChatGPT have highlighted the urgency of recognising the massive water footprint caused by AI models (George et al. 2023; Microsoft 2022).

Additionally, artificial intelligence relies on data to work. At present, cloud computing eats up energy at a rate somewhere between the national consumption of Japan and that of India (Greenpeace International 2011; Murdock and Brevini 2019). Today, data centres' energy use averages 200 terawatt hours each year (Jones 2018; IEA 2017): more than the national energy consumption of many countries, including Iran. Moreover, the information and communications technology (ICT) sector, that includes mobile phone networks, digital devices and television, accounts for 2% of global emissions (Jones 2018). Greenhouse gas emissions from ICT could grow from roughly 1-1.6% in 2007 to exceed 14% worldwide by 2040, accounting for more than half of the current relative contribution of the whole transportation sector. Additionally, data centres require large, continuous supplies of water for their cooling systems, raising serious policy issues in places like the US and Australia where years of drought have ravaged communities (Brevini 2021; Sensorex 2022).

Thirdly, AI development is based on a model of surveillance capitalism (Brevini 2021: 45; Zuboff 2019) enhanced by data extraction, analysis and monetisation. This, in turn, has contributed to the development of increased consumption habits. Facilitated by decades of unregulated capitalism, AI services and products bear major responsibility for generating the uberconsumerism which surrounds digital services and the destructive hyperconsumption that leads to unattainable energy demands (Brevini 2021). New developments in AI – especially neural networks – place high demands on energy while the gains in efficiency currently achieved in data centres have proved very slow in compensating for the escalating demands of computational power.

Lastly in the global AI supply chain, when communication and computational machines are discarded they become electronic waste, saddling local municipalities with the

challenge of safe disposal. This task is so burdensome that it is frequently offshored and many countries with developing economies have become digital dumping grounds for more privileged nations, as the case of Kenya demonstrates (Napainoi 2021).

To make things worse, while holding out the promise of solving the Climate Emergency, AI companies are marketing their offers and services to coal, oil and gas companies, thus compromising efforts to reduce emissions and divest from fossil fuels. A new report on the future of AI in the oil and gas market published by Zion Market Research found that AI in oil and gas is expected to reach around 4 billion dollars globally by 2025 from 1.75 billion in 2018 (Zion Market Research 2019).

5. Conclusion

Developing an Eco-political economy of AI that entails a focus on its global production/supply chain enables us to grasp its real environmental toll.

We need to ask who should own and control the essential infrastructures that power artificial intelligence and, at the same time, be sure to place the Climate Emergency at the centre of the debate. For what purposes, and with what consequences for collective wellbeing, should we shape artificial intelligence? What values should guide its development if we want to address the Climate Emergency? At the time of writing, there are a number of international agreements, position papers and guidelines that are being discussed, initiated in global forums or at national levels, illustrating that progress is being made. For example, UNESCO's recently adopted recommendation on artificial intelligence explicitly clarifies that 'if there [is a] disproportionate negative impact of AI systems on the environment (…) they should not be used' (UNESCO 2021).

There is a clear need to demand climate accountability from those who own cloud computing operations and data centres. One crucial intervention could be to adopt government-mandated green certifications for server farms and centres to achieve zero emissions, given AI's increasing computing capabilities.

Moreover, a Tech Carbon Footprint Label, providing information about the entire global supply chain of the AI devices we use, from the raw materials used, the carbon costs involved and the recycling options that are available, could be implemented. This would result in stronger public awareness about the implications of adopting a piece of smart technology.

Making transparent the energy used in producing, transporting, assembling and delivering the technology we use daily would enable policymakers to make more informed decisions and the public to make more informed choices. Added to this could be policy intervention which requests manufacturers lengthen the lifespan of smart devices and provide spare parts to replace faulty components. Global policymaking should encourage educational programmes to enhance green tech literacy and raise awareness of the costs of hyperconsumerism as well as the importance of responsible energy consumption. Green tech literacy programmes should also entail interventions

to ban the production of products that are too demanding in data terms and that deplete energy too significantly.

As artificial intelligence, like all technologies, is always in 'a full sense social' (Williams 1981: 227), the choice to develop the kind of 'green AI' that can enhance environmental sustainable goals rests with us. Unfortunately, the current development of AI does not display the kind of environmental commitment that is needed to address the Climate Emergency we are facing. An Eco-political economy of AI could, however, lead us in the right direction.

References

Belkhir L. and Elmeligi A. (2018) Assessing ICT global emission footprint: Trends to 2040 and recommendations, Journal of Cleaner Production, 177, 448–463. https://doi.org/10.1016/j.jclepro.2017.12.239

Brevini B. (2020a) Black boxes, not green: Mythologizing artificial intelligence and omitting the environment, Big Data and Society 7 (2). https://doi.org/10.1177/2053951720935141

Brevini B. (2020b) Introduction, in Brevini B. and Swiatek P. (eds.) Amazon: Understanding a global communication giant, Routledge, 1–6.

Brevini B. (2021) Is AI good for the planet?, Polity.

Brevini B. and Murdoch G. (2017) Carbon capitalism and communication: Confronting climate crisis, Palgrave Macmillan.

European Commission (2020) White paper on artificial intelligence: A European approach to excellence and trust, COM (2020) 65 final, 19.2.2020. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission (2022) Communication from the Commission to the European Parliament and the Council, 2022 Strategic Foresight Report: Twinning the green and digital transitions in the new geopolitical context, COM(2022) 289 final, 29.6.2022. https://commission.europa.eu/strategy-and-policy/strategic-planning/strategic-foresight_en

George A.S., George A.H. and Martin A.G. (2023) The environmental impact of AI: A case study of water consumption by Chat GPT, Partners Universal International Innovation Journal, 1 (2), 97–104. https://doi.org/10.5281/zenodo.7855594

Greenpeace International (2011) How dirty is your data? A look at the energy choices that power cloud computing. https://www.greenpeace.org/static/planet4-international-stateless/2011/04/4cceba18-dirty-data-report-greenpeace.pdf

Guzman A. and Lewis S. (2019) Artificial intelligence and communication: A human–machine communication research agenda, New Media and Society, 22 (1), 70–86. https://doi.org/10.1177/1461444819858691

IEA (2017) Digitalisation and energy, International Energy Agency. https://www.iea.org/reports/digitalisation-and-energy

IEA (2019) World energy outlook 2019, International Energy Agency. https://www.iea.org/reports/world-energy-outlook-2019

IEA (2020) Global energy review 2020, International Energy Agency. https://www.iea.org/reports/global-energy-review-2020

IPCC (2022) The evidence is clear: The time for action is now. We can halve emissions by 2030, The Intergovernmental Panel on Climate Change. https://www.ipcc.ch/2022/04/04/ipcc-ar6-wgiii-pressrelease/

Jones N. (2018) How to stop data centres from gobbling up the world's electricity, Nature News Feature, 12 September 2018. https://www.nature.com/articles/d41586-018-06610-y

Madianou M. (2019) Technocolonialism: Digital innovation and data practices in the humanitarian response to the refugee crisis, Social Media and Society, 5 (3). https://doi.org/10.1177/2056305119863146

Microsoft (2022) Microsoft environmental sustainability report 2022. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW15mgm

Mosco V. (2004) The digital sublime: Myth, power, and cyberspace, MIT Press.

Mosco V. (2017) The next Internet, in Brevini B. and Murdock G. (eds.) Carbon capitalism and communication, Palgrave Macmillan, 95–107.

Murdock G. and Brevini B. (2019) Communications and the capitalocene: Disputed ecologies, contested economies, competing futures, The Political Economy of Communication, 7 (1), 51–82.

Naipanoi L. (2021) Dumped e-waste threatens Kenyan lives, contributes to global warming, The Elephant, 6 November 2021. https://www.theelephant.info/features/2021/11/06/dumped-e-waste-threatens-kenyan-lives-contributes-to-global-warming/

NRDC (2022) Lithium mining is leaving Chile's indigenous communities high and dry (Literally), Natural Resources Defense Council. https://www.nrdc.org/stories/lithium-mining-leaving-chiles-indigenous-communities-high-and-dry-literally

Sensorex (2022) Data centre water usage challenges. https://sensorex.com/2022/08/16/data-center-water-usage-challenges/

Strubell E., Ganesh A. and McCallum A. (2019) Energy and policy considerations for deep learning in NLP, Cornell University. https://doi.org/10.48550/arXiv.1906.02243

Williams R. (1981) Communication technologies and social institutions, in Williams R. (ed.) Contact: Human communication and its history, Thames and Hudson.

World Economic Forum (2018) Harnessing artificial intelligence for the earth. https://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf

UN (2022) New UN weather report "a chronicle of chaos": UN chief, News, 6 November 2022. https://news.un.org/en/story/2022/11/1130237

UNESCO (2021) UNESCO member states adopt the first ever global agreement on the ethics of artificial intelligence, Press Release, 25 November 2021. https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-in

WMO (2022) Provisional state of the global climate report, World Meteorological Organization. https://storymaps.arcgis.com/stories/5417cd9148c248c0985a5b6d028b0277

Zion Market Research (2019) Global AI in oil and gas market will reach to USD 4.01 Billion by 2025, GlobeNewswire, 18 July 2019. https://www.globenewswire.com/news-release/2019/07/18/1884499/0/en/Global-AI-In-Oil-and-Gas-Market-Will-Reach-to-USD-4-01-Billion-By-2025-Zion-Market-Research.html

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

All links were checked on 26.01.2024.

Cite this chapter : Brevini B. (2024) An Eco-political economy of AI: environmental harms and what to do about them, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 7
'End-to-end' ethical AI. Taking into account the social and natural environments of automation

Antonio A. Casilli

1. Introduction

In the last few decades, AI applications have largely been driven by machine learning paradigms. Instead of coding rules for every possible occurrence, machine learning algorithms can identify patterns in vast amounts of data to detect trends, find solutions and predict future events. Large language models like ChatGPT, recommendation algorithms on e-commerce platforms and operating systems for autonomous vehicles all use these techniques to learn from data.

Science, politics and industry have also focused on the ways in which these machine learning models influence behaviour. Although fears concerning artificial general intelligence are unrealistic, and an AGI (Artificial General Intelligence) that rivals or surpasses human cognition remains largely science fiction, AI's scale is especially concerning. Privacy and surveillance concerns are related to the data that these technologies require. Automation improves as information becomes richer, while data used for the development of machine learning models may include private or personally identifiable data. It is not only governments and police forces who conduct surveillance, but also businesses and the private sector.

The application of machine learning can, furthermore, pose a challenge to the transparency and democracy of decision-making since even the programmers cannot systematically determine which characteristics of the data the system has used in order to generate solutions. Machines can 'learn' from historical data that is biased to recognise, for instance, light-skinned men more accurately than dark-skinned women, to select men over women in recruitment processes or to predict reoffending risks that are higher for black than white defendants (Müller 2021).

Due to these recent advances, efforts to formulate ethical guidelines have mushroomed. In its crowdsourced AI Ethics Guidelines Global Inventory, the NGO AlgorithmWatch includes 173 such documents (AlgorithmWatch 2023). However, the most comprehensive analysis of 84 ethics charters has been published by three researchers at ETH Zurich, in Switzerland (Jobin et al. 2019). They identify five recurrent themes in their corpus: transparency; justice and fairness; non-maleficence; responsibility; and privacy. But each document defines these ethical domains differently and, more importantly, these principles tend to be operationalised mathematically and implemented as technical measures. Almost all the existing guidelines focus on machine learning's ability to resolve ethical issues (Hagendorff 2020). By reducing ethics principles to their technical

aspect, the proponents of ethical AI guidelines are neglecting broader socioeconomic, geographical and institutional factors.

The implications are profound and we must therefore take a holistic view, looking at the whole system in which AI is incorporated. The question to ask is: 'whose values are prioritised in these AI ethics guidelines?'. According to Anna Jobin and her colleagues, the geographical distribution of the issuers of ethics guidelines highlights hotspots and hubs in Europe and the United States, as well as Japan and India (China, which is another major player, was not included in the study). All these countries are heavily investing in artificial intelligence. On the contrary, countries from the Global South are absent from their map.

It is not a surprise, then, that AI ethics charters overlap geographically with producers of AI solutions in general, even though this may result in conflicts of interest. For example the principle of responsibility, if applied effectively, may interfere with the continuous exponential expansion of technologies and commercialisation. By the same token, minimising privacy risks may disrupt data collection and severely hamper machine learning. A strict adherence to ethical principles would clash with free market ideologies. A key responsibility of AI ethicists is to minimise such clashes.

Therefore, ethical AI is consistent with the tradition of science that addresses industry needs without challenging corporate motives. Existing guidelines are developed by or with the contribution of technology companies. As another means of serving the industrial status quo, AI ethicists refer to elite engineers as arbiters of 'bias' while excluding scholars and advocates who denounce power dynamics and economic imbalances. Meredith Whittaker suggests that corporate actors use ethical AI to 'co-opt and neutralise critique'. This is done in part by funding the 'weakest critics, often institutions and coalitions that focus on so-called AI ethics, and frame issues of tech power and dominance as abstract governance questions' (Whittaker 2021: 54).

Companies' implicit or explicit involvement in AI ethics fills the current regulatory vacuum – and ultimately contributes to it. The emphasis on in-house AI ethics prevents the development of legally mandated standards and enforcement mechanisms. The tech industry assumes it can formulate ethical norms for AI and ensure compliance, but the absence of any discussion of the effective ways of enforcing ethics standards in these charters proves that this assumption does not hold. Even when ethical principles are operationalised through specific tools, they fail to address existing imbalances. Another study conducted on 169 ethics documents finds that only 39 include AI ethics tools such as lists of best practice, checklists and adapted software applications. The most important aspect, however, is that key stakeholders were excluded from the design of these tools and that there was no external auditing (Ayling and Chapman 2022).

Admitting that AI ethics discourses are connected to the commercial interests of tech companies is the first step towards acknowledging that AI is industrially produced using human and natural resources. This industrial system requires appropriate corporate structures, capital investments and institutional backing. Within what context is AI produced? Who contributes and for what? When viewed from the standpoint of society

as a whole, what are its production costs? Virtually none of these questions are addressed in AI ethics charters which implicitly assume that voice assistants, recommendation engines and self-driving cars raise ethical concerns only when consumers use them. As in other areas of sociopolitical research on AI, ethics has historically put its emphasis on the possible effects of technology only at the deployment and in the marketing phases.

2. The AI production process

The production process behind AI is crucial long before the deployment of products and solutions. This is particularly clear if we consider the example of autonomous vehicles (Tubaro and Casilli 2019). Besides engineers, software developers and designers, safety drivers are also needed. These drivers travel inside the car, monitor the trip, provide feedback to the technical team and are expected to take control of the vehicle if necessary. The development of self-driving cars, however, also requires a vast army of hidden workers. Some refer to these as 'data workers' since they manage information; others as 'microworkers' since their tasks are fragmented and are viewed as less important than those of data scientists and software developers. For autonomous vehicles to be safe, computer-vision algorithms must be capable of recognising pedestrians crossing the street, for example. How does the car know what a pedestrian looks like? Generally, these examples are based on large sets of image data. The images routinely taken by cameras and sensors mounted on autonomous cars provide precisely this. However, these images need to be labelled before they can be used. In a traffic photo, the computer needs to read tags indicating 'pedestrian', 'bike', 'traffic light', 'bus', etc. Human workers are paid to add these tags, thus making everything visible to the AI. It is a huge job because the car's algorithm cannot learn from small amounts of data. Annotation would be tedious and lengthy if only a few workers were involved. But by fragmenting these large batches into many short, one-shot tasks, and assigning them to many data workers, each of whom perform just one or two, the goal can be achieved.

Human work is performed off-street by data annotators who remotely tag the images for the development of autonomous cars. Under what conditions are these human workers recruited, remunerated, managed and facilitated in exercising their rights?

Autonomous cars do not represent an isolated case when it comes to production-related ethics issues, but rather exemplify the trends seen throughout AI. Human labour and the environment are the two main issues that arise.

Despite the high value the AI industry places on its visible workforce of software developers, data scientists and computer engineers, it tends to ignore and to render invisible its lower-level microworkers who are indispensable, despite performing repetitive and often unqualified data tasks. It is these invisible humans that intervene at various stages during the development process of machine learning models: the initial training of the models (data generation and annotation); the verification of their

outputs after deployment; and, sometimes, performing the real-time correction or 'impersonation' of AI systems.[1]

Mechanical Turk, Amazon's microtasking platform, popularised human-powered data work for AI in the middle of the first decade of the 2000s. The name comes from an Ottoman-dressed chess-playing automaton from the eighteenth century. This 'proto-AI' could supposedly simulate the cognitive processes of a real chess player. The original mechanical Turk, however, was a hoax, controlled by a hidden operator. Today, it serves as a metaphor for the 'human-in-the-loop' principle that governs AI production. Human workers still prepare, test and sometimes pose as autonomous systems, but on a much larger scale. Amazon Mechanical Turk has hundreds of thousands of freelance workers who perform human intelligence tasks (HITs) which are fairly straightforward for humans but difficult for machines: recognising objects, creating lists, transcribing short sentences, etc. This microwork has been described by Amazon's founder Jeff Bezos, without a hint of irony, as 'artificial artificial intelligence' (Casilli and Posada 2019).

There have been many up and coming companies entering this market since Mechanical Turk launched. In addition to the Australian Appen, the American Remotasks or the German Clickworker, several international platforms provide microworking services on demand. Technology multinationals have created their own microworking services where they act as the sole recruiter, such as Microsoft's UHRS (Universal Human Relevance System) and Google's RaterHub. Companies that operate as BPO (business process outsourcing) vendors can also recruit workers for their clients in countries where the workforce is cheaper, such as India or Venezuela. Among them are some very big companies and platforms; others are smaller and sometimes specialised, for instance IsAHit and Wirk in France. A few smaller start-ups have been acquired by larger companies, such as Mighty AI, which is dedicated to the automotive industry and was acquired by Uber Advanced Technologies Group in 2019. Subsequently, the entire ATG was acquired by a startup, Aurora, in which Uber holds a 40% ownership stake. Companies, platforms and startups are all involved in complex arrangements which shows that tech companies need to be agile. A large number of intermediaries must be used by AI developers to recruit disposable workers.

3. Invisibility and the working conditions of microworkers

The conditions under which this type of work is performed are problematic.

Microworkers rarely figure on the payroll: typically, platforms bind them through membership or participation contracts similar to the general terms of service that are routinely agreed to by consumers of internet services and mobile apps. This leaves workers vulnerable to market volatility and without social protection. In addition, they work remotely from anywhere, which opens them up to worldwide competition and lower wages. The common practice of paying by the piece rather than by the hour or by a monthly salary makes it difficult to control, even though some platforms (such as

1. This three-part conceptual framework is further developed in Tubaro et al. (2020).

Clickworker) recommend paying the minimum wage. Microtasks can be paid as little as a few cents in the most dire cases. In a report published in 2018, the International Labour Organization estimated that, on average, microworkers earn 3.31 dollars per hour (counting the time they spend searching for new tasks to perform), well below the minimum wage in most countries (Berg et al. 2018). TIME revealed in January 2023 that workers on one of OpenAI's subcontracting platforms in Kenya were earning only between 1.34 and 2 dollars per hour when they were annotating data for ChatGPT (Perrigo 2023).

Economic necessity often motivates the workers who perform these data tasks in spite of the low wages. The situation is not limited to low- and middle-income countries. Other research has found that low wages and poverty-level workers are over-represented even among microworkers in France, a high income country (Casilli et al. 2019). Women with young children often work part-time and supplement their income with microtasks. There are no clear indications that they can derive additional benefits in terms of career progression; in the future their skills won't make them particularly attractive to employers (Tubaro et al. 2022).

This contradicts some of the fundamental principles outlined in the AI ethics charters discussed above, including fairness, justice and transparency – and even privacy, as workers are sometimes asked to provide personal data, such as selfies and voice recordings, for the creation of machine learning datasets. The industry cannot even meet its own standards, regardless of how vague they may be in the first place.

Microwork's invisibility, low wages and precarious status are problematic, as its geographical distribution makes all the more clear. In contrast to charters and guidelines, which are usually published in high income countries, ongoing research on data production shows what a flipped image of ethical AI looks like. The majority of microworkers are located in the Global South, despite it being Global North countries that have attracted the most research attention (Difallah et al. 2018). Efforts to map their global distribution suggest that they reproduce legacy inequalities based on wealth, power and geographical influence (Graham et al. 2017), although this is perhaps a natural conclusion for results that were obtained by analysing English-speaking platforms, where clear links appear with former British colonies, such as India, or former zones of US hegemony, such as the Philippines (Gray and Suri 2019).

Based on these pioneering studies, team members at DiPLab (Digital Platform Labour at the Institut Polytechnique de Paris) have conducted extensive research in French-speaking African countries (including Madagascar, Cameroon, Mali, Senegal, Morocco and Egypt), as well as Portuguese and Spanish-speaking Latin American countries which mainly serve North American tech companies (Le Ludec et al. 2023; Viana Braz et al. 2023).[2] Venezuela, Argentina (Miceli and Posada 2022) and Brazil (Grohman and Fernandes Araújo 2021) are among the most active countries in this international market.

2. Results were obtained within the framework of DiPLab's projects HUSH (Human Supply Chain of Smart Technologies), funded by the French National Research Agency (ANR); and TRIA (El Trabajo de la Inteligencia Artificial), funded by the French Centre for Scientific Research (CNRS).

Figure 1 **Global flows of annotated data from microwork providers to AI solution providers**

Source: DiPLab's HUSH and TRIA projects. Author's elaboration.

On the basis of current evidence, Figure 1 schematically represents the global flows of data and work that are feeding the development of AI. The Latin American workforce serves technology producers in North America (particularly the United States) and Europe. From South and Southeast Asia, work primarily goes to North America but also to China. Africa provides data work to Europe and China. The latter has substantial internal flows, despite little knowledge of them, as do both Europe (with some east-west flows) and the United States. As the exact size of the flows is still not documented in many cases, the map is only indicative. However, it reveals a major flaw in current AI ethics approaches: the severe underrepresentation of the Global South in the creation of charters and guidelines, as well as its overrepresentation within the 'human supply chain' that produces AI and undermines diversity and cultural awareness – and, consequently, reduces the voice of workers.

4. Natural resources

Along with the labour force required to annotate and enrich data, AI consumes natural resources such as energy, minerals and metals in its construction as well as energy in the performance of heavy calculations. This raises questions about sustainability, a topic rarely mentioned in the charters and guidelines.

For AI to be effective, it must respect the natural environment where resources are used as well as the human environment where labour is produced – which explains why recent years have seen a rise in attention to the environmental costs of AI. The challenge has been met in two fundamentally different ways: by quantifying the carbon footprint

of computational processes; and by denouncing the extractivist logic underlying the tech industry.

Several efforts have been made to measure the environmental cost of machine learning. Training one large Natural Language Processing (NLP) transformer model generates nearly five times the amount of carbon dioxide of a single car's annual emissions, or 50 times the amount emitted by one individual in a lifetime (Strubell et al. 2019). Various trackers and impact measurements help researchers assess the energy use of their tools and make actionable recommendations to reduce emissions (Bannour et al. 2021). Many companies, including Alphabet, are investing heavily in green energy sources and developing more efficient ways to cool their data centres. AI training can also be enhanced with 'green algorithms', efficient architectural settings and smaller models (Cai et al. 2019). Even though these efforts are commendable, it must be acknowledged that, once again, they rely on self-regulation and assume that the tech industry can and will follow. Therefore, one must assume that the technological systems that pollute and waste natural resources also have the potential to mitigate them.

Yet, over-reliance on portmanteau concepts such as 'environment', 'climate' and 'energy' obscures the concrete human and material substratum that is supporting the transformation of natural resources into AI. Other researchers have thus developed a more radical theoretical perspective centred on digital extractivism to analyse the materiality of AI, its reliance on natural resources and its links to capitalism (see Brevini in this volume; Iyer et al. 2021). Rather than referring to extraction as a plundering of natural resources, they use the notion to describe how capital interacts with and draws on human, political, economic and social activity. In contemporary capitalism, extractivism is ubiquitous, spanning not only traditional sectors like logistics and agriculture but also intangible activities like finance.

Data production and AI are new frontiers of extraction in this sense. Algorithms and data would not function without the minerals and metals that form the core hardware components that host and compute them. Several countries, including Zimbabwe, Madagascar, Bolivia, Argentina and Chile, mine minerals like cobalt, nickel and lithium for batteries. Indonesia is a source of tin for high income regions and countries like Europe and the US, while those that produce semiconductors, like Taiwan and South Korea, can be seen in the same light. Author Kate Crawford calls the 'mineralogical layer of AI' the foundation of the informational infrastructure that fuels intelligent solutions (Crawford 2021). The environmental costs of this must also be taken into account in the moving of minerals, metals, fuel, hardware and final products internationally.

A map of the locations where most natural resources are extracted for the AI industry overlaps significantly with the trajectories of microwork flows. As data workers from the Global South are mobilised for the AI industry of the North, those who work in mining, transport and electronic waste management follow the same global patterns. The production of information technologies is also part of an international division of labour (Fuchs 2016). Information and contents circulating from remote servers to our screens are not just produced by data processing activities. Instead, they revolve around the extraction and transportation of the minerals used in electronics.

Today's production processes for AI have significant limitations that would go unnoticed if we only focused on the principles emphasised in current public debates and ethics guidelines. But there are ways of overcoming these deficiencies. In accordance with the analyses advanced by Kate Crawford (2021), Gunay Kazimzade and Milagros Miceli (2020), a new, alternative research programme would adopt a dual approach to AI by taking into account not just its technological but its social, economic and political context as well. In addition to the now popular depiction of AI as a data intensive technology, it raises the important question of how such data is created.

5. Conclusion

End-to-end ethical AI requires a consideration of the conditions of production of the data, tools and equipment used to manufacture and market these systems. A number of consumer products already apply this kind of ethical reasoning. For example, footwear and processed food manufacturers do not discriminate against their consumers based on gender, location or other factors – they are not as biased as AI in this respect. However, that does not make them ethical. A company is considered ethical if it respects workers' rights, provides decent working conditions and minimises its environmental impact. In a similar vein, an ethical AI must minimise negative externalities both within and outside human communities and the same with regard to its production processes.

Taking account of the human, social, political and economic contexts of today's AI technologies can provide novel insights and suggest directions for future action. It is important to protect the rights of remote AI workers. They should be able to resist unsuitable conditions if they have the opportunity. They should be able to protest against technological systems they consider ethically problematic and they should be able to refuse to contribute to them. In such a scenario, AI ethics shifts from promoting the interests of producers-owners to promoting the ethical agency of producers-workers. For this to succeed, it is necessary to acknowledge the invisible work of preparation, verification and impersonation of automated systems in order to establish a stronger, more fundamental approach to AI ethics. As a result, workers should be provided with methods of protection. Other workers directly and indirectly involved in the supply chain of modern computing devices could benefit from the same principles.

References

AlgorithmWatch (2023) AI ethics guidelines global inventory. https://inventory.algorithmwatch.org/

Ayling J. and Chapman A. (2022) Putting AI ethics to work: Are the tools fit for purpose?, AI and Ethics, 2, 405–429. https://doi.org/10.1007/s43681-021-00084-x

Bannour N., Ghannay S., Névéol A. and Ligozat A.L. (2021) Evaluating the carbon footprint of NLP methods: A survey and analysis of existing tools, Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, 11–21. https://aclanthology.org/2021.sustainlp-1.2/

Berg J., Furrer M., Harmon E., Rani U. and Silberman M.S. (2018) Digital labour platforms and the future of work: Towards decent work in the online world, ILO. https://www.ilo.org/global/publications/books/WCMS_645337/lang--en/index.htm

Cai H., Gan C., Wang T., Zhang Z. and Han S. (2019) Once-for-all: Train one network and specialize it for efficient deployment. https://doi.org/10.48550/arXiv.1908.09791

Casilli A.A. and Posada J. (2019) The platformization of labor and society, in Graham M. and Dutton W.H. (eds.) (2019) Society and the Internet. How networks of information and communication are changing our lives, 2nd ed., Oxford University Press, 293–306. https://doi.org/10.1093/oso/9780198843498.003.0018

Casilli A.A., Tubaro P., Le Ludec C., Coville M., Besenval M., Mouhtare T. and Wahal E. (2019) Le micro-travail en France. Derrière l'automatisation, de nouvelles précarités au travail ?, Projet de recherche DiPLab.

Crawford K. (2021) Atlas of AI: Power, politics, and the planetary costs of artificial intelligence, Yale University Press.

Difallah D., Filatova E. and Ipeirotis P. (2018) Demographics and dynamics of mechanical Turk workers, Proceedings of the eleventh ACM international conference on web search and data mining, 135–143. https://doi.org/10.1145/3159652.3159661

Fuchs C. (2016) Digital labor and imperialism, Monthly Review, 67 (8), 14–24. https://doi.org/10.14452/MR-067-08-2016-01_2

Graham M., Hjorth I. and Lehdonvirta V. (2017) Digital labour and development: Impacts of global digital labour platforms and the gig economy on worker livelihoods, Transfer, 23 (2), 135–162. https://doi.org/10.1177/1024258916687250

Gray M.L. and Suri S. (2019) Ghost work: How to stop Silicon Valley from building a new global underclass, Houghton Mifflin Harcourt.

Grohmann R. and Fernandes Araújo W. (2021) Beyond mechanical Turk: The work of Brazilians on global AI platforms, in Verdegem P. (ed.) AI for everyone? Critical perspectives, University of Westminster Press, 247–266. https://library.oapen.org/bitstream/handle/20.500.12657/58191/1/ai-for-everyone.pdf#page=256

Hagendorff T. (2020) The ethics of AI ethics: An evaluation of guidelines, Minds and Machines, 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Iyer N., Achieng G., Borokini F. and Ludger U. (2021) Automated imperialism, expansionist dreams: Exploring digital extractivism in Africa, Pollicy. https://archive.pollicy.org/wp-content/uploads/2021/06/Automated-Imperialism-Expansionist-Dreams-Exploring-Digital-Extractivism-in-Africa.pdf

Jobin A., Ienca M. and Vayena E. (2019) The global landscape of AI ethics guidelines, Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kazimzade G. and Miceli M. (2020) Biased priorities, biased outcomes: Three recommendations for ethics-oriented data annotation practices, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 71. https://doi.org/10.1145/3375627.3375809

Le Ludec C., Cornet M. and Casilli A.A. (2023) The problem with annotation. Human labour and outsourcing between France and Madagascar, Big Data and Society, 10 (2). https://doi.org/10.1177/20539517231188723

Miceli M. and Posada J. (2022) The data-production dispositif, Proceedings of the ACM on Human-Computer Interaction, 6 (CSCW2), 1–37. https://doi.org/10.1145/3555561

Müller V.C. (2021) Ethics of artificial intelligence and robotics, in Zalta E.N. and Nodelman U. (eds.) The Stanford encyclopedia of philosophy, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/

Perrigo B. (2023) Exclusive: OpenAI used Kenyan workers on less than $2 per hour, Time, 18 January 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/

Strubell E., Ganesh A. and McCallum A. (2019) Energy and policy considerations for deep learning in NLP. https://doi.org/10.48550/arXiv.1906.02243

Tubaro P. and Casilli A.A. (2019) Micro-work, artificial intelligence and the automotive industry, Journal of Industrial and Business Economics, 46 (3), 333–345. https://doi.org/10.1007/s40812-019-00121-1

Tubaro P., Casilli A.A. and Coville M. (2020) The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence, Big Data and Society, 7 (1). https://doi.org/10.1177/2053951720919776

Tubaro P., Coville M., Le Ludec C. and Casilli A.A. (2022) Hidden inequalities: The gendered labour of women on micro-tasking platforms, Internet Policy Review, 11 (1), 1–26. https://doi.org/10.14763/2022.1.1623

Viana Braz M., Tubaro P. and Casilli A.A. (2023) Microwork in Brazil: Who are the workers behind AI?, Research Report DiPLab and LATRAPS. https://hal.science/hal-04140411

Whittaker M. (2021) The steep cost of capture, Interactions, 28 (6), 50–55. https://doi.org/10.1145/3488666

All links were checked on 29.01.2024.

Cite this chapter: Casilli A.A. (2024) ‘End-to-end’ ethical AI. Taking into account the social and natural environments of automation, in Ponce del Castillo A. (ed.) Artificial intelligence, labour and society, ETUI.

Part 3

Technological perspectives

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).

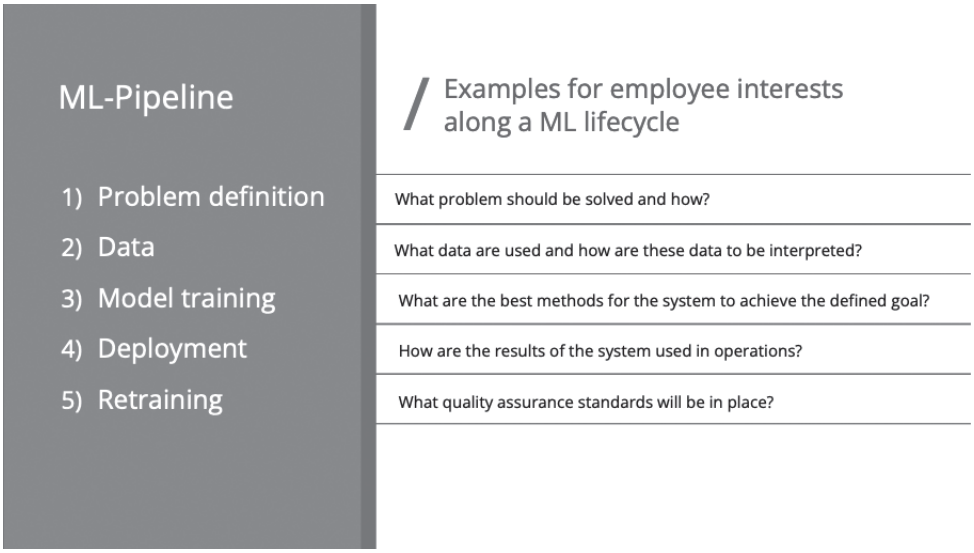
In the same way as choice of algorithm might affect transparency and robustness, other risks can be attributed to respective steps in the planning process or to specific technical components (Suresh and Guttag 2021). Because of the data-driven learning process, Machine Learning systems are, for instance, 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**



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.

Chapter 9
Measuring work is hard. Subcontracting it won't help. Explainable AI won't help

Sandy J. J. Gould

1. Introduction

An employer finds itself brought to an employment court by staff backed by their union. The staff are unhappy with the workplace monitoring that they are subject to, claiming that it constitutes excessive surveillance. To try and make sense of where monitoring becomes surveillance, Ponce del Castillo and Molè's chapter in this volume would be a very good place to start. In this chapter, though, I want to think about the decisions made by the employer that led to this position of conflict. Specifically, I want to explore what things an employer might decide to monitor, how it is that these decisions are made and how the availability of technology-based tracking might influence decision-making processes. These are important issues to consider as workplaces are increasingly instrumented with technologies like algorithmic management that depend on machines' abilities to sense and measure.

2. What to measure?

Researchers have written about 'digital Taylorism' and 'neo Taylorism' (e.g. Crowley et al. 2010; Goods et al. 2019), based on the idea that modern methods of technology-enhanced workplace surveillance take their cue from Taylor's Scientific Management. Much of this work has focused on workers' experiences of these new forms of surveillance. The focus on workers makes sense; they are often the ones at the 'sharp end' of surveillance. Writing on neo-Taylorism, Gautié et al. (2020: 788) identify that it 'is not so much performance measurement itself that has been politicized as the way in which it is linked to rewards and sanctions, and particularly to pay.' In this view, neo-Taylorism is a means of coercive control over staff, rather than a way of building evidence for business process design and planning. The prominence of worker control in contemporary narratives about workplace monitoring makes sense; the ways that technology constructs work are increasingly individualising (Tassinari and Maccarrone 2020), favouring individual measures of atomised output (Alkhatib et al. 2017), while they are also being fed into systems of algorithmic management that influence individual behaviour (Lee et al. 2015). The nature of these changes implies a focus on the individual experiences of workers.

The focus on the externalities felt by workers after technology has changed work has been accompanied by a broader analysis of the political economy of modern work. What has, perhaps, not received as much focus as it should is organisations' decision-making at the point at which choices about monitoring are being made. How do organisations

decide, specifically, what it is important and necessary for them to measure? Did that imaginary company brought to the tribunal realise what it was collecting and consider what the implications of that would be? How has technology, now one of the main mediating factors in workplace monitoring, changed how decisions are made?

Objectivist science (which Scientific Management and its descendent manifestations are trying to ape) seeks to measure phenomena in the world. To do this, appropriate measures must be developed and then operationalised such that they can be practicably made. 'Practicably' means having to accept resource constraints on what is measured, as well as the ontological limitations (some things are never going to be open to 'direct' measurement by any conceivable means). The extent to which the measure that we have been able to operationalise actually measures our phenomenon of interest pinpoints its construct validity. Higher construct validity means we have come closer to measuring our phenomenon of interest, but this is highly contestable – one of the things that scientists might focus on when they are reviewing their colleagues' work is the extent to which the latter have plausibly measured the thing they have set out to measure. Developing new measures with high construct validity is error prone and laborious. We might sometimes find that, despite our best efforts, we have not measured anything at all of consequence.

If, as that employer, I find myself in front of an employment tribunal, could I at least comfort myself with the knowledge that, even if what I had done was illegally intrusive and unethical, I had actually measured the thing about my staff's work that I wanted to measure? That, if nothing else, it had at least been accurate? Of course, this is not the way to think if you want to avoid ending up at a tribunal. But it does focus us on a question that has been lacking in discussions of neo-Taylorism. To what extent can new, technology-mediated ways of measuring things at work be successful even on their own terms let alone ones based on the legal, ethical and worker welfare dimensions?

Measuring things about work is difficult, especially work that does not lend itself easily to being broken down into independent atomic parts. How do you know if a member of staff is being productive? You could measure them using something like income from clients. But how much of that measure would be confounded by chance? What is it about those who bring in the most income that makes them effective? How could other staff be trained to be better at these things? These are difficult questions and, while sectors of the economy where work is more proceduralised, as with some kinds of manufacturing, have sought answers to all of these questions and sometimes returned with reasonably accurate answers, in less proceduralised sectors (e.g. some kinds of services work) it is not obvious that answers have been forthcoming. Managers' frustrations over staff working at home during the Covid-19 pandemic are a testament to this. Sometimes, the solution has simply been to accept inadequate measures and hope that, say, ranking on a bad measure can still produce a good result. Brankovic (2022) argues persuasively against these kinds of zero-sum approaches.

Given the fundamental epistemological challenges of measuring anything, rather than accepting the need, perhaps, for more interpretivist approaches to understanding work, it is easy (whether as a scientist or as a manager) to wonder whether technology might

be useful in closing the construct validity gap between what we are currently able to measure and what we might perhaps be able to measure if we had more technology. This technology might involve physical sensors. Perhaps we could use microphones to measure ambient noise levels? Perhaps we could put pressure sensors in seats to keep an eye on desk occupancy? Perhaps, where we find that a physical sensor can't access the information that we want, we can instead turn to virtual sensors. These might sense how many emails someone sends, or how many windows they have open on their computer. At times, might we want to combine physical and virtual sensors into a mash-up sensor that can tell us something that we couldn't have known with either alone? Maybe a system that combines microphones with metadata from virtual meetings to work out who spends more time speaking and how efficient they are with their speech? And where the technology can't measure the thing we want, perhaps we could get 'human-in-the-loop' sensors? We get people to measure things with their senses and require them to feed these into computer systems for further processing.

To what extent can technology get us closer to measuring the 'reality' of work? To what extent is it just another manifestation of technology solutionism – bringing more technology into complex sociotechnical systems in the hope that technology can solve hard problems with things that don't even lend themselves to objectivist measurement? Of course, as with all difficult problems, the conclusion is that it's a little complicated. However, we find ourselves in a moment when datafication is increasingly common and where the idea that data science can solve otherwise intractable problems is becoming more prevalent in business and politics. Given this zeitgeist, it seems apt to focus on the ways that technology cannot help solve the fundamental questions put by any neo-Taylorist system: what do we want to measure and what's the closest that we can get to measuring it?

3. Surveillance-as-a-Service

Deciding how to measure something is difficult, even after the decisions about whether something should or can be measured have been taken. Science, broadly interpreted, is a challenging undertaking, even for those who are suitably trained and experienced. I have written about the challenges that academic researchers face when making decisions about what to measure and how to measure it (Gould 2022), noting that these decisions are often 'contracted out' to other researchers or to other organisations. Building on the work of others is necessary in order to be able to make substantive progress, but it also sets up a consumption relationship in which the appeal of contracting-out is not having to understand what is happening 'under the hood' (see Anthony 2021 for examples of this happening at investment banks). Given the challenges for the trained and experienced, then, it is perhaps not a surprise to see a similar commodification and consumption of measurement in workplace contexts.

There are many organisations that now offer 'Surveillance-as-a-Service' (SaaS). West (2019) has used this phrase to describe the experiences of users of consumer technology (e.g. voice assistants), but here I am using it to describe organisations consuming

the workplace surveillance technologies provided by third parties. Time Doctor[1] has 250,000 users and includes 'distraction alerts', screen recording and 'productivity' measurement. Hubstaff[2] has nearly 600,000 active users and offers 'productivity' measurement through time-on-app tracking, GPS location tracking and device screenshot collection. Tools sold to employers will let them monitor staff on webcams, or produce a number next to every member of staff to say how much of a security risk they are (Corbyn 2022). These measures can be used by employers to different ends. They might be used to 'gamify' a particular activity, with leader boards and prizes used to incentivise more of a particular kind of productivity (O'Donnell 2014). They might also be used for performance management or promotion purposes (Ball 2010).

Surveillance tools are often sold as being a benefit to individual staff, with claims that they will help staff understand employer expectations and allow them to receive 'advice' from the system on meeting them. Workplace surveillance often comes under the guise of helping staff (Edwards et al. 2018). There are plenty of other tools like this, measuring which applications workers have open, how long they have them open for, how many emails they send and how quickly they respond to messages, and which also capture webcam feeds. Furthermore, there is significant overlap in the kinds of things that these systems can measure, which is what you'd expect from commodity off-the-shelf tools. They measure things that computing devices find it easy to measure: the name of a window open on the device; for how long it has been in focus; how much a mouse has moved. Sophisticated measures of productivity, however, require a nuanced understanding of work processes, and commodity software does not usually demonstrate the capacity for such nuance.

SaaS has significant implications for the way that workplaces are surveilled. As consumers of these services, organisations need to 'buy-in' to the measures for which these tools are able to collect data. To think they can tell you something about workplace productivity, you need to believe that time spent in a browser or the number of times someone looks away from their screen are reliable indicators of productivity. Perhaps these tools simply operationalise better the measures you are already trying to use. Or perhaps you're not sure how to work out how staff are productive; in this case, this software offers you two solutions: one to the problem that you don't know what your staff are doing to be productive; and another that measures these things and produces well ordered lists. What would you say when the tribunal asked why you had been collecting the data that you had? Would you just point to your third party supplier and claim they had understood productivity in your organisation better than you?

As an employer, taking third party measures of performance and using them for making commercial decisions comes with risks. Those risks are potentially further magnified by the use of algorithmic management in these tools. Algorithmic management involves not only the collection of data but automated reasoning and decision-making in relation to it. This might involve being directed to take certain trips in a ridesharing application (Lee et al. 2015; Möhlmann and Zalmanson 2017), or it might direct careworkers' time

1. Time Doctor: https://www.timedoctor.com
2. Hubstaff: https://hubstaff.com/

and attention (McCormick 2021). More subtly, we see this happening with 'focused inbox' tools that measure something about emails and then decide which ones to put in front of workers.[3]

Not only does using commodity workplace tracking tools risk measuring things that are not good measures of productivity, there is a bigger risk, which I have explored in other work (Gould et al. 2023), that organisational effort then coalesces around such measures. Work becomes designed around algorithmic management (Parent-Rocheleau and Parker 2022). So not only is productivity not being accurately measured, resources are then devoted to maximising performance on these weak measures, thus throwing good money after bad.

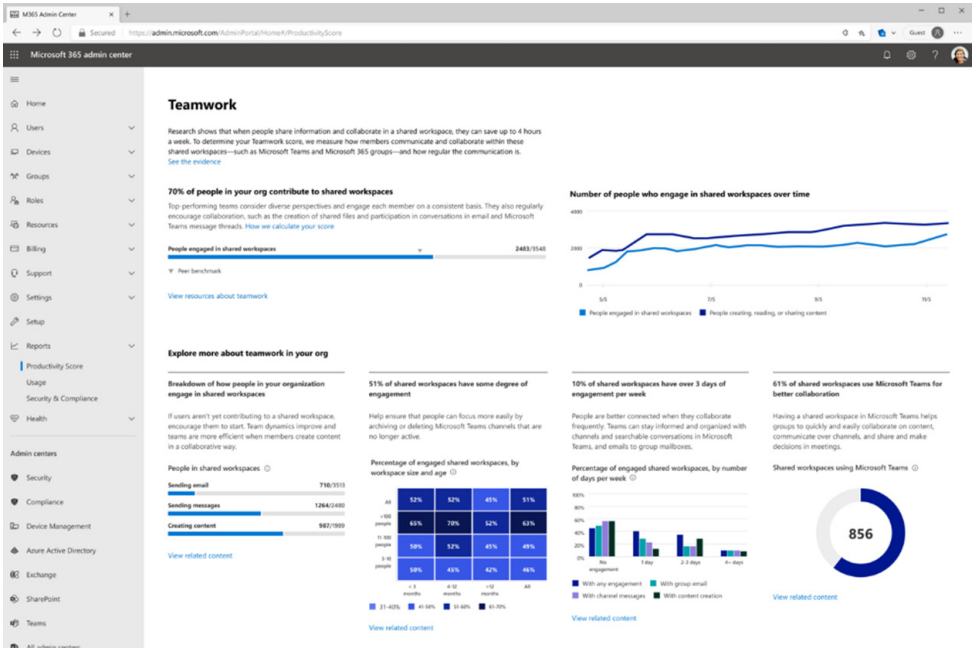
This issue is compounded further by another feature of these SaaS tools: 'analytics'. Decisions about what to measure can be outsourced through SaaS. But the process of drawing conclusions from those measures can also be outsourced. Microsoft's Office 365 package comes with 'productivity' tools[4] that produce reports about workers, an example of which is shown in Figure 1. In this example, on-device interactions are used to produce an analysis of a group's teamworking performance. To do this, it pulls data from the things it can monitor: emails, instant messages and digital meetings. It comes with a 'peer benchmark', or an implicit rank against other groups on those measures. After admitting to the tribunal that you had outsourced measurement of productivity to a third party, how would you explain that you'd also outsourced the interpretation of those measurements, too? It would be difficult to make a convincing case, so you'd need to work hard on your post hoc rationalisation.

Users of SaaS tools, whether employers or workers, can find that, even if they wanted to understand how these systems had produced their analyses, they could not. Many of the more cutting-edge tools make use of machine learning as part of their analyses. This might be to map some set of measures to some performance benchmark (e.g. sales) to identify why one team was producing more than another. Or it may use models to read correspondence and feed that into a performance analysis. Often there is no way to know how these systems are working, not just because they are proprietary to the third party SaaS provider but also because these models simply cannot produce explanations in terms that can be debated.

3. 'What is Focused Inbox?' https://support.microsoft.com/en-us/office/what-is-focused-inbox-16b24373-dfa9-4139-ab19-08aa753a6055

4. 'Microsoft Productivity Score and personalized experiences—here's what's new to Microsoft 365 in October' https://www.microsoft.com/en-us/microsoft-365/blog/2020/10/29/productivity-score-and-personalized-experiences-heres-whats-new-to-microsoft-365-in-october/

Figure 1 **A screenshot of Microsoft's 'Productivity Score' tool, nominally showing an analysis of 'teamwork' based on emails, instant messages and meetings**



Can the use of these SaaS technologies be regulated effectively, or even the datafication of the workplace more generally? To begin to think about effective regulation, we'd need first to understand what data it was that organisations were collecting and how they were using it. Organisations are unlikely to want to share information about this. When asked to produce data collected about workers, organisations have claimed that it is 'commercially sensitive' and declined to release it (van Doorn and Badger 2020). There is a second issue; even if an organisation wanted to release this kind of information, could it? As I have claimed, one of the reasons that you might wish to make use of SaaS as an organisation is because you don't understand what data your organisation holds about staff and prefer to contract this work out to other organisations. These organisations complete the reification of measurement by substituting what-it-is-desired-to-measure for what-it-is-possible-to-measure.

Convergence on off-the-shelf SaaS solutions could make regulation easier if it yielded standardised 'commodity' measures more amenable to regulation. If organisations did indeed work to these measures, then regulation of them might actually mean the regulation of the reality of work. It feels unlikely that such a purification would ever be completed, though; we could instead end up with significant gaps between work-as-specified (i.e. with commodity measures) and work-as-done which might demand 'real' measures (i.e. measures with good construct validity) in, say, a particular unit of an organisation. All the while, it is work-as-specified that gets regulated rather than the reality. It is not obvious that trying to regulate particular kinds of measurement would be fruitful.

4. 'Explainable AI'

A response to criticisms about the use of opaque learning systems might be 'use Explainable AI!' Many AI systems are opaque – there is no way to inspect how they produce a particular decision any more than putting a manager in an MRI scanner would tell you why they had chosen to use a particular piece of software. Recognising this limitation, Explainable AI (XAI) is something that researchers have been advocating. The idea is that these systems are able to produce a reasoning for a given output from a given input, and that this reasoning will be intelligible to human users of the system.

Explainable AI and its role in workplace surveillance (and attendant algorithmic management) has yet to see significant coverage in the literature. Tsiakas and Murray-Rust discuss the role of XAI in workplaces (2022), focusing on the potential benefits. The benefits are likely to come, Tsiakas and Murray-Rust propose, from augmenting human performance. There is an argument to be made for XAI in some contexts where a particular worker is able to refine AI outputs into their decision-making processes by using explanations from those same AI tools.

To what extent could XAI help employers make better decisions about what to measure at work or what constitutes work? That is not quite so obvious. XAI systems are still a form of task-focused intelligence which reason about the things they have been directed to reason about. Their 'explanations' are constrained by the parameters that have been set and the particular datasets that they have been trained on.

It seems that, in the context of understanding how work is happening, XAI does not yet have a role. Workplaces where measures of work are well operationalised with good construct validity might be able to use AI tools to identify patterns in seemingly disparate measures. An XAI tool might be able to explain the connections it has made, but no-one could expect it to answer the question of why those relationships exist. Even generative AI tools (e.g. conversational agents, image generators) are still relying on production from the latent space of a dataset. We don't have datasets for workplaces like we have for images or text corpuses. Without any authentic understanding of what happens in each workplace, of where productivity comes from, these systems (in their current form) are never going to be able to make the kinds of deductions that will improve a workplace. Right now, only people can do that and it is imperative that they do not abandon their responsibility for doing so – the consequences could well be a degraded working environment with worse productivity.

5. Conclusion

Technology for monitoring/surveilling/tracking workers, even sophisticated AI tools that can explain their reasoning, cannot solve the measurement question. Not everything important to productivity in the workplace is susceptible to measurement. Even the things that seem like they might be are not accessible through measures that are technically possible, ethical or legal. Actual neo-Taylorist approaches to work management would be very much more sophisticated in the approach to how work

was measured and analysed. Some very large technology companies seem to have the capacity to do this (for better or worse) and make competitive advantages out of it. But most of the commodity Surveillance-as-a-Service is a simulacrum of Scientific Management; there is nothing scientific about using commodity collection and analysis tools for complex work environments where productivity is hard to operationalise. Scientists struggle with the trade-offs associated with construct validity every day; there are no answers to be had from magic boxes taken off shelves.

References

Alkhatib A., Bernstein M.S. and Levi M. (2017) Examining crowd work and gig work through the historical lens of piecework, Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 4599–4616. https://doi.org/10.1145/3025453.3025974

Anthony C. (2021) When knowledge work and analytical technologies collide: The practices and consequences of black boxing algorithmic technologies, Administrative Science Quarterly, 66 (4), 1173–1212. https://doi.org/10.1177/00018392211016755

Ball K. (2010) Workplace surveillance: An overview, Labor History, 51 (1), 87–106. https://doi.org/10.1080/00236561003654776

Brankovic J. (2022) Why rankings appear natural (but aren't), Business and Society, 61 (4), 801–806. https://doi.org/10.1177/00076503211015638

Corbyn Z. (2022) 'Bossware is coming for almost every worker': The software you might not realize is watching you, The Guardian, 27 April 2022. https://www.theguardian.com/technology/2022/apr/27/remote-work-software-home-surveillance-computer-monitoring-pandemic

Crowley M., Tope D., Chamberlain L.J. and Hodson R. (2010) Neo-Taylorism at work: Occupational change in the post-Fordist era, Social Problems, 57 (3), 421–447. https://doi.org/10.1525/sp.2010.57.3.421

Edwards L., Martin L. and Henderson T. (2018) Employee surveillance: The road to surveillance is paved with good intentions. https://doi.org/10.2139/ssrn.3234382

Gautié J., Jaehrling K. and Perez C. (2020) Neo-Taylorism in the digital age: Workplace transformations in French and German retail warehouses, Relations industrielles/Industrial Relations, 75 (4), 774–795. https://doi.org/10.7202/1074564ar

Goods C., Veen A. and Barratt T. (2019) 'Is your gig any good?' Analysing job quality in the Australian platform-based food-delivery sector, Journal of Industrial Relations, 61 (4), 502–527. https://doi.org/10.1177/0022185618817069

Gould S.J.J. (2022) Consumption experiences in the research process, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 326, 1–17. https://doi.org/10.1145/3491102.3502001

Gould S.J.J., Rudnicka A., Cook D., Cecchinato M.E., Newbold J.W. and Cox A.L. (2023) Remote work, work measurement and the state of work research in human-centred computing, Interacting with Computers, 35 (5), 725–734. https://doi.org/10.1093/iwc/iwad014

Lee M.K., Kusbit D., Metsky E. and Dabbish L. (2015) Working with machines: The impact of algorithmic and data-driven management on human workers, Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 1603–1602. https://doi.org/10.1145/2702123.2702548

McCormick E. (2021) What happened when a 'wildly irrational' algorithm made crucial healthcare decisions, The Guardian, 2 July 2021. https://www.theguardian.com/us-news/2021/jul/02/algorithm-crucial-healthcare-decisions

Möhlmann M. and Zalmanson L. (2017) Hands on the wheel: Navigating algorithmic management and Uber drivers' autonomy, Proceedings of the International Conference on Information Systems (ICIS), 3. https://aisel.aisnet.org/icis2017/DigitalPlatforms/Presentations/3

O'Donnell C. (2014) Getting played: Gamification and the rise of algorithmic surveillance, Surveillance and Society, 12 (3), 349–359. https://doi.org/10.24908/ss.v12i3.5017

Parent-Rocheleau X. and Parker S.K. (2022) Algorithms as work designers: How algorithmic management influences the design of jobs, Human Resource Management Review, 32 (3). https://doi.org/10.1016/j.hrmr.2021.100838

Tassinari A. and Maccarrone V. (2020) Riders on the storm: Workplace solidarity among gig economy couriers in Italy and the UK, Work, Employment and Society, 34 (1), 35–54. https://doi.org/10.1177/0950017019862954

Tsiakas K. and Murray-Rust D. (2022) Using human-in-the-loop and explainable AI to envisage new future work practices, Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA '22), 588–594. https://doi.org/10.1145/3529190.3534779

van Doorn N. and Badger A. (2020) Platform capitalism's hidden abode: Producing data assets in the gig economy, Antipode, 52 (5), 1475–1495. https://doi.org/10.1111/anti.12641

West S.M. (2019) Data capitalism: Redefining the logics of surveillance and privacy, Business and Society, 58 (1), 20–41. https://doi.org/10.1177/0007650317718185

All links were checked on 01.02.2024.

Cite this chapter: Gould S.J.J. (2024) Measuring work is hard. Subcontracting it won't help. Explainable AI won't help, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 10
Standardising AI – a trade union perspective

Natalia Giorgi

1. Introduction

The European Commission's proposal for a regulation laying down harmonised rules on artificial intelligence,[1] the AI Act, is an internal market piece of legislation. Its legal basis, Article 114 of the Treaty on the Functioning of the European Union (TFEU),[2] provides for the adoption of measures to ensure the establishment and functioning of the internal market. As such, the AI Act is meant to form part of the Union harmonisation legislation for products[3] including a number of legislative initiatives that are mainly sector specific, for example the regulation on machinery products, the medical devices regulation, and the regulation on personal protective equipment. The Union harmonisation legislation aims to ensure that products placed on the European single market are safe; that they meet high health, safety and environmental requirements;[4] and that these products can circulate freely throughout the Union. The AI Act will follow the 'New Approach' and the New Legislative Framework which together form a legislative technique that is specific and unique to Europe under which standards developed through a public-private partnership are used for the technical implementation of the legislation.

The choice of legal basis is a decisive one because, as an internal market regulation, the AI Act is not meant to regulate the use of AI, for which it refers to existing European legislation, but essentially the design and development of AI systems, as products and services, before they are placed on the market and then put into use. Standardisation plays an essential role in the Union's harmonisation legislation as, in effect, it operationalises the legal acts.[5]

---

1. The proposal for a regulation laying down harmonised rules on artificial intelligence (COM(2021) 206 final) is entering, at the time of writing, negotiations with the Council on the final form of the law. The analysis provided in this chapter focuses on the Commission's proposal as it assesses the legal basis of the legislation and its relationship with standardisation neither of which have changed subsequent to the Council's General approach (2021/0106(COD), 6 December 2022) and the Parliament's negotiating position (P9_TA(2023)0236, 14 June 2023).
2. OJ C 202, 2016.
3. OJ C 247, 2022.
4. According to Article 114(3) of the TFEU (OJ C 202, 2016), the Commission, in its proposals envisaged in paragraph 1 concerning health, safety, environmental protection and consumer protection, will take as a base a high level of protection, taking account in particular of any new development based on scientific facts. Within their respective powers, the European Parliament and the Council will also seek to achieve this objective.
5. Regulation 1025/2012 on European standardisation highlights the important role of European standards within the internal market, 'for instance through the use of harmonised standards in the presumption of conformity of products to be made available on the market with the essential requirements relating to those products laid down in the relevant Union harmonisation legislation' (OJ L 316, 14.11.212. Recital 5).

In this context, it is worth reflecting on the conditions that need to be put in place and ensured, from a trade union perspective, when standardising AI to support the future legislation and whether ultimately this will be sufficient and adequate to ensure the protection of workers' rights as prescribed by the AI Act.

2. The New Legislative Framework; a legislative technique unique to Europe

The internal market legislation that falls under the New Approach'[6] and the New Legislative Framework[7] sets out the essential requirements a product or service must comply with in order to be placed on the market. The technical measures substantiating the legal requirements are then provided in standards. These provide a presumption of conformity to the law in the sense that, if a product is compliant with the standards, then it is presumed safe in accordance with the legislation and can be placed and circulate freely within the EU.

The AI Act, as an internal market regulation following the New Legislative Framework, sets out the essential requirements that AI systems must comply with in order to be placed on the market. Although it is horizontal by nature and aims to address all AI systems, the requirements it lays down are applicable to the design and development of AI systems that qualify as high risk.[8] In relation to the specific deployment of AI systems in a workplace context, those intended to be used in this area are considered high risk, in accordance with Annex III of the AI Act which sets out critical areas and use cases for systems considered as such. These systems will therefore need to comply with the legal requirements and go through the specific technical measures that operationalise them before they can be used in Europe.

To translate this into practice, if a company wishes to develop and place an AI system in the EU single market, it will first need to assess whether the system it aims to develop qualifies as high risk according to the AI Act. If it does, because for instance it is intended to be deployed for recruitment purposes (Annex III point 4 (COM(2021) 206 final), then the AI system will need to be compliant with the legal requirements and its design and development will need to be carried out in line with specific technical standards.

---

6. OJ C 136, 4.6.1985.

7. The new legislative framework is a package of measures that aim to reinforce the application and enforcement of internal market legislation. It consists of Regulation (EC) No 765/2008, which established the legal basis for accreditation and market surveillance and consolidated the meaning of the CE marking; Decision No 768/2008/EC which sets up a model to be used in preparing and revising Union harmonisation legislation with the aim of updating, harmonising and consolidating the various technical instruments already in use; in existing Union harmonisation legislation, and Regulation (EU) 2019/1020 on market surveillance and compliance of products (OJ C 247, 2022, p. 12).

8. The AI Act (COM(2021) 206 final) follows a risk-based approach. It determines the prohibited AI systems that pose unacceptable risks (Article 5); the AI systems considered as high risk that are permitted but subject to requirements and ex ante conformity assessment (Article 6); and the permitted AI systems with either a certain transparency risk, that are permitted but need to comply with information and transparency obligations, and others that pose minimal to no risk, which are permitted in the Union with no restrictions. The risk-based approach remains unchanged by the Council's General approach and the Parliament's positioning.

2.1 Harmonised standards

The standards which provide the technical requirements needed by manufacturers to meet the law are not ordinary ones; they are harmonised standards. To develop them, the European Commission prepares and issues standardisation requests addressed to the three recognised European standardisation organisations (ESOs) – that is, CEN, CENELEC and ETSI – which are governed by private law. These European standards are then adopted and referenced in the Official Journal of the European Union (OJEU) as providing a 'presumption of conformity' with the law. (This is in line with Regulation 1025/2012 on European standardisation.[9])

European standardisation is organised by and for the stakeholders concerned based on national representation.[10] In effect, participation in the work of CEN and CENELEC follows the 'national delegation principle' under which a national standardisation body (NSB) holds voting rights. Participation in European standardisation activities is therefore channelled via the NSBs in which 'mirror committees' are established to reflect the work of the European technical committees (ETUC 2022: 12).

It is important to note that, although the European Standardisation System (ESS) is open to the participation of all interested stakeholders, it is dominated by industry and remains characterised by an important deficit in terms of representation (ETUC 2020: 18). An independent review of the European standardisation system noted that 'industry remains the core element of the European standardisation system, being the main standards user and, at the same time, leading the contribution to technical standardisation work' (European Commission 2015: 102). Indeed, large companies collect direct benefits from participating in standardisation work and generally have the necessary resources and expertise to bring to the technical process. Organisations representing societal stakeholders, including trade unions, however, have no financial interest in participating in standardisation (European Commission 2015: 102). Their engagement only aims at representing and defending the interests of their constituencies.

2.2 Their legal effects

According to the 2016 ruling of the Court of Justice of the European Union in *James Elliot*,[11] harmonised standards produce legal effects.[12] Even though these standards are privately produced, they are part of EU law. The CJEU states in its judgment that, while the development of a harmonised standard:

9. OJ L 316, 14.11.2012, Art. 10.

10. OJ L 316, 14.11.2012, Recital 2.

11. Case C-613/14, p. 10. The court asks in this case 'about the legal nature of European harmonised standards for construction products and their relevance in contractual relationships between two private parties where reference is made to a national standard adopted pursuant to a harmonised standard in a contract for the supply of goods'.

12. Case C-613/14, para. 42.

> … is indeed entrusted to an organisation governed by private law, it is nevertheless a necessary implementation measure which is strictly governed by the essential requirements defined by that directive, initiated, managed and monitored by the Commission, and its legal effects are subject to prior publication by the Commission of its references in the 'C' series of the Official Journal of the European Union.[13]

The Commission plays an important role in this process as it issues a mandate, approves the ESO work programme, decides on the compliance of the draft harmonised standard with the mandate and finally confers its legal effects by publishing the reference to it in the Official Journal.

3. The European standardisation request in support of safe and trustworthy AI

European standardisation requests mandating the development of harmonised standards are in principle issued by the Commission following the adoption of the respective legislation. As regards the AI Act, the Commission is looking to speed up the process so as to have standards available by the time the regulation is applicable.[14]

To that aim, and despite the AI Act not having yet been adopted, the Commission has already prepared standardisation request M/593 in support of safe and trustworthy artificial intelligence (European Commission C (2023)3215). This was given a positive position in May 2023 by CEN and CENELEC. As the standardisation request is based on the Commission's proposal, it is expected that a revised request will be adopted after the entry into force of the AI Act to reflect its final text (European Commission C (2023)3215 Annex II:3).

Annexes 1 and 2 of the request provide a list of standardisation deliverables to be developed with requirements regarding the risk management system; data and data governance; record keeping through logging capabilities; transparency and information to users; human oversight; accuracy, robustness and cybersecurity specifications; and quality management systems for providers of AI systems, including a post-market monitoring process.

The request further mandates that:

> … the policy objectives of the Commission in the field of artificial intelligence should be taken into account when drafting European standards and European

13. Case C-613/14, para. 43.

14. Para. 5 of the standardisation request for AI notes that 'to advance technical harmonisation in the field of artificial intelligence and to prepare the necessary technical environment for the implementation of the Artificial Intelligence Act, it is necessary to start the work on European standards and European standardisation deliverables to support the key technical areas covered by the Artificial Intelligence Act proposal. Such standards should include the specifications for the design and the development of AI systems identified as high-risk in the proposal, for AI providers' quality management systems and for the conformity assessment of AI systems.' European Commission (2023), C (2023)3215, p. 2.

> standardisation deliverables in reply to this request. Such policy objectives include ensuring that AI systems placed on the market or put into service in the Union are safe, are used in compliance with fundamental rights and in full respect of Union values (European Commission C (2023)3215: para. 14).

As explained above and in accordance with Regulation 1025/2012, harmonised standards are produced by the recognised European standardisation organisations. However, harmonised standards can effectively be international standards transposed into European standards which are, in turn, referenced in the OJEU. This is particularly important in that most of the standards on AI are being developed at international level, essentially in IEEE and ISO/IEC. Indeed, what changes significantly with AI is that, given the global nature of digitalisation and the emerging digital technologies, international cooperation in standardisation is acknowledged by the Commission and prioritised by industry as a means of remaining competitive in the global market.[15] There is a race to influence international standards on AI because those who own the standards own the technology.

With this in mind, it is worth noting that the standards developed at international level are not required to be aligned with the EU legislation or to support EU values and principles, as mandated in the European Commission's standardisation request. Should the international standards on AI be transposed in Europe, and used to support the AI Act, they pose the risk of falling short of supporting European policy objectives. This concern was underlined in the recent EU strategy on standardisation where it was highlighted that 'in sensitive areas, like lithium batteries, facial recognition or the digital twin, other world regions are taking the lead in international technical committees promoting their technological solutions, which are often incompatible with the EU's values, policies and regulatory framework' (European Commission, COM(2022)31: 6).

Moreover, the standardisation request emphasises the need to have the public interest occupy a prominent executive role and expects that European standardisation organisations will facilitate the effective participation of relevant stakeholders, including but not limited to small and medium-sized enterprises and societal stakeholders, including such as consumer organisations and trade unions (European Commission C (2023)3215, para. 15).

If European standardisation is to live up to the objectives set out in the AI Act and in the underpinning standardisation request, it is essential to ensure the participation of all relevant stakeholders, including trade unions, in the development of those standards. Indeed, standards mandated by the Commission to show compliance with rules imposed in the public interest cannot be left in the hands of technical experts from large companies. Standards called upon to support EU policies and legislation, but where responsibility is left to industry alone to determine, through the ESOs, their technical

15. The risk of losing competitiveness was highlighted in the EU Strategy on standardisation which states: 'In the global race for digital leadership, the ability to shape international standards for digital products, processes and services as global benchmarks is essential for the EU's competitiveness', European Commission COM(2022)31: 1.

content, in practice gives private actors, who have a financial interest in developing the standards, the responsibility to protect the public interest.

The role of the Commission in providing the necessary checks and balances to ensure that harmonised standards are fully in line with Union legislation is fundamental; while ensuring the participation of the public interest in the standardisation activity seems equally important. Considering that standardisation forms part of the overall framework being established in Europe to address AI, this raises the concern around the role trade unions can truly play in the standards setting process.

4. Inclusiveness in the European standardisation system

The growing use in Europe of harmonised standards to support legislation and policy objectives has, over time, raised important questions concerning legitimacy and the ability of the standardisation bodies to ensure the balanced representation of all relevant stakeholders (ETUC 2022: 9). The recent European Strategy on Standardisation even called on the ESOs 'to make proposals [...] addressing uneven and intransparent representation of industrial interests and increasing the involvement of SMEs, civil society and users' (European Commission COM(2022)31: 4).

In Europe, Regulation 1025/2012 on standardisation has recognised the value that societal stakeholders, including trade unions, bring to the development of European standards and sets rules about their participation. The European standardisation organisations are required to encourage and facilitate the appropriate representation and effective participation of all relevant stakeholders (Art. 5) and firm steps have been implemented in this direction. The Regulation further grants the European stakeholder organisations representing consumers, environmental interests, trade unions and SMEs, the so-called Annex III organisations,[16] financial support and formal access to the annual European Union standardisation work programme and the Commission's drafts of standardisation requests, as well as to the work of the technical committees drafting the standards, albeit without voting rights which remain a prerogative of the national delegations (ETUC 2020: 18). Moreover when preparing the deliverables mandated in standardisation requests, the participation of societal stakeholders, including trade unions, should be ensured at all stages; evidence of such involvement is required from the ESOs in the reports they provide to the Commission (European Commission SWD(2015) 205 final: Part 3).

However, and bearing in mind that a significant amount of international standards are likely to be transposed in Europe to support the AI Act, it raises the concern of the potential lack of representation of societal stakeholders in international standard setting activities. This is so especially considering that the measures in terms of facilitating inclusiveness applicable to the international standardisation organisations, mainly ISO

16. Annex III organisations refer to the European societal stakeholders receiving Union financing for standardisation activities and listed under Annex III of the regulation (OJ L 316, 14.11.2012). The Annex identifies four specific categories of stakeholder interests that should be included: SMEs; consumers; environmental; and social interests. The ETUC belongs to the latter category.

and IEC, are different from the European ones in so far as they are not required to comply with Regulation 1025/2012 on standardisation.

Hence one of the prerequisites for the use of standards to support the implementation of the AI Act in Europe should be to ensure that, if standards are developed at international level, the requirements regarding the representation and effective participation of all relevant stakeholders, including trade unions, are fulfilled. From a trade union perspective, it is essential that assessments on the selection and the content of the international standards to be transposed as European standards are conducted in consultation with trade unions. As trade unions were not necessarily involved in the development of the standards at international level, this will contribute to ensuring that they are in line with the AI Act and are in full consistency with existing Union legislation applicable to sectors where high risk AI systems are already used, or likely to be used, such as in employment.

5. Conclusion

The use of algorithms in the world of work must not come at the cost of workers' rights and interests. It is essential to have the right policy settings and a regulatory framework that will not only protect the rights and interests of users but also, and most importantly, the rights and interests of those who are and who will be affected by the use of AI systems, such as workers. The AI Act, as an internal market regulation, fails to address the workplace dimension sufficiently.

The technical guidance provided by the standards supporting the legislation can help in the design and development of AI systems that are in conformance with the requirements, specifications, guidelines and characteristics to ensure that AI technologies and systems meet critical objectives concerning functionality and interoperability. However, standards alone cannot guarantee that the use of AI systems, in particular in the workplace, will promote and protect fundamental rights, including workers' rights, as prescribed by the AI Act. AI is very much context dependent. There is potentially no good or bad AI; it is how it is deployed that will determine whether it poses risks to health, safety and fundamental rights.

That being said, and following the analysis presented in this chapter, there are three basic elements that need to be considered when developing standards in support of the AI Act from a trade union perspective.

First, there needs to be a general understanding of what standardisation, a private and market-driven activity, is meant to do within the overarching framework addressing AI. Technical standards are certainly necessary to set a level playing field in terms of safety and interoperability, but they only provide common technical rules and guidelines for the design, development and safety of AI systems at that level: the standards do not address their implementation, deployment and use in practice. In the workplace, the implementation of AI systems and the data selected to contribute to the systems must be subject to the scrutiny of worker representatives and trade unions through

the existing legal framework on the rights of workers to information, consultation and participation. This will not be achieved through standards.

Second, the participation of all relevant stakeholders in the development of standards which support the legislation is key to ensuring that they take into account the public interest. Considering that the use of AI systems affects – and will increasingly affect – workers and the employment relationship, it should be required that the development of standards to support the AI Act should see the participation of those directly concerned. And although legislation and collective bargaining are, for trade unions, the first and foremost legitimate avenues to address work related aspects, standards play an important role in supporting the legislation for AI and they should therefore be involved in that specific process as well.

Lastly, it is crucial to take into account the international dimension when reflecting on standardisation for AI. Unlike in Europe, where the European standardisation organisations are called upon to ensure the effective participation of societal stakeholders, international standardisation bodies are not required to ensure such representation. Moreover, the standards developed at international level are not particularly intended either to be aligned with EU legislation or to support EU values and principles, and therefore risk falling short in supporting European policy objectives. In this context, it is key to assess the international standards on AI before their transposition in Europe to ascertain that they are in line with European policy objectives. This should ensure that the AI systems placed on the market, or put into service in the Union, are safe, are used in compliance with fundamental rights and in full respect of Union values (European Commission M/593, para. 15).

References

Council of the European Union (1985) Council resolution of 7 May 1985 on a new approach to technical harmonization and standards, Official Journal of the European Union, C 136, 14.6.1985. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31985Y0604(01)

Court of Justice (2017) Judgment of the Court (Third Chamber) of 27 October 2016, James Elliot Construction Limited v Irish Asphalt Limited, (Case C-613/14), Official Journal of the European Union, C 006, 9.1.2017. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62014CA0613

ETUC (2020) The role of international and European standards in shaping the world of work in the European service sector. https://tinyurl.com/4pyxyebm

ETUC (2022) Trade union access to national standardisation committees. https://tinyurl.com/yk93v99f

European Commission (2015a) Independent review of the European standardisation system: Final report - annexes. https://op.europa.eu/en/publication-detail/-/publication/9d26ee6e-c146-4fca-b7d8-c537e0a1965b

European Commission (2015b) Commission staff working document, Vademecum on European standardisation in support of Union legislation and policies, Part III: Guidelines for the execution of standardisation requests, SWD(2015) 205 final, 27.10.2015.

https://single-market-economy.ec.europa.eu/single-market/european-standards/vademecum-european-standardisation_en
European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence, COM(2021) 206 final, 21.04.2021. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence
European Commission (2022a) Commission notice - The 'Blue Guide' on the implementation of EU product rules, Official Journal of the European Union, C 247, 29.06.2022. https://single-market-economy.ec.europa.eu/news/blue-guide-implementation-product-rules-2022-published-2022-06-29_en
European Commission (2022b) An EU strategy on standardisation – Setting global standards in support of a resilient, green and digital EU single market, COM(2022) 31 final, 2.2.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022DC0031
European Commission (2023) Commission implementing Decision of 22.5.2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence, COM(2023) 3215, 22.05.2023. https://ec.europa.eu/growth/tools-databases/enorm/mandate/593_en
European Parliament (2012) Regulation (EU) No 1025/2012 of the European Parliament and of the Council on European standardisation, Official Journal of the European Union, L 316, 14.11.2012. https://eur-lex.europa.eu/eli/reg/2012/1025/oj
WTO (2000) Principles for the development of international standards, guides and recommendations, World Trade Organization. https://www.wto.org/english/tratop_e/tbt_e/principles_standards_tbt_e.htm

All links were checked on 01.02.2024.

Cite this chapter: Giorgi N. (2024) Standardising AI – a trade union perspective, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Part 4

Legal perspectives

Chapter 11
Automated work and workers' rights: platform work and AI work management systems

Mario Guglielmetti

1. Introduction

This chapter discusses the risks to and opportunities for workers stemming from the recent legislative initiatives of the EU in the area of platform work and the use of automated decision-making (ADM) systems, including those using artificial intelligence. In doing this, it points to the solutions envisaged by the co-legislators having regard to the proposal for a directive on improving working conditions in platform work (PWD),[1] looking especially at the report of the European Parliament Employment and Social Affairs Committee.[2] It discusses the coexistence of privacy and data protection on the one hand with labour and social protection, and health and safety objectives, on the other. In relation to both sets of objectives, the PWD must ensure effective oversight and redress for workers. The chapter also considers the regulation of AI work management systems provided by the proposal for an Artificial Intelligence Act (AI Act)[3] in the light of Parliament's draft compromise amendments,[4] identifying some gaps in workers' protection and recommending possible solutions which, at least partly, address this gap.

Automated decision-making is currently applied to work organisation and management in almost every workplace. Personal data is used to enhance the algorithmic systems of work patterns and control ('algorithmic management') (Baiocco and Fernández

1. Proposal for a directive of the European Parliament and of the Council on improving working conditions in platform work, COM(2021) 762 final.
2. Report on the proposal for a directive of the European Parliament and of the Council on improving working conditions in platform work (COM(2021)0762 – C9-0454/2021 – 2021/0414(COD)), Committee on Employment and Social Affairs, Rapporteur: Elisabetta Gualmini, adopted on 21 December 2022. The European Parliament approved its negotiating position on the PWD on 2 February 2023.
3. 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.
4. Draft compromise amendments on the draft report proposal for a regulation of the European Parliament and of the Council on harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 16 May 2023, available at: https://artificialintelligenceact.eu/wp-content/uploads/2023/05/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-16-May-2023.pdf

Macías 2022),[5] to allocate tasks and working time, to establish wages (Dubal 2023a) and to evaluate workers' performance. ADM is thus ubiquitous in working life. This 'datafication' of work is described by some authors as 'techno-normative' control (Griesbach et al. 2019), often based on the 'gamblification' of platform work (Dubal 2023b).[6] The automation of work is indeed grounded on a human workers 'data cycle', described in the following terms: in the first stage, information from workplace and workers is gathered and analysed in real time to create representations of work; in the second stage the information is assessed in accordance with a set of objectives and aligned to standards of performance; in the third stage, interventions are made to seek to change workers' behaviour to ensure standards of performance are met (Gilbert and Thomas 2021).

One of the sectors where algorithmic management is commonly used is platform work, notably in the transport and logistics sectors (Hassel and Özkiziltan 2023). Hence, these two aspects (platformisation and algorithmic management) are almost symbiotic. Interestingly, platform work, as crowdwork,[7] is not only subject to algorithmic management but is also used to train artificial intelligence systems.

2. Automated decision-making in the proposed directive on platform work

This section examines the PWD with an emphasis on the report of the European Parliament (the EP Report), which states that:

> Algorithm-based technologies, including automated monitoring and decision-making systems, have enabled the emergence and growth of digital labour platforms but can produce power imbalances and opacity about decision-making, as well as technology enabled surveillance which could exacerbate discriminatory practices and entail risks for privacy, workers' health and safety and human dignity

5. The Joint Research Centre policy brief provides a description of algorithmic management and of its effects: 'Algorithmic management depends on the collection, transmission and processing of data on the workers and on the economic process. Therefore, algorithmic management relies on several enabling digital technologies that allow for intensive data collection and processing. However, algorithmic management is not linked to a specific technology, but it is better understood as a particular combined use of technologies which are widely available in the digital era' (Baiocco and Fernández Macías 2022: 2). The JRC report goes on to state: 'Algorithmic management can contribute to fissured employment relations. Employment relations can be deteriorated, by resorting to precarious forms of contracts, such as short term or zero hours contracts, especially for more replaceable workers. Also, employment relations can be shifted to market transactions, when the organisation, rather than hiring, 'buys' services from externals (either individual workers or other organisations) for non-core functions. In both cases, labour and social protection can be affected. Ultimately, algorithmic management can undermine labour standards, as it has been argued already in the case of digital labour platforms.' Furthermore, 'Likewise, algorithmic management can have important implications for occupational health and safety, because it intervenes on risk factors that may lead to physical and psychosocial disorders or diseases, such as anxiety, stress, sleep deprivation, depression, musculoskeletal pains, cardiovascular diseases' (Baiocco and Fernández Macías 2022: 4-5).

6. Dubal (2023b) refers to gamification as wage manipulation, the 'gamblification' of wages and labour management via algorithmic wage discrimination.

7. The EP Report introduces, in recital 17c, a definition of crowdwork as 'the organizing of outsourcing or allocation of tasks potentially provided to a large pool of customers or employers, through online platforms.' It also specifies that digital labour platforms organising crowdwork should fall within the scope of the PWD.

> and may lead to adverse consequences for working conditions and the exploitation of workers. (Recital 4, as amended)

The PWD aims at addressing the increasing power and information asymmetry between the digital employer and the worker.[8] Other key issues addressed by the PWD which are generally outside the focus of this chapter, are: the correct determination of the employment status of the person carrying out platform work; and ensuring collective bargaining and workers' representatives role in the context of platform work. Having regard to the first issue, the PWD establishes, under Article 4, a legal presumption of an employment relationship. Moreover, it lays down measures aiming at ensuring the effective implementation of this, including strengthening controls and cooperation between different national authorities as well as – according to the amendments proposed in the EP Report – measures to avoid the circumvention of the safeguards established under the PWD in relation to subcontracting.[9] Concerning the second issue, the PWD provides among others that digital labour platforms should not only ensure human oversight of automated decision-making but also evaluate its impact on working conditions, health and safety, and fundamental rights and freedoms including dignity, together with workers' representatives. Notably, the EP Report introduces a new Article (10a) on collective bargaining in platform work, which encompasses bargaining on the features of automated monitoring and decision-making systems to improve working conditions.

In order to tackle the issue of the incorrect determination of employment status, but also to remedy the power and information asymmetry between the platform and the worker, the EP Report provides the obligation for Member States to determine a national target for the number of inspections to be carried out and to ensure adequate powers for the appropriate authorities to carry out these inspections, including the provision of sufficient staff with the skills and qualifications required.[10]

One important aspect of algorithmic management is the processing of personal data related to workers. Right from the initial step of workers' identification, in relation to which the EP Report specifies that employers should always provide workers with identification methods less intrusive than biometric identification,[11] up to the monitoring of workers' performance, personal data related to identifiable persons is processed. The issue of increased levels of worker surveillance (see Ponce Del Castillo and Molè, and Gould; both in this volume) is also a focus of the PWD, and this has also been reported in the media as a worrying trend (Barbaro 2022). The issue of workplace surveillance is connected to the problem of the definition of the so-called 'data perimeter', understood as the types of personal data that should not be processed by the employer due, for instance, to the risk of damaging the dignity of the worker. The processing of personal data in the context of the PWD falls under the scope of the

---

8. See recital 8, as amended by the EP Report: 'persons performing platform work subject to such algorithmic management often do not have information on how the algorithms work, which personal data are being used and how their behaviour affects decisions taken by automated systems.'

9. See Article 12b, Subcontracting liability; as well as recital 26, as amended by the EP Report.

10. See Article 4(3)(d), as amended by the EP Report.

11. See Article 6(5) point d a (new).

General Data Protection Regulation (GDPR).[12] In this regard, it is worth remarking that the GDPR does not preclude the establishment of specific, context-related, safeguards for the persons concerned. This reasoning is even more applicable in the workplace since Article 88 GDPR expressly lays down the possibility for Member States to provide more specific rules to ensure the protection of workers' rights and freedoms in respect of the processing of their personal data.

Article 6(5) of the proposed PWD provides that digital labour platforms must not process any personal data concerning platform workers that are not intrinsically connected to and strictly necessary for the performance of the contract between the worker and the platform. It also specifies certain categories of personal data which must not be processed, namely: (a) any personal data on the emotional or the psychological state of the worker; (b) any personal data relating to the health of the worker, except in cases referred to in Article 9(2), points (b) to (j), of the GDPR; (c) any personal data in relation to private conversations, including exchanges with workers' representatives; and (d) any personal data in relation to the time when the worker is not offering or performing platform work.

The EP Report prohibits the processing of personal data inferring the emotional and psychological state of the worker; personal data revealing racial or ethnic origin, migration status, political opinions, religious or philosophical beliefs, disability or state of health, including chronic disease or HIV status, or trade union membership; genetic and biometric data for the purpose of uniquely identifying a person; and data concerning a person's sex life or sexual orientation.[13]

This is welcome since it counters the trend of increased, continuous and invasive surveillance at the workplace and aims at fulfilling the requirements of necessity and proportionality concerning the processing of personal data in a 'horizontal' (business to citizen) dimension;[14] and, ultimately, to preserve the worker's dignity, specifying these requirements in a way that is easy to operationalise.

The PWD, as amended, also requires continuous assessment of automated decision-making: digital labour platforms, with the involvement of workers' representatives, must carry out an assessment, regularly and at least annually, of the impact of the individual decisions taken or supported by automated monitoring and decision-making systems on working conditions, health and safety, and fundamental rights.[15] Moreover, digital labour platforms must provide for the human oversight of all decisions affecting working conditions.[16] The EP Report introduces, in addition to this broader assessment,

12. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4 May 2016, pp. 1-88.

13. Article 6(5), point c a (new).

14. As opposed to the 'vertical' (state to citizen) dimension, in relation to which necessity and proportionality of the processing of personal data have been assessed by the Court of Justice of the European Union in several judgments.

15. Article 7(1), as amended by the EP Report.

16. Article 7(1) (new), as introduced by the EP Report.

an obligation for digital labour platforms to perform a data protection impact assessment (DPIA), here also with the involvement of those affected by the processing of personal data.[17]

When it comes to automated decision-making in the context of platform work, similarly to cases of algorithmic decision-making in other regulated activities, such as ADM for online content recommendation[18] or for consumer credit decisions,[19] considerations related to the protection of privacy and personal data coexist with aspects related to compliance with sectoral law (concerning platform work: health and safety, non-discrimination and working time, among others) which fall within the oversight of labour inspectorates. It is therefore essential to consider both sets of – mutually reinforcing – objectives.

From a data protection perspective, the GDPR provides important safeguards for workers since it empowers them, as subjects of the processing of personal data, concerning the right to access their personal data, and to rectify and have these erased, as well as the right to data portability and to be informed about the processing of personal data.

Moreover, Article 22 GDPR provides that, where ADM and profiling is allowed, the data controller (in this case, the digital labour platform) must implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, including at least the right to obtain human intervention, to express his or her point of view and to contest the decision.[20] According to Article 13(2)(f) and 14(2)(g) of the GDPR, the controller must provide the data subject with information on the existence of automated decision-making, including profiling, and meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. In this regard, the provisions of the GDPR on ADM and profiling are functional to allowing the data subject some control over the decision-making processes that significantly affect him or her (Bygrave 2020).

Meanwhile, from a sectoral law perspective, transparency on ADM is key to the control of the impact on platform workers' working conditions and of the compliance of such systems with national law or practice, including with regard to the role of data protection authorities (DPAs), and applicable collective agreements. At the same time, the auditing of the functioning of ADM can provide useful information on the degree of autonomy of the worker, potentially misclassified as independent.

---

17. Article 6(5a) (new).

18. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277, 27 October 2022, pp. 1-102.

19. Proposal for a directive of the European Parliament and of the Council on consumer credits, COM/2021/347 final.

20. See Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679, adopted on 3 October 2017, and as last revised and adopted on 6 February 2018, available at: https://ec.europa.eu/newsroom/article29/items/612053

The rationale underpinning the provisions of sectoral law (in this case, the PWD) on ADM transparency is strengthened as a result of the EP Report.[21] Digital labour platforms must provide the platform worker with a written statement of the reasons for any decision supported by an automated decision-making system to restrict access to work assignments or to restrict, suspend or terminate a platform worker's account; any decision to refuse remuneration for work provided by the platform worker; any decision on the platform worker's contractual status; and any decision producing an effect on the agreed terms of the employment relationship or which has similar effects. This statement of reasons has the function of allowing the worker and the administrative or judicial authorities to have control over the compliance of a decision supported by ADM with national law or practice and applicable collective agreements.

When it comes to the interface between the GDPR and sectoral law regulating the activity in which ADM is being used, the policy objective of the GDPR can contribute to ensure compliance with the rules-of-the-art applicable to the business activity where ADM is deployed, but it is not sufficient (Matsumi and Solove 2023; Abraha 2023; Kelly-Lyth and Thomas 2023).[22]

The correct allocation of competences between the sectoral authorities and the GDPR supervisory authorities, in this case the DPAs and the labour authorities, is neither obvious nor easy. Nonetheless, it is a key aspect of the regulatory framework. Regarding ADM and work management regulated under the PWD, there are significant limits to DPAs' possible scope of action.

These limits become particularly evident with reference to the specific provisions of the PWD: a DPA – competent for the enforcement of privacy and data protection violations – cannot assess the overall impact of ADM on workers' working conditions, for instance whether these comply with maximum working time or minimum wages, or occupational safety and health standards; nor can DPAs order businesses to replace or correct ADM to assert the exact decision, work patterns, etc. Furthermore, a DPA cannot issue decisions on access to work assignments, workers' earnings or their occupational safety and health, working time, promotion or their contractual status, including the restriction, suspension or termination of their account for work-related reasons. Finally a DPA cannot even fully assess the grounds for decisions to restrict,

21. Article 8(1) subparagraph 2.

22. On the conditions and limits of GDPR as a safeguard in algorithmic decision-making, see Matsumi and Solove (2023). More specifically, in relation to algorithmic work management, Halefom Abraha comments thus: 'The enforcement of data protection rules at work falls under the regulatory remits of DPAs, who are not labour experts. Multiple reports show that DPAs are under-resourced and understaffed. Compounding the lack of resources and expertise is the lack of interest on the part of DPAs to prioritise data protection in the employment context' (Abraha 2023: p. 186). Aislinn Kelly-Lyth and Anna Thomas provide specifications on the interplay between algorithmic risk and impact assessments (ARIAs) and DPIA: 'Although data protection can operate as a gateway to access other rights and freedoms relevant to algorithmic management, the data protection regime rests on assumed human ability to control and manage information (human 'sovereignty' over data), an idea that has been challenged by the latest wave of algorithmic management tools. Further, there will be some situations in which tools used to inform the managerial prerogative do not require a DPIA simply because they do not process personal data. For example, where non-personal (anonymised) data on supply chain efficiency leads to an entire team being relocated, there may be no DPIA obligation on the employer – even though the team move is an exercise of the managerial prerogative with impacts on workers.' (Kelly-Lyth and Anna Thomas (2023: p. 248).

suspend or terminate a worker's account, or to refuse remuneration for work done by the worker, or on the worker's contractual status and, therefore, it cannot review such decisions and rectify them.

Moreover, a DPA is not in a position to monitor that the person charged by the employer with responsibility as a data controller has the necessary competence, training and authority to monitor the overall and granular impact of ADM on working conditions. Neither can it ensure that the employer provides the worker with a written statement of the reasons for any decisions taken or supported by an automated decision-making system in these areas and which is fully meaningful and respondent to labour law obligations (Dubal 2023b; Fink and Finck 2022).[23]

In contrast, the EP Report aims to enhance cooperation between DPAs and other competent authorities. In Article 7, paragraph 3 a (new), the text specifies that, when an impact assessment (on working conditions, health and safety and fundamental rights, to be submitted to DPAs and labour authorities as well as to workers' representatives), is found to be non-compliant with Article 7(1), the health and safety, data protection, labour and other competent authorities shall take coordinated measures to enforce those provisions.

The EP Report also introduces provisions on cooperation between labour, social protection and tax authorities in cross-border cases.[24] Taking into account these considerations on the limits of DPAs' scope of action, the EP Report appropriately adds the wording 'together with national labour authorities' in Article 19(1) on the supervision of the compliance by digital labour platforms with the provisions of Article 6, 7(1) and (3), 8, 10 and 15 of the PWD. According to recital 48, since ADM in the context of platform work raises issues of data protection as well as labour and social protection law, DPAs and the relevant labour and social protection authorities should cooperate, including at cross-border level, in the enforcement of the PWD, including by exchanging relevant information with each other.

---

23. As Dubal comments: 'For example, through a GDPR data request, Worker Info Exchange succeeded in gaining access to data collected by Amazon, as well as a guidance document from Amazon Flex. Nevertheless, this knowledge has not ended digitalized variable pay or control for DSPs in Europe. In other words, firm transparency or a worker right to algorithmic explainability — while crucial to understanding the logic of existing practices — does not by itself shift the power dynamics that enable algorithmic wage discrimination. Nor does it do much to mitigate the culture of labor gamblification described in Part II that is becoming endemic to the on-demand economy – and to more conventional workplaces' (Dubal 2023b: 47). Relying on 'GDPR only' (namely, providing DPAs with exclusive competence in the area of ADM for work management) would be detrimental to the effective protection of labour, social protection, health and safety rights provided to workers under the EU acquis: it might render the enforcement of these rights and safeguards more difficult in practice. Moreover, it is also clear that, in the absence of specific provisions on supervision by DPAs in the proposal for PWD, the GDPR applies in any case to all processing of personal data in the context of the PWD. Having regard to the use of AI systems for ADM, Fink and Finck observe: 'it is paradoxical that discussions around the explainability of AI have focused almost exclusively on data protection law, neglecting not only obligations in administrative law, but also other areas of EU law where similar obligations exist, such as public procurement law, consumer protection law, and financial regulation. The acknowledgement that explanation obligations already exist in other areas of EU law is important more generally, especially in the context of claims that EU data protection law should 'introduce' explanation requirements. Only then can the interplay between general and sectoral requirements, as well as the advantages of one versus the other, be properly evaluated' (Fink and Finck 2022: 389).

24. Article 12 a (new).

Furthermore, the new recital 48a specifies that platform workers should have meaningful access to reporting and redress mechanisms with the relevant national authority, be it the DPA or the labour inspectorate. They should also be able to report possible infringements of the PWD and have the right to be heard and to be informed about the outcome of their complaint, in addition to the right to a timely decision.

These amendments are key to addressing the limits of the legislation in terms of the supervision by DPAs (Article 19 of the PWD proposal) as well as the right to redress (Article 13) and the procedures on behalf or in support of workers engaged in platform work (Article 14). These provisions would, however, jeopardise the enforcement of the PWD and workers' rights if they were adopted in such a way that establishes the exclusive competence of DPAs as supervisory authorities and of the GDPR in terms of the right of redress and to representative actions under the PWD. The amendments in the EP Report on concurrent supervision (i.e. cooperation between the authorities as regards oversight) and the cumulative applicability of GDPR and labour law forms of redress are steps in the right direction since they would address these specific gaps. However, these aspects should be further specified and clarified in the enacting terms of the PWD.[25]

3. The AI Act as regulation: CE marking for AI work management systems

The proposed AI Act regulates as 'high risk'[26] certain artificial intelligence systems for work management. Listed among high-risk AI systems in Annex III, point 4, is:

> Employment, workers management and access to self-employment: (a) AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests; (b) AI intended to be used for making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.

25. Having regard to the proposal for PWD, Aída Ponce Del Castillo and Diego Naranjo note: 'For instance, the obligations established by Articles 6, 7(1) and (3), 8 and 10 fall under the competence of national DPAs, but we believe they should fall under the competence of labour authorities' (Ponce Del Castillo and Naranjo 2022: 6).

26. Briefly, the AI Act distinguishes four categories of different risk levels regarding AI systems: (a) unacceptable risk; (b) high risk; (c) limited or minimal risk; (d) low risk. Systems with unacceptable risk and hence prohibited, are, except for specific purposes and where accompanied by prior authorisation: AI systems using subliminal techniques or exploiting vulnerabilities causing physical or psychological harm; for social scoring; and real-time remote biometric identification systems in publicly available spaces for the purpose of law enforcement. High-risk AI systems are included as two sub-categories of AI systems: first, AI systems that are safety components of products already covered by certain Union health and safety harmonisation legislation (such as toys, machinery, lifts or medical devices); second, 'stand-alone' AI systems ('AI for services') specified in Annex III for use in eight areas: biometric identification and categorisation of natural persons; the management and operation of critical infrastructure; educational and vocational training; employment, worker management and access to self-employment; access to and enjoyment of essential private services and public services and benefits; law enforcement; migration, asylum and border control management; and the administration of justice and the democratic process.

The essential requirements that the AI Act establishes for high-risk systems relate in particular to training, validation and the testing of data sets; record-keeping; providing information to the users of AI systems, including on the intended purpose and level of accuracy; human oversight; robustness; and security.

Providers[27] of these systems must conduct a conformity assessment, draw up an EU declaration of conformity and affix a CE marking. In the case of 'AI for services', this is a self-assessment control procedure which allows the AI system to be placed on the market and put into service, and then to move freely within the internal market. As observed by some authors (Ebers 2022; Veale and Zuiderveen Borgesius 2021), the AI Act builds on the legal framework for the safety of products.

It is important to note that, as a rule, an AI system which is in conformity with standards – once such standards have been issued – will be considered as being in conformity with the requirements for high-risk AI.[28] However, it is unclear whether and how compliance with such requirements would ensure that the AI work management system is aligned to the EU acquis and to national labour law (protection from dismissal, access to the minimum wage, maximum working hours, health and safety) and with the Charter of

27. A 'provider' of an AI system is defined in the AI Act (Article 3, definition (2)) as 'a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge.'

28. Carlo Colombo and Mariolina Eliantonio observe that 'new governance forms, of which standardization constitutes a pre-eminent example, have much to offer and are indeed essential in an era of framework norms. Standardization has proved an effective market integration tool, which has served to overcome technical barriers to trade when political agreement on these issues seems unattainable and it is a system which is able to keep pace with the fast and complex technological and scientific changes of our current society. However, we cannot overlook that this peculiar regulatory structure, operating 'in the shadow of hierarchy', gives rise to a form of complex normativity that combines hard and soft law instruments, together with European and national regulatory levels, in a way that challenges the essence of EU law. It is indeed frequently the case that these governance forms cut across established categories of public law, making their essential nature difficult to capture or distil. Looking back at our point of departure, and attempting an evaluation of the overall legitimacy of the standardization process, we can safely conclude that there is still a long way to go before we can speak of a fully legitimate system. This is because, in the current system, ex post legitimacy is not ensured: standards seem not to be judicially reviewable at EU level, neither directly nor indirectly, by affected persons, thus somewhat weakening the catalyst function that has allowed courts to address the challenges of other instances of new governance mechanisms. James Elliot has, from this perspective, closed a door to this possibility. In addition, the lack of judicial control is not compensated by a sufficient degree of ex ante legitimacy: participation by societal stakeholders only seems to work on paper, while the reality depicts a much more "elitist" system, in which consumer or environmental interests hardly have a voice. Similarly, while the Commission's control over the process seems to resemble a "paper tiger", safeguard measures are not at the disposal of affected persons.' (Colombo and Eliantonio 2017: 340). See also Martin Ebers: 'Conclusions: The proposed rules of the AIA for high-risk systems raise serious concerns. For these systems, the European Commission primarily wants to rely on an ex ante conformity assessment, which is not carried out by external third parties, but by the companies themselves – combined with the presumption of conformity, if the provider follows harmonized standards, which are to be developed by ESOs in accordance with the NLF. However, ESOs are clearly overburdened by this task. The standardization of AI systems is not a matter of purely technical decisions. Rather, a series of ethical and legal decisions must be made, which cannot be outsourced to private SDOs, but which require a political debate involving society as a whole' (Ebers 2022). Moreover, Ebers observes, especially having regard to standards of AI systems for services, that 'it is difficult to separate technical from political aspects. Issues of fairness, the acceptable level of accuracy, transparency of the systems: these are also political aspects. Moreover, standards are closed-access, subject to copyright; and despite the societal impact of standards, civil society and impacted communities cannot easily engage in their drafting. Therefore, on the one hand, standards are of course a factor of legal certainty and progress for industry; on the other hand, they might not always be fit for purpose.' The AI Act itself, according to Ebers, should specify for instance what types of bias are prohibited and how and to what extent they should be mitigated (what is the 'acceptable bias').

Fundamental Rights of the European Union.[29] It is also unclear whether, as an outcome of this certification process, 'the core sphere of privacy', and ultimately workers' dignity, would be adequately protected once the AI system has been placed on the market and put into service.

Moreover, the algorithmic calculation of platform workers' wages based on AI systems that include dynamic pricing, surge pricing and bidding systems which pick up the lowest wage/availability (Griesbach et al. 2019: 5) could, in most cases, be in violation of EU and Member State legislation on adequate minimum wages. Nonetheless, AI work management systems are not assessed in the context of certification in this area under the AI Act. Additionally, it is unclear if and how the requirements for algorithmic work management under the PWD (for instance the prohibition on the processing of data which seeks to infer the emotional and psychological state of the worker) would be taken into account in the declaration of the conformity of AI work management systems since the AI Act does not contain a prohibition of emotion recognition systems.[30]

In addition, the absence of independent, third party audits[31] of the AI work management system is a factor that will not lead to an increase in the level of trust in AI systems by users and workers.

More broadly, as a result of all these factors, the certification of work management systems under the AI Act will not, unless it is integrated with the sectoral law

29. See Article 31, Fair and just working conditions: 1. Every worker has the right to working conditions which respect his or her health, safety and dignity. 2. Every worker has the right to limitation of maximum working hours, to daily and weekly rest periods and to an annual period of paid leave.

30. Under the proposed AI Act, emotion recognition systems are not prohibited and are not considered high-risk AI systems per se. They are, however, subject to the transparency obligations under Article 52 according to which users of an emotion recognition system must inform those who are exposed to it about its use.

31. Authors have pointed to the need for external, third party audits in advance of the CE marking of high-risk systems. As Mauritz Kop comments: 'Self-assessment too non-committal (non-binding)? First, it is crucial that certification bodies and notified bodies are independent and that no conflicts of interest arise due to a financial or political interest. In this regard, I wrote elsewhere that the EU should be inspired by the modus operandi of the US FDA. Second, the extent to which companies can achieve compliance with this new AI 'product safety regime' through risk-based self-assessment and self-certification, without third party notified bodies, determines the effect of the Regulation on business practices and thus on the preservation and reinforcement of our values. Internally audited self-assessment is too non-committal given the high risks involved. Therefore, I think it is important that the final version of the EU AI Act subjects all high-risk systems to external, independent third party assessment requirements. Self-regulation in combination with awareness of the risks via (voluntary or mandatory) internal AI impact assessments is not enough to protect our societal values, since companies have completely different incentives for promoting social good and pursuing social welfare, than the state. We need mandatory third party audits for all High-Risk AI Systems.' (Kop 2021: 8). In the absence of a clear definition of 'work management system' in the AI Act, there is also some lack of clarity about what exactly is going to be CE-marked and against which parameters.

requirements,[32] provide the necessary trust in the conformity of the use of such systems with the PWD or with the EU acquis on labour and social protection law.

This 'protection gap' might become more apparent in terms of the outcomes of either an assessment by the digital labour platform (as the user)[33] of the impact of an AI work management system on working conditions, including the health and safety and labour and social protection law requirement introduced under Article 7 paragraph 3 a (new) of the EP Report, or the DPIA to be performed pursuant to Article 6(5a) (new). In this same regard, it is also notable that the EP Report introduces an obligation for the digital labour platform immediately to cease use of a system when the impact assessment to be performed under the PWD finds risks to health and safety or the fundamental rights of workers that cannot be avoided or mitigated.[34]

4. The European Parliament's draft compromise amendments to the AI Act: worker-related changes

The European Parliament's draft compromise amendments introduce recitals and modifications to specific articles that recognise the need to respect workers' rights.

Notably, they provide for the prohibition of emotion recognition systems in the workplace, adding to Article 5 of the proposed AI Act a prohibition of 'the placing on the market, putting into service or use of AI systems to infer emotions of a natural person in the areas of law enforcement, border management, in workplace and education

32. In the AI Act, the certification of AI work management systems does not cover compliance with the labour law acquis nor oversight by labour inspectorates. In contrast, for some financial services, reference to sectoral law and to its oversight (by financial oversight authorities) is provided in the AI Act. As regards financial institutions, the Council compromise text on the AI Act specifies that, for providers that are financial institutions subject to requirements regarding their internal governance, arrangements or processes under Union financial services legislation, the obligation to put in place a quality management system is considered to be fulfilled by complying with the rules on internal governance arrangements or processes pursuant to the relevant Union financial services legislation (Article 17(3)). Moreover, Article 63(4) lays down that, for high-risk AI systems placed on the market, put into service or used by financial institutions regulated by Union legislation on financial services, the appropriate market surveillance authority is the national authority responsible for the financial supervision of those institutions under that legislation in so far as placement on the market, putting into service or the use of the AI system is in direct connection with the provision of those financial services.

33. The 'user' of an AI system is defined in the AI Act (Article 3, definition (4)) as 'any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity.'

34. Article 7, paragraph 2 b (new). These important safeguards would, however, only apply when the AI work management system falls under the scope of the PWD.

institutions.'[35] This would ensure alignment between the provisions of the PWD and of the AI Act on the prohibited use of certain types of data.

Furthermore, the draft compromise amendments require that 'prior to putting into service or use a high-risk AI system at the workplace, deployers shall consult workers representatives with a view to reaching an agreement and inform the affected employees that they will be subject to the system' [sic].[36] This is welcome since it introduces the right for workers' representatives to be 'at the table' when AI management systems are introduced, in line with the amendments proposed in the EP Report on the PWD according to which digital labour platforms, with the involvement of workers' representatives, must conduct an assessment, regularly and at least annually, of the impact of the individual decisions taken or supported by automated monitoring and decision-making systems on working conditions, health and safety, and fundamental rights.[37]

Another welcome amendment to the AI Act specifies that it does not preclude Member States or the Union from maintaining or introducing laws, regulations or administrative provisions which are more favourable to workers in terms of protecting their rights concerning the use of AI systems by employers, or from encouraging or allowing the application of collective agreements which are more favourable to workers.[38]

Finally, recital 61 of the AI Act, as modified by the EP draft compromise amendments, states that, when AI systems are intended to be used in the workplace, harmonised standards should be limited to technical specifications and procedures. Since it seems that standards as developed today are not restricted to technical specifications (see Giogi, this volume), it would be difficult to reconcile this specification with current standard-setting as provided by the proposed AI Act.

---

35. Article 5 (dc). See also recital 26c: 'There are serious concerns about the scientific basis of AI systems aiming to detect emotions, physical or physiological features such as facial expressions, movements, pulse frequency or voice. Emotions or expressions of emotions and perceptions thereof vary considerably across cultures and situations, and even within a single individual. Among the key shortcomings of such technologies are the limited reliability (emotion categories are neither reliably expressed through, nor unequivocally associated with, a common set of physical or physiological movements), the lack of specificity (physical or physiological expressions do not perfectly match emotion categories) and the limited generalisability (the effects of context and culture are not sufficiently considered). Reliability issues and, consequently, major risks for abuse may especially arise when deploying the system in real-life situations related to law enforcement, border management, workplace and education institutions. Therefore, the placing on the market, putting into service, or use of AI systems intended to be used in these contexts to detect the emotional state of individuals should be prohibited.'

36. Article 29, 5a.

37. Article 7(1), as amended by the EP Report.

38. Article 2, 5c. See also recital 2d: 'In line with Article 114(2) TFEU, this Regulation complements and should not undermine the rights and interests of employed persons. This Regulation should therefore not affect Community law on social policy and national labour law and practice, that is any legal and contractual provision concerning employment conditions, working conditions, including health and safety at work and the relationship between employers and workers, including information, consultation and participation. This Regulation should not affect the exercise of fundamental rights as recognized in the Member States and at Union level, including the right or freedom to strike or to take other action covered by the specific industrial relations systems in Member States, in accordance with national law and/or practice. Nor should it affect concertation practices, the right to negotiate, to conclude and enforce collective agreement or to take collective action in accordance with national law and/or practice. It should in any case not prevent the Commission from proposing specific legislation on the rights and freedoms of workers affected by AI systems.' See also recital 36 which, as amended, concludes: 'This Regulation applies without prejudice to Union and Member State competences to provide for more specific rules for the use of AI-systems in the employment context.'

5. Conclusions

The draft directive on platform work, notably as amended by the EP Report, represents a significant advance towards regulating automated work in a way that avoids a 'race to the bottom' regarding working conditions. It provides a legal framework that establishes a level playing field for employers and which respects the EU acquis on labour and social protection and health and safety at work, and enhances the protection of personal data provided by the GDPR. Notably, it clearly defines prohibitions on the processing of certain personal data in the specific context of work.

Although the AI Act, which encompasses the regulation of high-risk AI, including artificial intelligence systems for work management, does not explicitly refer to the requirements outlined in sector-specific labour laws, such as the PWD, there is a clear interface between the AI Act and the PWD. Unless integrated with these requirements, however, the certification of work management systems under the AI Act would not provide trust on the conformity of the use of such systems with the PWD or with the EU acquis on labour and social protection law. This would hardly be considered a success story for the CE marking regime.

At the same time, the EP draft compromise amendments to the proposed AI Act would increase the level of protection for workers, notably stating a prohibition on the use of emotion recognition systems and by providing an obligation on the deployers of AI systems in the workplace to consult workers' representatives. Nevertheless, as is evident, the level of protection for workers, as well as the level of legal certainty and consistency between the PWD and the AI Act, may vary significantly depending on the final text of both legal instruments.

References

Abraha H. (2023) Regulating algorithmic employment decisions through data protection law, European Labour Law Journal, 14 (2), 172–191. https://doi.org/10.1177/20319525231167317

Baiocco S. and Fernández Macías E. (2022) The algorithmic management of work: A basic compass, European Commission. https://publications.jrc.ec.europa.eu/repository/handle/JRC129819

Barbaro M. (2022) The rise of workplace surveillance. Is your productivity being electronically monitored by your bosses?, New York Times, 24 August 2022. https://www.nytimes.com/2022/08/24/podcasts/the-daily/workplace-surveillance-productivity-tracking.html

Bygrave L.A. (2020) Article 22. Automated individual decision-making, including profiling, in Kuner C., Bygrave L.A. and Docksey C. (eds.) The EU General Data Protection Regulation (GDPR): A commentary, Oxford University Press, 522–542. https://doi.org/10.1093/oso/9780198826491.003.0055

Colombo C. and Eliantonio M. (2017) Harmonized technical standards as part of EU law: Juridification with a number of unresolved legitimacy concerns? Case C-613/14 James Elliot

Construction Limited v. Irish Asphalt Limited, EU:C:2016:821, Maastricht Journal of European and Comparative Law, 24 (2), 323–340. https://journals.sagepub.com/doi/full/10.1177/1023263X17709753

Dubal V. (2023a) On algorithmic wage discrimination, UC San Francisco Research Paper. http://dx.doi.org/10.2139/ssrn.4331080

Dubal V. (2023b) The house always wins: The algorithmic gamblification of work, LPE Project. https://lpeproject.org/blog/the-house-always-wins-the-algorithmic-gamblification-of-work

Ebers M. (2022) Standardizing artificial intelligence. A critical assessment of the European Commission's proposal for an artificial intelligence act, Robotics and AI Law Society. https://blog.ai-laws.org/standardizing-artificial-intelligence/

Fink M. and Finck M. (2022) Reasoned A(I)dministration: Explanation requirements in EU law and the automation of public administration, European Law Review, 47 (3), 376–392. https://hdl.handle.net/1887/3439725

Gilbert A. and Thomas A. (2021) The Amazonian era: How algorithmic systems are eroding good work, Institute for the Future of Work. https://www.ifow.org/knowledge-hub-items/the-amazonian-era-how-algorithmic-systems-are-eroding-good-work

Griesbach K., Reich A., Elliott-Negri L. and Milkman R. (2019) Algorithmic control in platform food delivery work, Socius, 5. https://doi.org/10.1177/2378023119870041

Hassel A. and Özkiziltan D. (2023) Governing the work-related risk of AI: Implications for the German government and trade unions, Transfer, 29 (1), 71–86. https://doi.org/10.1177/10242589221147228

Kelly-Lyth A. and Thomas A. (2023) Algorithmic management: Assessing the impacts of AI at work, European Labour Law Journal, 14 (2), 230–252. https://doi.org/10.1177/20319525231167478

Kop M. (2021) EU artificial intelligence act: The European approach to AI, Transatlantic Antitrust and IPR Developments, Stanford Law School. https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/

Matsumi H. and Solove D.J. (2023) The prediction society: Algorithms and the problems of forecasting the future, GWU Legal Studies Research Paper 2023-58, George Washington University Law School. http://dx.doi.org/10.2139/ssrn.4453869

Ponce del Castillo A. and Naranjo D. (2022) Regulating algorithmic management: An assessment of the EC's draft Directive on improving working conditions in platform work, Policy Brief 2022.08, ETUI. https://www.etui.org/publications/regulating-algorithmic-management

Veale M. and Zuiderveen Borgesius F. (2021) Demystifying the draft EU artificial intelligence act: Analysing the good, the bad, and the unclear elements of the proposed approach, Computer Law Review International, 22 (4), 97–112. https://doi.org/10.9785/cri-2021-220402

All links were checked on 14.02.2024.

Cite this chapter: Guglielmetti M. (2024) Automated work and workers' rights: platform work and AI work management systems, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 12
Automating employment: a taxonomy of the key legal issues and the question of liability

Teresa Rodríguez de las Heras Ballell

1. Setting the scene. AI and employment: a proposal for a taxonomy

Automation is probably the hallmark of our time. Automation is pervading society and extensively penetrating business activity and economic relationships, encouraged by numerous expected benefits and perceived efficiencies. It arguably does provide efficiency, dramatically reduce transaction costs, optimise processes, enable the efficient processing of an enormous amount of data and assist decision-making in complex and uncertain contexts.

Within this trend, automation has also penetrated the workplace and employment relationships. Algorithmic and AI-driven systems are extensively and increasingly employed in recruiting, promoting, monitoring or evaluating employees' work performance, planning and allocating assignments and, in a variety of industrial activities, developing mechanical procedures and conducting predictive and calculation-based tasks in the workplace.

Several legal issues arise from the intensive and extensive use of automation for such a variety of purposes related to employment, and a cascade of legitimate legal questions has been triggered. Is automation licit in employment? Should the use of AI be acknowledged and permitted by the law? Are automated decisions valid and enforceable, even for decisions susceptible to having an impact on workers' rights? To whom should the legal effects of such automated actions and decisions be attributed? Who is responsible for monitoring the outputs or mitigating the biases of the AI system? And who is liable for the damage caused by AI in the workplace or when used for employment purposes?

1.1 Assembling the pieces of a legal framework governing the use of AI in employment

A complete, consistent and solid legal regime providing for rules governing the use of automation in employment-related contexts is lacking. Some rules, scattered in various pieces of legislation, do provide guidance on the construction of a reasonable legal framework. But the effort to fill the gaps, infer principles, accommodate general rules to the employment context and even propose new provisions has still to be made to establish a sound and predictable body of rules governing the use of AI in employment.

Thus, the EU's future AI Act[1] does not provide a complete body of such rules and only incidentally refers to employment in qualifying certain uses as 'high-risk' AI systems. At the core of the AI Act lies a risk-based model that classifies AI systems on the basis of intended use (unacceptable risk/high risk/limited or minimal risk/low risk). The list of high-risk AI systems includes AI for work management.[2] Nevertheless, the sectoral law requirements applicable to such systems, even if they are harmonised at EU level, are not incorporated in the certification process, based as it is on self-assessed conformity with EU rules and CE marking. AI systems classified as high-risk are permitted on the European market subject to compliance with certain mandatory requirements in relation to data and data governance, documentation and record keeping, transparency and the provision of information to users, human oversight, robustness, accuracy and security, as well as to the ex ante conformity assessment.

The proposed directive on platform work explicitly addresses algorithmic management, providing for seemingly attractive rules and solutions, but the scope of application seems to be restricted to platform work[3] whereas AI is intensely used in a multitude of employment contexts that do not qualify as platform work and, indeed, in those where no platforms are involved at all. The scope of application of the proposed directive is defined by the concept of platform work – that is, as Recital 5 clarifies, work performed by individuals that, through the infrastructure of digital labour platforms, provides a service to customers. Nonetheless, it is not completely clear whether platform workers are all workers working for a platform or only those whose work 'is organised through a digital labour platform'. Thus, while it is evident that the work done by drivers or couriers for popular platforms is defined by the proposed directive as platform work, the definition of other tasks and job positions related to the running of such platforms is rather ambiguous and uncertain (regarding, for example, workers in warehouses or in other corporate departments). The connection with algorithmic management seems to invite the consideration of a narrower definition of platform work which excludes the latter.

1. 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.

2. As per the drafting of the text of 14 June 2023 – P9_TA(2023)0236 Artificial Intelligence Act. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1 (Ordinary legislative procedure: first reading), Annex III, paragraph 1, point 4.b): 'AI systems intended to be used to make or materially influence decisions affecting the initiation, promotion and termination of work-related contractual relationships, task allocation based on individual behaviour or personal traits or characteristics, or for monitoring and evaluating performance and behavior of persons in such relationships'.

3. Pursuant to Article 2(1) of the directive on platform work, the provisions apply to work organised through 'digital labour platforms' which are defined as follows: 'digital labour platform' means any natural or legal person providing a commercial service which meets all of the following requirements: (a) it is provided, at least in part, at a distance through electronic means, such as a website or a mobile application; (b) it is provided at the request of a recipient of the service; (c) it involves, as a necessary and essential component, the organisation of work performed by individuals, irrespective of whether that work is performed online or in a certain location.

Article 22 of the General Data Protection Regulation (GDPR),[4] referring to decisions based solely on automated processing, including profiling – the centrepiece of the EU's legal approach to automated decision-making (ADM) – may certainly apply in employment contexts, although without any specificity and provided that they fall under the GDPR's scope and to the extent of its main policy goals: that is, compliance with data protection and not social security protection or protection against unfair dismissal, overwork or unfair wages.

Even the platform-to-business regulation (P2B Regulation)[5] might become relevant in the provision of ranking services if they are used in employment agreements or in the workplace to determine working conditions in some way. Yet, the recently adopted Digital Services Act[6] and the Digital Markets Act,[7] as well as the European Commission proposals on liability rules – the AI Liability Directive[8] and the revised directive on product liability (revPLD)[9] – do not contain employment-specific rules but an array of rules of a potential general character that may be suited to employment relations or be properly adapted.

Against such a backdrop, this chapter explores the main scenarios where automation is used in the workplace and for employment, devises a taxonomy to identify key legal problems and provides guidance on the construction of a consistent body of rules governing the automation of employment relationships and the workplace by assembling, combining and contextualising existing rules and principles, and proposing gap-filling solutions.

Automation includes both purely algorithmic systems ('deterministic' models or 'symbolic' AI) and learning systems driven by AI techniques and approaches ('sub-symbolic', 'stochastic' or 'machine-learning based' AI). In both cases, the terminology used to describe any such system is ADM – automated decision-making – that is, it is defined as a (computational) process and includes AI techniques and approaches that, fed by inputs and the data received or collected from the environment, can generate, given a set of pre-defined objectives, outputs in a wide variety of forms (content, ratings, recommendations, decisions, predictions, etc.) (Rodríguez de las Heras Ballell

4. Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.
5. Regulation (EU) 2019/1150 of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services [2019] OJ L186/57.
6. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Act) (text with EEA relevance). OJ L 277, 27 October 2022, pp. 1-102.
7. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (text with EEA relevance). OJ L 265, 12 October 2022, pp. 1-66.
8. Proposal for a directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM/2022/496 final.
9. Proposal for a directive of the European Parliament and of the Council on liability for defective products, COM/2022/495 final.

2022). This explanation is largely based on the (ongoing) definition of AI systems,[10] for the purposes of the future AI Act, that identifies the key elements which enable the formulation of a working definition of ADM itself: inputs (these can be human-based inputs, machine-generated data or interactions with the environment); pre-defined objectives; techniques and approaches to achieve those objectives; and outputs.

1.2 Devising a taxonomy for classifying the possible uses of automation in employment

In the process of mapping the risk scenarios and identifying the most relevant legal problems, a two-layer taxonomy is proposed.

1.2.1 First layer: from algorithmic management to smart workplaces

The first layer classifies the possible uses of automation in employment-related contexts into two main categories.

First, the automation of any decision-making process that is likely to affect workers' rights and working conditions to any extent. E-recruiting programmes and all algorithmic management systems belong to this category under which decision-making is fully or partially automated in the context of employment.

The common denominator is that the 'affected person' (someone affected by a decision taken or supported by ADM)[11] is the worker while the 'user' or deployer (who uses or relies on ADM outputs) is the employer or its agents, collaborators or contractors. The user is the person who is in control of the ADM system and benefits from its operation in the context of carrying out an economic activity. Hence, the employer (and collaborators) deploying ADM for algorithmic management purposes is the user. The proposed AI Act uses a broader definition of user including provider, deployer, authorised representative, importer and distributor of an AI system. The term 'deployer' describes any natural or legal person, public authority, agency or other body using an AI system under authority except where the AI system is used in the course of a personal non-professional activity. Accordingly, in line with the terminology of the proposed AI

10. Article 3(1) of the AI Act reads: 'Artificial intelligence system (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with'. The 'joint compromise' text unveiled at the end of November 2021, 2021/0106(COD), proposed some changes to this definition. In the preamble, the joint compromise clarifies that the proposed amendments are intended to make explicit that an AI system, unlike traditional software, should be capable of determining how to achieve a given set of human defined objectives by learning, reasoning or modelling. The revised definition is the following: 'artificial intelligence system (AI system) means a system that: (i) receives machine and/or human-based data and inputs; (ii) infers how to achieve a given set of human-defined objectives using learning, reasoning or modelling implemented with the techniques and approaches listed in Annex I; and (iii) generates outputs in the form of content (generative AI systems), predictions, recommendations or decisions which influence the environments it interacts with.' Pursuant to the latest version of the AI Act of 14 June 2023, 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations or decisions that influence physical or virtual environments.

11. As per the drafting of the text of 14 June 2023, Article 3 (8a) defines 'affected person' as any natural person or group of persons who are subject to or otherwise affected by an AI system.

Act, an employer implementing algorithmic management solutions is the deployer while the worker affected by the decisions adopted by such AI systems used by the employer is the affected person.

Second, the automation of tasks, activities and mechanical procedures in the workplace to assist workers in the performance of their jobs or as a means of fully or partially replacing workers at any stage of the working process (industrial robots, automated quality controls, autonomous vehicles, smart warehouses, predictive maintenance). This second category mostly, but not exclusively, covers purely algorithmic processes driving the performance of repetitive and mechanical tasks. Nonetheless, learning systems providing predictions, recommendations or calculations to assist or replace workers in their work are also, and increasingly, included in this category. So under these second category deterministic and non-deterministic (learning) systems are included.

The distinctive feature of this second category is that the worker is frequently the operator of ADM in a specific situation – using ADM calculations to perform a task, relying on ADM predictions in doing so or planning work on the basis of ADM recommendations. Certainly, both the employer and the worker may benefit, with different intensity, from the efficiencies of automation. However, in the case of malfunctions, the worker as well as the employer can be negatively affected with a range of potential damage. While the employer may suffer loss of profits, reputational harm and economic losses arising from a breach of contracts with clients, or late delivery, the worker may be affected by poor performance due to the errors of the system but also by physical harm or bodily injury caused by the operation of a machine or use of a device driven by the system.

1.2.2 Second layer: from ADM systems supporting decisions to ADM systems taking final decisions

The second layer distinguishes from the categories described above two further sub-categories.

ADM can produce or deliver a myriad of outputs from a rating to a bonus award decision, from a promotion recommendation to the allocation of work assignments and from a recruitment decision to the flagging of a worker as 'unreliable' leading to dismissal, the application of sanctions or a reduction in salary.

Against such a backdrop, and despite the ample variety of outputs, a distinction can be made between two situations: ADM systems that generate outputs that will be used as inputs into subsequent decision-making (feeding another ADM or supporting human decisions); or which produce a final decision that has a direct effect on the legal or contractual status of the affected person. In the former case, the classification of a worker as unreliable by a reputational system might support a subsequent (automated or human) decision to dismiss him or her; in the latter, a reputational system may demote a courier in a platform hierarchy or classify a job promotion applicant as ineligible, consequently having a direct impact on the working conditions of that courier or the labour status of that applicant.

This proposed two-layer taxonomy provides an analytical framework to discuss, explore and address the key legal issues arising from the use of ADM in employment-related contexts.

Thus, as further expounded below, ADM employed for purposes included in the first category may lead to discrimination, the violation of labour rights, data protection infringements or to unfair, unlawful or wrongful dismissal. In situations covered by the second category, a defective robot installed in the workplace may cause physical harm or bodily injury to the operating worker, material damage to inventory or other equipment or economic losses to the company. Hence, the taxonomy helps to group and classify use cases and traces an exploratory path towards potential legal problems and possible solutions.

2. Key legal issues: attribution of legal effects and the allocation of liability

The analysis of the main legal issues in this section spotlights solely private law-related matters – essentially, contractual aspects and liability – and is elaborated through the formulation of four questions:

1. Is automation permitted in employment contexts?
2. To whom can the legal effects of such permission be attributed?
3. Who is liable for the use of automation and on which grounds?
4. Are there already, or do we need, AI-specific principles and rules? Or, on the contrary existing liability-related rules suffice?

2.1 The principle of technology neutrality and non-discrimination against ADM

The most basic, but fundamental, question is whether automation (the use of ADM) is permitted in general terms in employment-related contexts either to adopt or support decisions such as recruitment, worker promotions, bonus programmes, task allocation or even dismissals; or to perform certain work activities or replace, totally or partially, workers in certain jobs. To provide an answer in the affirmative, two principles are instrumental: technology neutrality; and non-discrimination.

The principle of non-discrimination is widely recognised in international instruments on the use of electronic communications in international contracts. The United Nations Commission on International Trade Law (UNCITRAL) model laws on electronic commerce (1996), on electronic signatures (2001) and on electronic transferable records (2017) are all based on the principles of non-discrimination, technological neutrality[12]

12. The principle of technology neutrality is not usually defined in international texts in an affirmative manner but it may be inferred from acknowledgement of the non-discrimination principle ('legal effects shall not be denied solely on the grounds that certain technology is used') and it underlies the provisions recognising functional equivalence. For instance, Article 6 of the UNCITRAL Model Law on Electronic Commerce states: 'Where the law requires information to be in writing, that requirement is met by a data message (…)'. This statement entails a technology-neutral approach.

and functional equivalence. More precisely, the United Nations Convention on the Use of Electronic Communications in International Contracts (2005)[13] extends the principle of non-discrimination to the use of automated systems whose actions are not reviewed or triggered by natural persons.[14] Thus, in the absence of human intervention, the action performed by an automated system shall not be denied legal effect, validity or enforceability solely on the grounds that it is performed by automated means.

These international principles endorse the use of law-compliant ADM, unleashing the full potential of automation but without compromising the protection of rights and liberties.

A non-discrimination rule neither necessarily means that ADM-specific rules cannot be adopted nor implies that their use cannot be limited, exempted or subject to conditions, as discussed below. Indeed, the future AI Act classifies as high-risk systems those used for certain purposes in the area of employment, worker management and access to self-employment,[15] but it does not forbid their use in employment entirely. Accordingly, specific requirements and limitations will apply in high-risk cases, but only the general rules applicable to the 'equivalent non-automated situation' will govern other uses.

However, the key finding is that the use of ADM in employment should be evaluated with consideration to the specificity of the use case, including the data collection and processing upon which ADM relies. Regardless, the certification of the AI system should strive to be as comprehensive as possible, encompassing the sectoral law requirements and fully complying with all the relevant principles and legislation.

2.2 The allocation of legal effects

If automated decisions or ADM-supported decisions related to employment can be valid and enforceable, the next question is to whom to allocate the legal effects of such actions or decisions. Whether the decision is to hire a candidate, promote an employee, refuse the promotion request of a worker, grant a bonus, apply a disciplinary action or dismiss a worker, a decision-making process may be either fully or partially automated. When the decision is fully automated, to whom such a decision of the ADM system and its legal effects should be attributed is a legitimate, and critical, question. Should it be deemed a decision of the employer? The ADM system operates with a substantial level

13. United Nations Convention on the Use of Electronic Communications in International Contracts (New York, 2005) (adopted 23 November 2005, entered into force 1 March 2013).

14. Art. 12 of the UN Convention defines the use of automated message systems for contract formation thus: 'A contract formed by the interaction of an automated message system and a natural person, or by the interaction of automated message systems, shall not be denied validity or enforceability on the sole ground that no natural person reviewed or intervened in each of the individual actions carried out by the automated message systems or the resulting contract.'

15. Annex III, AI Act, as amended by Parliament: '(a) AI systems intended to be used for recruitment or selection of natural persons, notably for placing targeted job advertisements screening or filtering applications, evaluating candidates in the course of interviews or tests; (b) AI systems intended to be used to make or materially influence decisions affecting the initiation, promotion and termination of work-related contractual relationships, task allocation based on individual behaviour or personal traits or characteristics, or for monitoring and evaluating performance and behavior of persons in such relationships.'

of autonomy, without human intervention in each decision, and collects and processes data from several sources, driven by learning techniques and even, on a few occasions, as a result of flaws, errors or malfunctions. Is a decision then to be attributed to the system developer, the data providers, the software update provider, the users or the distributors?

A decision taken by an ADM system is deemed to be the decision of the user implementing, employing or applying that ADM in making or supporting decisions for the purposes and within the scope of its activity. Accordingly, the legal consequences of such a decision are to be attributed to the user, regardless of the decision being arrived at by automated means. For such purposes, the employer is the user.

Hence, the employer has to assume the legal effects and bear the consequences of an ADM decision; the employer cannot excuse itself from complying with an ADM decision or bearing its legal consequences solely on the grounds that the decision was made by automated means. The user also cannot deny that the decision may be attributed to it on the grounds that the ADM was developed by a third-party provider or that data was collected from third-party data providers. The decision is not attributed to the programmer, the system provider, the distributor or the data providers; it is the user that is responsible for ensuring that the ADM system is fit for its intended purpose and operates as it should.

2.3 Risk scenarios and liability rules

The use of ADM can provoke, trigger, aggravate or intensify risks but, either way, the question is one of to whom to allocate the legal consequences. Three liability scenarios can be explored.

2.3.1 Noncompliance with legal, collective and contractual terms

An employer who decides to use an ADM system for employment-related purposes or in the workplace must ensure that it operates in compliance with the law, the applicable collective labour agreement and the contractual terms of employment. Should the system deviate from either of these, there will be an infringement attributable to the employer leading to liability (legal remedies, administrative sanctions, compensation for damages or other contractual actions).

The following examples provide a few illustrations:

> If a bonus programme, as agreed in the employment contract, is based on five objective/quantitative performance criteria, the ADM cannot be programmed to use other factors, such as a reputational system departing from the agreed one.
>
> Should race become a relevant criterion (either intended and programmed, or subsequently learned by the ADM but which has passed unnoticed) for an

automated recruitment programme, an employee promotion campaign or a bonus calculation mechanism, the user may be committing racial discrimination.

If, pursuant to a collective bargaining agreement, employees are free to opt for full-time work or a part-time alternative with no differential treatment, an ADM system that allocates tasks (deliveries) preferentially to full-time workers over part-time ones and consequently penalises part-time workers for not reaching the minimum number of deliveries per week is, at the very least, violating the collective agreement.

Should robots installed in the workplace be improperly and non-regularly monitored and maintained, and cause personal injury to workers, the employer may be infringing rules on the prevention of labour risks.

2.3.2 Liability for damage for defective AI systems and AI-enabled products

Should a defective ADM system be employed in the workplace (industrial robots, smart warehouse, quality controls) or in the context of employment (task allocation, reputational systems, employee promotion programmes), damage can be caused to a company's property, while workers may be injured, data lost or corrupted and the manufacturing chain interrupted.

One of the challenging questions is whether and, if so, to which extent the classical product liability regime (PLD)[16] is applicable to ADM. The product liability rules do not easily accommodate ADM and there are several hurdles to go over: the concept of product; the meaning of defect; the rationale of some defences; and the liability of the economic user (as producer, importer, distributor or provider). That is precisely the aim of the proposal for a revised PLD published in September 2022.

Under the proposal's revised approach, AI systems and AI-enabled goods are 'products' for the purposes of the scope of the revPLD. That means, without the need to prove the manufacturer's fault, where an injured party manages to establish the defectiveness of a product, the existence of a causal link between that defectiveness and damage, and the level of damage itself (even if the burden of proof is alleviated by certain presumptions as per the revised Article 9), he or she can be compensated for the damage – that is, the material losses resulting from death or personal injury, or harm to property, to the extent provided for the proposed directive – which has been caused by the defective ADM system.

However, who is the injured person entitled to claim and which damages can be compensated? The proposal is not particularly clear in determining who can be an injured person entitled to claim compensation but combining the reference to 'natural persons' in Article 1 and the extent of compensable 'damages' in Article 4(6), workers are amply within the coverage of the product liability regime. However, it is very unlikely

16. Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. OJ L 210, 7 August 1985, pp. 29-33.

that employers can be compensated for the damage resulting from a defective ADM system that they have employed (i.e. for property harmed by a defective robot) insofar as Article 1 limits compensation to natural persons while Article 4 excludes harm to, or the destruction of, property used exclusively for professional purposes.

In terms of liability, in cases other than those where the damage entails a violation of labour safety prevention policies or there are other causes attributable to the employer,[17] the primary liable economic operator is the manufacturer. From there, a cascade of liable economic operators descends (importers, distributors, refurbishers, online platforms). The employer as a mere user is not in the cascade of liable economic operators on the grounds of product defectiveness. Even so, does the employer become a potentially liable economic operator when it is not simply a passive user? Interestingly, the proposal is not clear (in Art. 7(4)) in its definition of when a person who substantially modifies a product already placed on the market becomes liable. That assessment is critical to an understanding of whether an employer who customises, personalises, adapts or modifies an ADM system to make it fit for a specific purpose, after acquiring it from its manufacturer, might be liable.

Workers, as victims of the damage caused by a defective product, and protected by the product liability rules, might, however, be in practice rather defenceless if they have to raise a claim against the manufacturer, the importer or even the platform where the employer is not a liable person under the product liability regime (even if the employer may be liable on other liability grounds). Cost, procedural complexities, burden of proof, parties' locations or conflict-of-law aspects may provide a drastic deterrence to a claim. Therefore, the proposal adds some innovations to the PLD aimed at enhancing the protection of injured persons by alleviating the burden of proof and by enlarging the liability scope of the manufacturer when AI is used.

As far as the burden of proof is concerned, the traditional rule that the injured person has to prove the defectiveness of the product, the causal link between defectiveness and damage, and the damage itself is preserved. However, some presumptions have been provided to alleviate the burden, especially in cases of technical or scientific complexity. This may apply to ADM due to complexity, opacity and a certain level of unpredictability in the operation of learning systems. Thus, defectiveness would be presumed when the claimant establishes that the product does not comply with mandatory safety requirements or that the damage was caused by 'an obvious malfunction of the product during normal use or under ordinary circumstances'.

In relation to the scope of liability, a manufacturer shall not be exempted provided that the defectiveness is within its control or is due to a related service, a software problem, including software updates or upgrades, or reflecting a lack of software updates or upgrades necessary to maintain safety. Consequently, the manufacturer can be held liable if, upon placing the product on the market, it then becomes defective, as long as the defect is triggered by a subsequent update or upgrade, or lack of an update or

17. Council Directive of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work (89/391/EEC) – 'the Framework Directive'.

upgrade, within the control of the manufacturer. For instance, if the manufacturer fails to provide an update to prevent cybersecurity breaches or the compatibility of the software embedded in the product with a common operating system and, as a result, the product causes damage, the manufacturer may be liable.

It must be noted that the product liability rules do not cover all liability scenarios and types of losses, only those triggered by the defectiveness of a product in relation to damage compensable under the PLD and within the scope of the rules. Indeed, Article 2 explicitly acknowledges that the revPLD does not affect, in particular, any rights which an injured person may have under national rules concerning contractual liability or concerning non-contractual liability on grounds other than the defectiveness of a product. That leads to a third liability scenario to which the other recently published proposal – the AI Liability Directive – to be adopted as a package alongside the revPLD proposal, is related. The claims do not overlap as the grounds for liability as well as the potential liable economic operators and types of damage for which compensation can be claimed are different.

2.3.3 Fault-based liability as a general default liability system

An employer using ADM in the workplace and for employment purposes can be at fault for injuries caused to workers (in addition to, or alternatively to, contractual liability) or even to third parties. As far as workers are concerned, regarding non-contractual liability arising from accidents at work, in most cases there is strict liability at national level.

The rules on fault-based liability are largely non-harmonised and depend upon the national laws of Member States. The aim of the draft AI Liability Directive, which is much less ambitious than the preceding European Parliament Resolution of 2020 that provided for strict liability for listed high-risk AI systems,[18] is to provide common rules on the disclosure of evidence on high-risk AI systems and to establish the burden of proof in the case of non-contractual fault-based civil law claims brought before national courts for the damage caused by an AI system. A set of rebuttable presumptions is laid down to alleviate the perceived complexities surrounding claims for damages caused for AI systems due to their distinctive features (opacity, data dependence, vulnerability, learning capabilities, openness).

In practice, the injured person, either the worker or a third party, would claim compensation from the employer pursuant to the applicable national laws on (fault-based) civil liability, but they would benefit from the alleviation of the burden of proof provided by the common presumptions laid down in the directive. Whether the resulting burden of proof allows the person affected by the use of the high-risk AI system to overcome opacity or whether this is still an 'impossible burden' is an open question. In particular, once the provider has proven compliance with the requirements

18. European Parliament resolution with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)) [2020] OJ C404/107, that includes a proposal for a regulation of the European Parliament and the Council on liability for the operation of artificial intelligence systems.

of Chapter II of the AI Act, the worker must prove all four elements listed in the AI Liability Directive: namely, the fault of the provider; the causal link between the fault and the output of the AI system; the causal link between the output of the AI system and the damage; and the damage itself.

It is relevant to note how the draft AI Liability Directive builds a bridge with the future AI Act by connecting the non-compliance of harmful AI systems, having certain requirements for high-risk AI systems as laid down in the AI Act, with the rebuttable presumptions provided for in the AI Liability Directive. Thus, it meets head-on one of the criticisms of the AI Act; that it does not provide for individual rules of redress. The AI Act indeed makes no provisions in this area but the AI Liability Directive aims to facilitate compensation for damage according to non-contractual fault-based liability rules at national level by setting presumptions triggered by non-compliance with the requirements of the AI Act. The consequence is that the two future pieces of legislation will interplay as follows: non-compliance with some of the requirements of Chapter II of the AI Act may trigger the rebuttable presumptions of the AI Liability Directive; and, therefore, the burden of proof on the victim will be eased in a claim of fault-based liability against the user (the employer) for the damage caused by a non-compliant AI system pursuant to fault liability under national laws.

In order to ensure legal certainty and prevent a substantial reduction in the level of non-contractual protection for workers, the proposed AI Liability Directive should clearly stipulate that, where the business activity to which the high-risk AI is applied is subject to strict liability or another sectoral or specific non fault-based regime, the proposal does not affect this liability framework.

To conclude, the previously-analysed scenario can be put in practice in the following hypothetical situation:

> A smart warehouse is equipped by interconnected AI-enabled devices and smart equipment that, based on dynamic data and activity predictions, automatically handles inventory, packs deliveries, unloads an incoming vehicle and sends instructions to the personnel allocating tasks and requiring the confirmation of orders. Unexpectedly, one of the pieces of industrial equipment starts performing random movements, causing damage to stored goods, damaging other equipment and injuring two workers.
>
> It was proved that the equipment was defective, but it had not been maintained on a regular basis by the company due to the high cost. Additionally, it was proved that the damage was aggravated because the equipment was not fit for the assigned purpose and, therefore, there were interoperability issues.
>
> The manufacturer (unless located outside the EU, when the liability passes to importers or other users along the liability cascade) would be liable for the injuries caused by the AI-enabled device and smart equipment to the individual workers, to the extent covered by the revPLD. The burden of proof on the workers would be alleviated by the measures discussed above laid down in the revPLD. In parallel,

workers might be entitled to raise a claim on strict liability grounds against the employer (lack of monitoring, maintenance and unsuitable selection), in addition to any extra-contractual fault-based liability, where national laws provide for this in cases beyond the sphere of the employer's strict liability. Besides, the damage caused might be deemed to arise from non-compliance with labour risk prevention duties that would also form the object of contractual obligations for the employer, in which case the employer would be additionally liable on such grounds (contractual liability). The worker, in most cases, can invoke both the contractual and the extra-contractual liability to the extent allowed under Member State law.

3. Conclusion

This chapter explores the main scenarios where automation is used in the workplace and for employment, devises a taxonomy of automation situations to identify key legal problems, and provides guidance to build a consistent body of rules governing the automation of employment relationships and the workplace by assembling, combining and contextualising existing rules and principles, and proposing gap-filling solutions.

The possible uses of automation in employment-related contexts are classified into two main categories: the automation of any decision-making (algorithmic management); and the automation of tasks, activities and procedures in the workplace (smart workplace).

The main legal issues have been identified as centring on four questions: legality (is the use of AI permitted by the law?), validity (are decisions adopted by AI valid?), enforceability (can decisions and actions performed by AI enforceable?) and liability (who is liable?).

The primary finding is that a consistent and combined application of the rules provided for by the various texts related to AI in the EU is necessary to tackle in full the legal issues which arise from automation in employment. Thus, the future AI Act should be applied in combination with a number of other provisions established either by EU instruments or by national rules. In particular, the proposals of the directive related to liability issues are of special relevance insofar as they are aimed at addressing the specificities of the damage caused by AI systems. These rules are applicable to the extent that AI systems are used, but they should be aligned with the relevant sectoral legal framework. Furthermore, none of these AI-specific regimes are intended to repeal labour laws which will continue to be applied and which should not entail a diminishing of workers' protection solely on the grounds that automation is used in an employment-related context. On the contrary, it has been advocated that full and effective protection needs to be ensured for workers where AI is used in employment contexts from algorithmic management to smart workplaces.

References

Council of the European Union (1985) Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, Official Journal of the European Union, L 210, 7.8.1985. https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vitgbgh2ouzc

Council of the European Union (1989) Council Directive 89/391/EEC of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work, Official Journal of the European Union, L 183, 29.6.1989. https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vhckkmbs3qy5

European Commission (2021a) Proposal for a Directive of the European Parliament and of the Council on improving working conditions in platform work, COM(2021) 762 final 9.12.2021. https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vlolewihtkxw

European Commission (2021b) 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, 21.4.2021. https://www.eumonitor.eu/9353000/1/j4nvhdfdk3hydzq_j9vvik7m1c3gyxp/vli6mrzmjby8

European Commission (2022a) Proposal for a Directive of the European Parliament and of the Council on liability for defective products, COM(2022) 495 final, 29.9.2022. https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vlwrmdrl2dxc

European Commission (2022b) Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final, 28.9.2022. https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vlwrmdrw4nxd

European Parliament (2020) European Parliament resolution with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)), 20.10.2020. https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.html

European Parliament (2023) Artificial intelligence act. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1, 14.6.2023. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

European Parliament and Council of the European Union (2016) Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union, L119/1, 4.5.2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Parliament and Council of the European Union (2019) Regulation (EU) 2019/1150 of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services, Official Journal of the European Union, L186/57, 11.7.2019. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R1150

European Parliament and Council of the European Union (2022a) Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Act) (text with EEA relevance), Official Journal of the European Union, L 277, 27.10.2022. https://eur-lex.europa.eu/eli/reg/2022/2065/oj

European Parliament and Council of the European Union (2022b) Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (text with EEA relevance), Official Journal of the European Union, L 265, 12.10.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R1925

Rodríguez de las Heras Ballell T. (2022) Guiding principles for automated decision-making in the EU, ELI Innovation Paper, European Law Institute. https://europeanlawinstitute.eu/news-events/news-contd/news/eli-innovation-paper-on-guiding-principles-for-automated-decision-making-in-the-eu-now-available-f/

UN (2005) United Nations convention on the use of electronic communications in international contracts. https://uncitral.un.org/en/texts/ecommerce/conventions/electronic_communications#:~:text=Purpose,their%20traditional%20paper%2Dbased%20equivalents

All links were checked on 16.02.2024.

Cite this chapter: Rodríguez de las Heras Ballell T. (2024) Automating employment: a taxonomy of the key legal issues and the question of liability, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 13
Worker monitoring vs worker surveillance: the need for a legal differentiation

Aída Ponce Del Castillo and Michele Molè

1. Introduction

In recent years, the use of increasingly powerful technologies to monitor workers has become a prevalent and almost ubiquitous feature of the labour market. These technologies are used for a variety of purposes including monitoring productivity, measuring workers' performance, tracking movements and workers' health, and making profiles of the workforce used for multiple purposes. They have permeated all types of workplaces and sectors. No job, task, profession or sector is immune to this intrusive practice which reduces workers to a 'quantified self' (Lupton 2016; Moore 2018; Swan 2013). It can extend beyond the confines of the workplace, working time and one's working life (Hendrickx 2018). Against this background, the European Social Partners, as delineated in their Autonomous Framework Agreement on Digitalisation (2020), have identified worker surveillance as a pivotal topic for social dialogue.

The objective of this chapter is to point to the increasing interest in workers' permanent measurement provided by new surveillance products. It provides factual and legal arguments that support the need to draw a line between the concept of 'worker monitoring' and 'worker surveillance', arguing that such surveillance creates a situation of structural asymmetry between employers and workers in terms of information and control at work, while also leading to abuses of workers' rights. Against this background, it highlights the need to set boundaries to employers' monitoring prerogatives, themselves significantly enhanced by the development of new and powerful surveillance technologies. The analysis is based on literature from sociology, surveillance studies and law.

Section 2 describes the right of employers to monitor the workforce under the European Convention on Human Rights (ECHR). Although this monitoring power is governed by the principles of proportionality and necessity set down in Article 8 ('Right to respect for private and family life'), it contends that the rapid advance of technology and its promises has provided employers with an intrusive and invisible, yet intensive, monitoring power which should not find the European legal framework on worker monitoring unprepared.

Section 3, after describing how worker monitoring has evolved and transformed into surveillance, takes a closer look at the recent decisions taken by national data protection authorities (DPAs) and argues that privacy in the workplace is an issue that should more consistently be on their radar. It looks at the impact of surveillance on occupational safety and health, and argues that this has led to a gradual erosion of the principle

and culture of prevention. Section 4 addresses data, the driving force behind worker surveillance, and data analysis.

Section 5 lays the groundwork for a legal definition: it shows how surveillance can be considered an unnecessary monitoring power of the employer which often fails the suitability and necessity tests as developed by the jurisprudence of the European Court of Human Rights (ECtHR). Then this chapter concludes.

2. The monitoring power of employers: an open contractual clause

The right to monitor employees is one of the essential powers granted to employers 'to ensure the smooth running of the company', a concept discussed in the ECtHR landmark case *Bărbulescu v. Romania* (2017) (which went on to rule that the monitoring of an employee's corporate messaging account was a violation of the right to respect for private life and correspondence under Article 8 of the ECHR). Such a power includes a variety of monitoring activities covered by an employer's right to property and to conduct a business (Art. 16 and Art. 17 Charter of Fundamental Rights of the European Union (CFREU); Art. 1 ECHR Protocol 1). Managing an economic activity is linked with the legal and legitimate interest of the employer to oversee any work-related activity: company organisation; employee performance; compliance with occupational safety and health (OSH) laws; and protection of company assets. Therefore, the monitoring power of the employer is an implicit clause of the employment contract, defined here as a contractual clause that does not set specific details about the intensity, quality or pace of the monitoring that the employer will perform. Thus, in the contract, the employer has the ability to monitor the workforce with few limitations, allowing for direct or technology-based observation, e.g. cameras, sensors, geolocation systems or device tracking by default (such as laptops, smartphones, etc.) (Ball 2010; Carby-Hall 2003: 34; Coase 1937).

Such an employer power finds, however, either restrictive or permissive regulatory approaches in the various European Member States. As noted by Eurofound, legislative interventions on employer discretion in monitoring are focused, at national level, mainly on two purposes: they legitimise monitoring for occupational safety and health or security purposes; while they prohibit (or link to transparency duties) the direct monitoring of the employee's work performance (Riso 2020: 8-16).

A similar approach to limiting employer discretion comes from the European Court of Human Rights. With the *Niemietz* case (*Niemietz v. Germany*, 1992) came the first reference to the protection of an employee's privacy and personal data under the scope of Article 8 ECHR. Mr Niemietz was a lawyer who complained before the ECtHR that a search of his offices was an interference with his private life. There, the Court held that there exists in a workplace an individual right to personal development and to establish relationships with other individuals; a sphere that cannot be disproportionately reduced to accommodate an employer interest in monitoring.

Thereafter, the Court has addressed diverse and specific implementations of this monitoring power. Monitoring via cameras or GPS has often been found to be legitimate for preventing theft or the abuse of company property (*Florindo de Almeida Vasconcelos Gramaxo v. Portugal*, 2022; *López Ribalda and Others v. Spain*, 2019; *Köpke v. Germany*, 2010). In these cases, monitoring systems had been (legitimately) implemented to prevent the illegitimate use of a company car through GPS tracking (*Florindo*, 2022) or to provide video evidence of the theft of goods or money by shop assistants (*López Ribalda*, 2019; *Köpke* 2010). However, the scope of the power has often been found to contravene the right to privacy when cameras or other tracking systems have been employed to monitor an employee's work performance (*Antović and Mirković v. Montenegro*, 2018; *Bărbulescu v. Romania*, 2017; *Copland v. United Kingdom*, 2007; *Halford v. UK*, 1997). In *Antović and Mirković v. Montenegro* (2018), Ms Antović and Mr Mirković alleged that the unlawful installation and use of video surveillance equipment in the university auditoriums where they held classes had violated their right to respect for their private life. Similarly for Mr Bărbulescu's Yahoo Messenger account: opened for business purposes, this was monitored by his employer including his personal and non-professional communications. In the cases involving Ms Copland and Ms Halford, their right to private life was, according to the Court, disproportionately reduced by the covert monitoring of email and internet use (*Copland*, 2007) and of telephone conversations (*Halford*, 1997).

Looking through the ECtHR's rulings on worker monitoring, two principles set boundaries on a case-by-case basis to the 'openness' of the monitoring power: necessity and proportionality. The Court weighs the proportionality of the intrusion into the private sphere of employees against the right of employers to manage and preserve their economic activity. In addition, the practical implementation of the monitoring measure is also assessed: according to the principle of necessity, the employer must adopt the most suitable and least intrusive measures to express its legitimate interest in monitoring (Gerards 2013). The principles of proportionality and necessity of the monitoring measure are also reflected in Art. 5 GDPR ('Principles relating to processing of personal data'). Yet, the European Court of Justice (ECJ) has still not dealt with the monitoring power of employers and the application of the GDPR's principles (Mangan 2022: 321).

3. The transformation of monitoring into surveillance

As described above, there is extensive ECtHR jurisprudence on monitoring and privacy at work. Today, however, we are witnessing a gradual transformation of monitoring into surveillance, a process characterised by the acquisition of new features and purposes: employers use monitoring technologies that go beyond those analysed so far by the courts, such as video surveillance, GPS, biometric scans or tools that allow access to personal communications (Zuboff 2019). Powerful new technologies are creating new challenges for courts as to what can be considered proportionate and necessary in the digital age. Increasingly, labour scholars are calling for a 'technological contextualisation' of labour regulation in light of the rapid expansion of the market

for workplace surveillance technologies which threatens the enjoyment of fundamental rights at work (Aloisi and De Stefano 2022; Mangan 2022; Molè 2022).

This chapter refers to this new form of all-encompassing monitoring as 'worker surveillance' and argues that it operates on the basis of 'the three Is': it is intrusive on and invisible to the individuals it targets; and is characterised by the intensive collection of data.

The sociologist Gary T. Marx defines surveillance in the *International Encyclopedia of the Social & Behavioral Sciences* as the 'scrutiny of individuals, groups, and contexts through the use of technical means to extract or create information' (Marx 2015). In a post Covid-19 context, this definition remains valid, but the phenomenon has undergone exponential growth. It now operates on a practically limitless scale and generates constant and substantial amounts of fine-grained personal and sensitive data. Modern surveillance tools capture data points related to the worker's emotional state (anxiety, frustration, boredom, happiness, fear, insecurity, etc); safety (exposure to hazards, risk levels, movements, fatigue, microsleep episodes, etc); health (physiological data such as heart rate, blood pressure, breathing rate, temperature, ergonomic data such as 'good' or 'bad' posture, stress levels, possible burnout, etc.) (Al Jassmi et al. 2019; Jebelli et al. 2019; Moore 2018; Swan 2013); wellness (sleep patterns, fatigue management, level of physical activity, etc) (Dockser Marcus 2023); brain activity (Cheng et al. 2022; Farahany 2023; Wang et al. 2017); security (use of company assets, information leaks, risky behaviours, etc); and productivity (engagement with teammates, working time vs rest time, contents of e-mails, internet use, etc.) (Burnett and Lisk 2021). The collected data is then measured, analysed and processed for a variety of purposes.

Surveillance tools and techniques are often invisible and non-material, embedded within other technologies and devices. This, coupled with the broad range of data points collected, makes worker surveillance a markedly distinct practice from monitoring. As briefly anticipated in Section 2, a legal differentiation following the transformation of monitoring into surveillance has become imperative.

The gradual shift from monitoring to surveillance can be primarily attributed to technological advance and the enhanced capacity to gather data. First, the advent of new software, tools and technology considerably expanded the scope, until then relatively limited, of human resource management (HRM) (Gallup 2023; Laker et al. 2020). This led to the emergence of people analytics (PA), which HRM experts describe as a data-driven method analysing all the processes related to personnel in a company with the aim of achieving business success and increasing the organisation's efficiency and productivity (De Cremer and Stollberger 2022). The next generation of tools incorporated novel AI-based models to enable even more intricate surveillance and provide employers with more powerful insights in decision-making. These tools are used to assist managers when allocating rewards (salary rises, promotions) or imposing disciplinary measures (dismissals, suspensions of platform accounts, etc.).

The latest tier in this hierarchy of monitoring techniques is algorithmic management (AI Now Institute 2023), or automated monitoring and decision-making systems. Through

the use of third parties, vendors, networks, data brokers and transfer mechanisms, the power of the employer to process workers' data has dramatically increased. Surveillance tools not only capture data, to be processed and used by managers, but they can also be used to support decisions (Adams-Prassl 2019).

Algorithmic management is one of the building blocks of the platform business model (Stark and Pais 2020) and it is gradually encroaching on more traditional forms of employment. The proposed directive on improving working conditions in platform work aims to regulate automated decision-making and monitoring systems in the platform economy (European Commission 2021), but several challenges remain – especially when used in standard employment settings. These include the overall opacity of existing processes, the metrics employed, the implications of collecting data in nanoseconds, the use of surveillance techniques in determining workers' employability and the impact of behavioural analytics on workers' remuneration but which – as also noted by Bales and Stone (2020) – deter unionisation, enable subtle forms of employer blackballing, exacerbate employment discrimination, render trade unions ineffective and eradicate the protections of labour law.

3.1 Privacy and data protection in the workplace: a new issue of concern for data protection authorities

The Covid-19 pandemic brought the issue of worker surveillance into the spotlight and captured the attention of both the general public and the workforce. With companies wanting to monitor workers physically absent from the workplace, the use of analytical tools intensified. Surveillance became more prevalent, justified on the grounds of productivity and safety (Ball 2010), and started to be employed for a multiplicity of purposes going beyond workers' 'data perimeter' as identified by Mario Guglielmetti (this volume). Instances of misuse became frequent, as data collected from workers was used to penalise and discipline them, as well as to automate decisions that had adverse effects on them (Agosti et al. 2023; Rogers 2023). Against this backdrop, it is essential to consider whether the obligation to conduct a data processing impact assessment under Article 35 of the GDPR, involving the participation of worker representatives, has been adequately met.

Privacy in the workplace became an issue on the radar of national DPAs. Several issued recommendations to employers about the collection of data on remote workforces. In the UK, the Information Commissioner's Office (ICO)I reminded employers that, prior to starting processing, they must first assess whether the use of artificial intelligence (AI) is a necessary and proportionate solution to a problem (ICO 2022). In France, Commission Nationale Informatique & Libertés (CNIL; National Commission for Information Technology and Liberties) highlighted the existence of a particularly invasive piece of software which, when used, leads to a permanent and disproportionate surveillance of employees' activities. L'Autorité de protection des données/Gegevensbeschermingsautoriteit (APD/GBA; the Belgian DPA) reminded of the general prohibition on using cameras with AI to monitor the workforce.

In some cases, companies have been fined for excessive monitoring, including H&M's Service Centre (35 million euros) and notebooksbilliger.de (10.4 million euros), both in Germany (European Data Protection Board 2020a, 2020b).

Similarly, the Italian DPA fined the food delivery platform Foodinho 2.6 million euros, finding that 'the company had failed to adequately inform its employees on the functioning of the system and had not implemented suitable safeguards to ensure accuracy and fairness of the algorithmic results that were used to rate riders' performance' GPDP 2021). In other words, the Italian DPA questioned the covert surveillance and meaningfulness of the measurements carried out by the company regarding employees' activities: it found the data collected to be not relevant and used for discriminatory purposes, including communications in chats, emails and phone calls between couriers and customer care. Furthermore, technical evidence revealed that, when the mobile app was running in the background, it continued to send notifications of unassigned orders to all couriers, even those not on shift. It processed couriers' GPS location continuously and automatically – without verifying the actual need for such processing; sent data to the platform on the exact location of a courier, the speed and mobile phone battery level; and shared data, including GPS location, personal login, name and the courier's unique identifier, with third parties. It also produced a 'hidden' score for the courier, with no clear indication of the purpose of this value (Agosti et al. 2023).

Another relevant case is CNIL's investigation into the collection and analysis by Amazon France Logistics of data on its employees. In this case, Amazon was investigated on two indicators: 'the machine gun', which is the amount of time in which the worker puts away an item in less than 1.25 seconds; and 'idle time', when the worker does not store an item for 10 minutes. The CNIL has requested a fine of 170 million euros (Vitard 2023).

That DPAs are aware of the significant issue represented by privacy in the workplace is a welcome evolution. However, many types of surveillance practices continue to exist without being detected by any authority. In improving the enforcement of the GDPR, DPAs and labour authorities could intensify their level of cooperation, become allies and cross-fertilise their respective activities.

3.2 The impact of surveillance on occupational health and safety: a gradual erosion of prevention

One of the workplace dimensions most targeted by surveillance technologies has been occupational safety and health. Companies are increasingly resorting to technology not only regarding various aspects of their operations and work organisation but for the purposes of primary prevention and risk reduction as well as a mechanism for compliance with OSH legislation. Relying on the data of workers that that should not be being processed by the employer in the first instance, thereby trespassing workers' data perimeter, marketing strategies often feature promises such as 'streamline your safety processes and help create a culture of safety' (www.fluix.com); observing that 'automated prediction programs allow construction employees to minimize errors

during calculation, errors that could have created real risk during the building process' (www.kreo.net); while some seek to substitute the role of safety staff – 'by automating the repetitive and mundane aspects of the safety and health inspection, Intenseye enables safety inspectors and customers to utilize their skills for problem-solving while providing them with the relevant information regarding the problems' (www.intenseye.com). The promise is clear that sensors, cameras, IoT (internet of things) devices and AI systems can be promoted as tools that vigilantly monitor workplaces in real time, predicting potential risks, recognising unsafe behaviours, detecting unsafe conditions and suggesting recommendations to mitigate potential risks (Malik 2023; van Rijmenam 2023).

Even when AI systems could be helpful in risk management, the reliance on them needs to be carefully weighted since trusting in techno-solutions can embody a questionable shift of approach. The promotion of a culture of prevention, based on the active participation of and cooperation between managers, workers and their representatives, is a principle recognised and promoted by the International Labour Office in Convention 186 and enshrined in the European Framework Directive on Safety and Health at Work. Evidence shows that the effectiveness of safety management systems depends on a collective and positive health and safety culture that steers work towards optimising workers' physical and mental health (Wadsworth and Walters 2019; Nielsen 2014; Menéndez et al. 2009).

Following Gould (this volume), this chapter argues that technology-based surveillance is being pursued as another face of technology solutionism which finds it difficult to comprehend the reality of work, human factors and associated uncertainties. It not only goes beyond workers' 'data perimeter' but is increasingly being mistaken for, while also replacing, the culture of prevention that underpins occupational safety and health. Instead of establishing a cooperative and dynamic prevention culture, companies are opting for a 'machine-based' approach to 'advise', 'recommend', 'help' or even 'prevent' risks from materialising.

When using technology to pursue occupational safety and health, privacy is often not only seen as a trade-off but is quickly traded away. Yet privacy, and data protection, are an integral element of the right to health and to being safe, as well as a dimension of human integrity and ultimately human dignity, the foundation of all fundamental rights. The interconnection between health, safety, wellbeing and privacy is inseparable. Against such a backdrop, EU-OSHA, the European Agency for Safety and Health at Work, appears to endorse the use of automated measurement systems (comprising sensors, smart personal protective equipment, virtual and augmented reality, drones, etc.) to detect, minimise or eliminate risks (EU-OSHA 2023). The Agency has compiled eight case studies that exemplify how advanced robotics and AI-based systems can be used to automate physical and cognitive tasks in the workplace (EU-OSHA 2023).

Although an approach which sets out to minimise risks seems to be welcome, there are several implications to consider.

First, following Hildebrandt (2023), the automation of practices and even norms is a growing trend. There is an implicit validation of technology-based solutions that are being marketed to pursue preventive purposes, but without due diligence being undertaken in verifying design choices; identifying how well they fulfil their claims, functionalities and potential uses described in the terms of use; establishing how far they are reliable, effective or even helpful; and labelling their associations with third parties. As Hildebrandt (2022) sums up, 'the relevance of the solution always depends on purpose, context and agents'.

Second, the use of 'all-in-one' solutions which serve multiple purposes contradicts the coexistence of various fundamental legal dispositions. It moves away from a comprehensive approach to conducting risk assessments, characterised as a 'collaborative practice'. As various authors have observed, employers can, in their duty to assess risks systematically, work in collaboration with workers, their representatives and other experts in seeking to enhance worker protection encompassing diverse perceptions of risks, reviewing both qualitative and quantitative data and considering the variety of hazards and their severity (Castro and Ramos 2017; EU-OSHA 2007; Frick 2011, Ollé-Espluga et al. 2015). Also, as discussed in Section 2, it contravenes the core principles of necessity (assessment of the effectiveness of the measure in relation to the objective pursued) and proportionality (assessment of the appropriateness of the extent to which there is a logical link between the measure and the legitimate objective pursued) (EDPS 2019). Additionally, solutions often justify their deployment through ethical principles and rely on the consent of workers,[1] inadvertently disregarding that, when processing workers' personal data, often of a sensitive nature, informed consent does not constitute a lawful legal basis for such data processing (Article 29 Data Protection Working Party 2017).

Third, the risks associated with this situation are numerous both at individual and collective levels. Workers' autonomy and decision-making power, their capacity to be critical towards the use of technology and their capacity to bargain collectively can all be eroded (Hendrickx 2018). Its use may also conceal workers' tacit knowledge of their working environment and be a cause of deskilling and a reduction of agency.

One must also consider the implications for safety representatives and labour inspectors in this new environment. Despite the obligations under the Framework Directive, legal challenges have also been made related to liability and accountability in the event of system failures or accidents. This may provide some employers with the means to evade liability by shifting blame to the victim for non-compliance with safety rules or any failure to follow system recommendations, or otherwise by attributing the mistake to the system itself.

At work organisation level, the reliance on AI systems within intricate interconnected environments involving various actors such as employers, employees, technology providers, vendors and third parties may dilute the culture of prevention. Exposure to

1. For example, Intenseye's Ethics Principles state that: 'Intenseye requires its users to deploy their technology responsibly, clearly informing workers and limiting their use of technology to its intended and justified goals' https://www.intenseye.com/company/ai-ethics-statement

technologies that automate OSH objectives can have adverse effects on workers' overall health and wellbeing (Cabrelli and Graveling 2019; Schulte et al. 2020). An overly mechanical approach to prevention may prove less adaptive to changes in the working environment.

Machine-based prevention systems can, furthermore, lead to a fragmentation of organisational processes and damage the very purpose of occupational safety and health. Even when the Framework Directive provides for the employer to take the measures necessary for the safety of employees and the protection of their health, including the prevention of occupational risks, a reliance on AI systems might be questionable. AI systems depend heavily on the quality of data and the robustness of the predictive AI models they use. Conversely, one of its weakness comes from weak variables, the unpredictability of the external environment and the diversity and unpredictability of the 'human factors' and inputs which, if inadequately taken into account, can put occupational health and safety at risk (Badri et al. 2018; Reiman et al. 2021). AI systems are prone to inheriting inductive bias related to their training data, as well as to ethical and unlawful bias (Brynjolfsson et al. 2023; Hildebrandt 2023). If not carefully designed and monitored continuously, these systems may perpetuate or even exacerbate existing biases and discrimination in the workplace. In the process, OSH risks being transformed from a 'safety practice' into an 'ethical practice' or, worse, a compliance exercise.

Finally, it is imperative to stress the interconnectedness of health, safety, privacy and other human dimensions. Measuring and analysing them in isolation leads to fragmentation, undermining human integrity, dignity and the protection of fundamental rights. Prevention is a practice that simply cannot be automated: the complex context of the workplace environment is what is relevant.

3.3 In data we trust?

When trying to analyse the new phenomenon that is worker surveillance, the analyst keeps coming back to its engine or driving force: data. Worker surveillance is totally reliant on fine-grained personal data and, consequently, on the ability to measure things and to derive actionable conclusions from them. This is a fundamental concern that Sandy J. J. Gould addresses in his chapter in this volume, where he explains that metrics are hard to make and can be noisy, inaccurate and influenced by many factors, resulting in biased, missing information or simply wrong measurement. One example of this is made up of novel safety purposes, such as the monitoring of personal protective equipment compliance through AI, which raise concerns and expose possible limitations such as using an adequate metric, measuring the correct data and having accuracy in pattern recognition and in capturing the essence and subtlety of risk situations (Campero-Jurado et al. 2020).

In short, both data itself and the way it is analysed, which are the core elements of surveillance, should not always be trusted. When worker surveillance takes place, some data is collected through questionable and intrusive means and can be of a highly personal nature. This and other data related to the work environment (temperature, air

quality, machine parameters, noise levels, etc), are then analysed using metrics that are unknown to the workforce and which can generate inaccurate results. In other words, surveillance is built on a weak foundation of personal and sensitive data, intrusively collected and potentially inaccurately analysed, and which is hence likely to generate erroneous outcomes (Nath et al. 2017; Gould, this volume).

4. Towards a legal distinction between monitoring and surveillance

Worker surveillance, as described in previous sections, and the imbalance of power and the information it creates in favour of employers needs to be recognised as a self-standing concept in law and legal literature.

As explained in Section 2, employers have always had the ability to monitor workers as a result of an 'implicit' or 'open' clause in the employment contract. With the possibility of using new, intrusive, invisible and intensive data collection tools, employers can now bring their monitoring power to different levels, thereby establishing a one-sided power relationship with their workers, with little ability for the latter to counterbalance this move given the vast scope of employer oversight and the technical complexities of data science (Sewell and Barker 2001).

The reference to 'surveillance' is not casual. The concept, borrowed from sociology, implies a power-centred understanding of the power to monitor (Macnish 2018; Sewell et al. 2012; Marx 2002). Surveillance studies analyse such social interaction independently of legal categories and as a hierarchical structure with two actors: an observer and an observed. In today's workplace reality, the observer position is further enhanced by third party observers that give even greater monitoring power to the employer. A power-centred understanding of that power in the digital age can help to identify better its new features and actors, and the legal consequences for each of these.

The legal boundaries traditionally imposed on the power to monitor are now being extended in respect of the possibilities of supervision and the culture of the meticulous measurement of workers pushed by the growing market for surveillance technologies (Negrón 2021). The structural difference between traditional worker monitoring and worker surveillance should therefore lead to different legal implications. Worker surveillance has not yet been comprehensively analysed by the ECtHR, which has mainly dealt with traditional monitoring technologies (see Section 2). However, the shaping of such new hierarchies at work systematically presents issues as regards the necessity for these intense and intensive surveillance measures – to be balanced against workers' fundamental rights such as the right to privacy and data protection.

Currently, there is no case law from the ECtHR addressing data-intensive surveillance at work as such. Yet, the Court has clarified that the 'necessary' interference with Article 8 ECHR is a narrowly defined concept: it has none of the flexibility of expressions such as 'useful', 'reasonable' or 'desirable' but implies the existence of a 'pressing social need' for the interference in question (*Handyside v United Kingdom*, 1976, para. 48). Thus,

interference with Article 8 ECHR must correspond to a pressing social need and, in particular, must remain proportionate to the legitimate aim pursued.

Worker surveillance does not conform to a pressing social need under the ECHR. Thus, it can be described as an unnecessary monitoring power. Most of the new surveillance technology is not strictly necessary to achieve a legitimate employer purpose. Gould shows this point well in his chapter in this volume: 'measuring things about work is difficult, especially work that does not lend itself easily to being broken down into independent atomic parts'; this proves to be particularly relevant in services work in comparison with the manufacturing sector. How could a surveillance provider or an employer prove that looking at the time spent on Facebook, or at the number of times someone looks away from their screen, are representative of productivity or focus?

To prevent such innovations distorting employer authority into something intrusive and unjust, it is essential to ponder carefully the existence of a pressing social need under the ECHR. In several decisions (*Florindo*, 2022; *López Ribalda*, 2019; *Bărbulescu*, 2017), the Court has recognised that significant developments are underway in the field of workforce monitoring. In *Florindo*, the ECtHR in para. 93 explicitly refers to that particular form of surveillance coming under the analysis of the Court for the first time: the data in dispute is not images (see *Köpke*, 2010; *López Ribalda*, 2019; *Antović and Mirković*, 2018), electronic messages (see *Bărbulescu*, 2017) or computer files (see *Libert v. France*, 2018), but geolocation data. According to the Strasbourg judges, this novelty raises a topical question of the type and level of surveillance that is acceptable on the part of an employer with regard to its employees, to be counter-balanced against protection of the private life of employees.

Steering the fast-developing market for surveillance tools makes the so-called suitability (or effectiveness) test as described by Gerards (2013: 473) a topical criterion to be implemented in the Strasbourg Court's forthcoming case law. Any interference with an ECHR fundamental right ought to be realised by means which are effectively capable of realising the aims, or the ends, of the interference. This is a challenging analytical exercise for the Court: factual, statistical or empirical information ought to be scrutinised to ascertain the suitability of the measure. In this regard, Gould (this volume) gives more technical and empirical evidence about the actual usefulness of measurements in the context of worker surveillance. Referring to his arguments shows how easily worker surveillance tools fail the suitability test.

We can, however, already point to the dissenting opinion in *Florindo* where, in para. 17, the dissenting judges explicitly mention that the means of surveillance implemented by the employer (through the GPS) had been erroneously indicated by the majority of the Court as absolutely necessary to achieve the end pursued by the employer (monitoring the use of the company car). The dissenting judges also point out that less intrusive means than GPS were available and, moreover, indicated by Comissão Nacional de Proteção de Dados (CNPD; the Portuguese DPA). In this most recent ECtHR case on worker monitoring, therefore, an explicit reference to a suitability or effectiveness test is already present on the part of the dissenting judges (Molè and Mangan 2023). The GPS was capable of providing more data than necessary, causing an actual surplus of

data in the availability of the employer which was not needed to fulfil the purpose of the monitoring.

Via the outcome of this case, it is possible to introduce another test, now urgently needed to filter out unnecessary monitoring power: the least intrusive means test. When evaluating the suitability of the means of surveillance, the Court ought to carry out a factual and empirical assessment of the various means which are contextually available and determine which is the most effective and least intrusive. According to this test, the means to be chosen ought to be the one least harmful from the perspective of the individual rights at stake (Gerards 2013: 481-482). On this point, as explained in Section 3, a general consideration can be drawn. As surveillance is structurally more intrusive than standard monitoring, it is essential that the Court, in future years, undertakes efforts to compare surveillance technologies with others which are less intrusive yet which achieve the same legitimate goal pursued by the employer (for a general discussion on the balancing of ECHR fundamental rights by the ECtHR, see Greer 2004).

5. Conclusion

The power to monitor employees traditionally results from an 'open term' in the employment contract that grants employers a set of observational entitlements, in which privacy rights and economic freedoms often collide. This conflict between fundamental rights is being augmented today by new and more intrusive technologies. Hence, differentiating traditional employer monitoring against the background of today's surveillance technologies highlights the tensions at stake. Borrowing from sociological literature, contemporary monitoring amounts to the surveillance of workers, an activity where the asymmetry of power and information is structurally stronger and based on invisible, intrusive and intensive measurement which is likely to fail a suitability and least intrusive means test under Art. 8 ECHR. Worker surveillance shows features of a disproportionate and unlawful power, predicated on the belief that anything and everything that is measurable should be measured. The interconnectedness and natural nexus between autonomy and privacy (Hendrickx 2018) is broken. Human dignity itself is at stake since, as law professor Nita Farahany says 'there is no existing set of legal rights that protects us from employers scanning the brain or hacking the brain' (Dockser Marcus 2023).

ECtHR case law has been referenced in support of this line of argument with the aim of identifying the legal boundary beyond which worker monitoring becomes illegitimate worker surveillance. Here, a more favourable interpretation of Article 8 ECHR for workers might be suggested. Pointing to surveillance as a self-standing legal concept helps to tame some of the substantial risks of unnecessary and systematic monitoring under which worker autonomy is eroded, the principle of prevention in OSH is abandoned in favour of 'automated prevention' and fundamental rights are violated. All of this is fuelled by personal and sensitive data that is collected intrusively, yet this may be being inaccurately measured. The necessity and proportionality tests should become a legal obligation when introducing worker monitoring technologies. Action is urgently

needed not just to regulate or set limits on surveillance, but to define it legally as a prohibited practice which conflicts with the most essential fundamental rights at work.

References

Adams-Prassl J. (2019) What if your boss was an algorithm? Economic incentives, legal challenges, and the rise of artificial intelligence at work, Comparative Labor Law and Policy Journal, 41 (1), 123.

AI Now Institute (2023) Algorithmic management: Restraining workplace surveillance, 11 April 2023. https://ainowinstitute.org/publication/algorithmic-management

Al Jassmi H., Ahmed S., Philip B., Al Mughairbi F. and Al Ahmad M. (2019) E-happiness physiological indicators of construction workers' productivity: A machine learning approach, Journal of Asian Architecture and Building Engineering, 18 (6), 517–526. https://doi.org/10.1080/13467581.2019.1687090

Agosti C., Bronowicka J., Polidoro A. and Priori G. (2023) Exercising workers' rights in algorithmic management systems: Lessons learned from the Glovo-Foodinho digital labour platform case, Report 2023.11, ETUI. https://www.etui.org/publications/exercising-workers-rights-algorithmic-management-systems

Aloisi A. and De Stefano V. (2022) Essential jobs, remote work and digital surveillance: Addressing the COVID‐19 pandemic panopticon, International Labour Review, 161 (2), 289–314. https://doi.org/10.1111/ilr.12219

Badri A., Boudreau-Trudel B. and Souissi A.S. (2018) Occupational health and safety in the industry 4.0 era: A cause for major concern?, Safety Science, 109, 403–411. https://doi.org/10.1016/j.ssci.2018.06.012

Baiocco S., Fernández Macías E., Rani U. and Pesole A. (2022) The algorithmic management of work and its implications in different contexts, JRC Working Papers Series on Labour, Education and Technology 2022/02, European Commission. https://joint-research-centre.ec.europa.eu/publications/algorithmic-management-work-and-its-implications-different-contexts_en

Bales R.A. and Stone K.V. (2020) The invisible web at work: Artificial intelligence and electronic surveillance in the workplace, Berkeley Journal of Employment and Labor Law, 41 (1). https://doi.org/10.15779/Z380000085

Ball K. (2010) Workplace surveillance: An overview, Labor History, 51 (1), 87–106. https://doi.org/10.1080/00236561003654776

Benlian A. et al. (2022) Algorithmic management: Bright and dark sides, practical implications, and research opportunities, Business and Information Systems Engineering, 64 (6), 825–839. https://doi.org/10.1007/s12599-022-00764-w

Borle P., Reichel K., Niebuhr F. and Voelter-Mahlknecht S. (2021) How are techno-stressors associated with mental health and work outcomes? A systematic review of occupational exposure to information and communication technologies within the technostress model, International Journal of Environmental Research and Public Health, 18 (16), 8673. https://doi.org/10.3390/ijerph18168673

Burnett J.R. and Lisk T.C. (2021) The future of employee engagement: Real-time monitoring and digital tools for engaging a workforce, in Segalla M. (ed.) International perspectives on employee engagement, Routledge, 117–128.

Brynjolfsson E., Li D. and Raymond L.R. (2023) Generative AI at work, Working Paper 31161, National Bureau of Economic Research. https://doi.org/10.3386/w31161

Cabrelli D. and Graveling R. (2019) Health and safety in the workplace of the future, European Parliament. https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/638434/IPOL_BRI(2019)638434_EN.pdf

Campero-Jurado I., Márquez-Sánchez S., Quintanar-Gómez J., Rodríguez S. and Corchado J.M. (2020) Smart helmet 5.0 for industrial Internet of things using artificial intelligence, Sensors, 20 (21). https://doi.org/10.3390/s20216241

Carby-Hall J. (2003) The contractual nature of social law, Managerial Law, 45 (3/4), 23–107. https://doi.org/10.1108/03090550310770893

Castro I. and Ramos D.G. (2017) Understanding the management of occupational health and safety risks through the consultation of workers, in Arezes P.M. et al. (eds.) Occupational safety and hygiene V, CRC Press, 29–34.

Chan N.K. (2019) The rating game: The discipline of Uber's user-generated ratings, Surveillance and Society, 17 (1/2), 183–190. https://doi.org/10.24908/ss.v17i1/2.12911

Cheng B., Fan C., Fu H., Huang J., Chen H. and Luo X. (2022) Measuring and computing cognitive statuses of construction workers based on electroencephalogram: A critical review, IEEE Transactions on Computational Social Systems, 9 (6), 1644–1659. https://doi.org/10.1109/TCSS.2022.3158585

Coase R.H. (1937) The nature of the firm, Economica, 4 (16), 386–405. https://doi.org/10.1111/j.1468-0335.1937.tb00002.x

De Cremer D. and Stollberger J. (2022) Are people analytics dehumanizing your employees?, Harvard Business Review. https://hbr.org/2022/06/are-people-analytics-dehumanizing-your-employees

Dockser Marcus A. (2023) When your boss is tracking your brain, The Wall Street Journal, 15 February 2023. https://www.wsj.com/articles/brain-wave-tracking-privacy-b1bac329?mkt_tok=MTM4LUVaTS0wNDIAAAGJ_MWOqN2F8QfijbjrPEKP26_kBSJFTuoGXkacR02Cb7SKzIMT5hUfJrZaL1cQ51GnM7DlnL8ifJxaQGhubGvH4Hw4djn-dcXaLjaKnOHE10ki

EDPS (2019) Guidelines on assessing the proportionality of measures that limit the fundamental rights to privacy and to the protection of personal data, European Data Protection Supervisor. https://www.edps.europa.eu/sites/default/files/publication/19-12-19_edps_proportionality_guidelines2_en.pdf

EU-OSHA (2007) Factsheet 80 — Risk assessment — roles and responsibilities. https://osha.europa.eu/en/publications/factsheet-80-risk-assessment-roles-and-responsibilities

EU-OSHA (2023) Using AI for task automation while protecting workers: Eight case studies provide new insights. https://osha.europa.eu/en/highlights/using-ai-task-automation-while-protecting-workers-eight-case-studies-provide-new-insights

European Commission (2021) Proposal for a Directive of the European Parliament and the council on improving working conditions in platform work, COM(2021) 762 final, 9.12.2021. https://eures.ec.europa.eu/eu-proposes-directive-protect-rights-platform-workers-2022-03-17_en

European Data Protection Board (2020a) Hamburg commissioner fines H&M 35.3 million euro for data protection violations in service centre. https://edpb.europa.eu/news/national-news/2020/hamburg-commissioner-fines-hm-353-million-euro-data-protection-violations_en

European Data Protection Board (2020b) State commissioner for data protection in lower Saxony imposes € 10.4 million fine against notebooksbilliger.de. https://edpb.europa.eu/

news/national-news/2021/state-commissioner-data-protection-lower-saxony-imposes-eu-104-million-fine_en
European Social Partners (2020) European social partners framework agreement on digitalisation, ETUC. https://www.etuc.org/system/files/document/file2020-06/Final%2022%2006%2020_Agreement%20on%20Digitalisation%202020.pdf
Farahany N.A. (2023) The battle for your brain: Defending the right to think freely in the age of neurotechnology, St. Martin's Press.
Frick K. (2011) Worker influence on voluntary OHS management systems - A review of its ends and means, Safety Science, 49 (7), 974–987. https://doi.org/10.1016/j.ssci.2011.04.007
Gallup (2023) Gallup's employee engagement survey: Ask the right questions with the Q12® survey. https://www.gallup.com/workplace/356063/gallup-q12-employee-engagement-survey.aspx
Gerards J. (2013) How to improve the necessity test of the European Court of Human Rights. International Journal of Constitutional Law, 11 (2), 466–490. https://doi.org/10.1093/icon/mot004
GPDP (2021) Riders: Italian SA says no to algorithms causing discrimination. A platform in the Glovo group fined EUR 2.6 million, Garante per la Protezione dei Dati Personali https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9677377#english
Greer S. (2004) 'Balancing' and the European Court of Human Rights: A contribution to the Habermas-Alexy debate, The Cambridge Law Journal, 63 (2), 412–434. https://doi.org/10.1017/S0008197304006634
Hendrickx F. (2018) From digits to robots: the privacy-autonomy nexus in new labor law machinery, Comparative Labor Law and Policy Journal, 40 (3), 365–388.
Hildebrandt M. (2022) The issue of proxies and choice architectures. Why EU law matters for recommender systems, Frontiers in Artificial Intelligence, 5. https://doi.org/10.3389/frai.2022.789076
Hildebrandt M. (2023) Sustainable software: Issues of bias, proxies and ground truthing in machine learning. https://www.cohubicol.com/assets/uploads/hildebrandt-api-keynote.pdf
ICO (2022) Employment practices: Monitoring at work draft guidance, Information Commissioner's Office. https://ico.org.uk/media/about-the-ico/consultations/4021868/draft-monitoring-at-work-20221011.pdf
Jebelli H., Choi B. and Lee S. (2019) Application of wearable biosensors to construction sites. I: Assessing workers' stress, Journal of Construction Engineering and Management, 145 (12), 04019079. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001729
Laker B., Godley W., Patel C. and Cobb D. (2020) How to monitor remote workers—Ethically, MIT Sloan Management Review. https://sloanreview.mit.edu/article/how-to-monitor-remote-workers-ethically/
Lupton D. (2016) The quantified self: A sociology of self-tracking, Polity.
Macnish K. (2018) The ethics of surveillance: An introduction, Routledge.
Malik A. (2023) Artificial intelligence in health and safety, SafetyPedia. https://safetypedia.com/safety/artificial-intelligence-in-health-and-safety/
Mangan D. (2022) From monitoring of the workplace to surveillance of the workforce, in Gyulavári T. and Menegatti E. (eds.) Decent work in the digital age: European and comparative perspectives, Bloomsbury, 311–329.
Marx G.T. (2002) What's new about the 'new surveillance'? Classifying for change and continuity, Surveillance and Society, 1 (1), 9–29. https://doi.org/10.1016/B978-0-08-097086-8.64025-4

Marx G.T. (2015) Surveillance studies, International Encyclopedia of the Social and Behavioral Sciences, 733–741. https://doi.org/10.1016/B978-0-08-097086-8.64025-4

Menéndez M., Benach J. and Vogel L. (2009) The impact of safety representatives on occupational health: A European perspective, Report 107, ETUI.

Molè M. (2022) The Internet of things and artificial intelligence as workplace supervisors: Explaining and understanding the new surveillance to employees beyond art. 8 ECHR, Italian Labour Law E-Journal, 15 (2), 87–103. https://doi.org/10.6092/ISSN.1561-8048/15598

Molè M. and Mangan D. (2023) 'Just more surveillance': The ECtHR and workplace monitoring, European Labour Law Journal, 14 (4), 694–700. https://doi.org/10.1177/20319525231201274

Moore P.V. (2018) Tracking affective labour for agility in the quantified workplace, Body and Society, 24 (3), 39–67. https://doi.org/10.1177/1357034X18775203

Nath N.D., Akhavian R. and Behzadan A.H. (2017) Ergonomic analysis of construction worker's body postures using wearable mobile sensors, Applied Ergonomics, 62, 107–117. https://doi.org/10.1016/j.apergo.2017.02.007

Negrón W. (2021) Little tech is coming for workers. A framework for reclaiming and building worker power, Coworker.org. https://home.coworker.org/wp-content/uploads/2021/11/Little-Tech-Is-Coming-for-Workers.pdf

Nielsen K.J. (2014) Improving safety culture through the health and safety organization: A case study, Journal of Safety Research, 48, 7–17. https://doi.org/10.1016/j.jsr.2013.10.003

Ollé-Espluga L., Vergara-Duarte M., Belvis F., Menéndez-Fuster M., Jódar P. and Benach J. (2015) What is the impact on occupational health and safety when workers know they have safety representatives?, Safety Science, 74, 55–58. https://doi.org/10.1016/j.ssci.2014.11.022

Ponce del Castillo A. (2020) COVID-19 contact-tracing apps: How to prevent privacy from becoming the next victim, Policy Brief 5/2020, ETUI. https://www.etui.org/publications/policy-briefs/european-economic-employment-and-social-policy/covid-19-contact-tracing-apps-how-to-prevent-privacy-from-becoming-the-next-victim .

Reiman A., Kaivo-Oja J., Parviainen E., Takala E.-P. and Lauraeus T. (2021) Human factors and ergonomics in manufacturing in the industry 4.0 context–A scoping review, Technology in Society, 65, 101572. https://doi.org/10.1016/j.techsoc.2021.101572

Riso S. (2020) Employee monitoring and surveillance: The challenges of digitalisation, Publications Office of the European Union. https://www.eurofound.europa.eu/en/publications/2020/employee-monitoring-and-surveillance-challenges-digitalisation

Rogers B. (2023) Workplace data is a tool of class warfare, Boston Review. https://www.bostonreview.net/articles/workplace-data-is-a-tool-of-class-warfare/

Schulte P.A. et al. (2020) Potential scenarios and hazards in the work of the future: A systematic review of the peer-reviewed and gray literatures, Annals of Work Exposures and Health, 64 (8), 786–816. https://doi.org/10.1093/annweh/wxaa051

Sewell G. and Barker J.R. (2001) Neither good, nor bad, but dangerous: Surveillance as an ethical paradox, Ethics and Information Technology, 3 (3), 181–194. https://doi.org/10.1023/A:1012231730405

Sewell G., Barker J.R. and Nyberg D. (2012) Working under intensive surveillance: When does 'measuring everything that moves' become intolerable?, Human Relations, 65 (2), 189–215. https://doi.org/10.1177/0018726711428958

Stark D. and Pais I. (2020) Algorithmic management in the platform economy, Sociologica, 14 (3), 47–72. https://doi.org/10.6092/issn.1971-8853/12221

Swan M. (2013) The quantified self: Fundamental disruption in big data science and biological discovery, Big Data, 1 (2), 85–99. https://doi.org/10.1089/big.2012.0002

van Rijmenam M. (2023) The transformative role of AI in revolutionising workplace health and safety, The Digital Speaker, 3 August 2023. https://www.thedigitalspeaker.com/transformative-role-ai-revolutionising-workplace-health-safety/#:~:text=Predictive%20modelling%20uses%20sophisticated%20algorithms,measures%20and%20protect%20their%20workforce

Vitard A. (2023) Amazon risque une amende de 170 millions d'euros pour sa gestion des données de productivité des salariés, L'Usine Digitale, 18 Sptember 2023. https://www.usine-digitale.fr/article/amazon-risque-une-amende-de-170-millions-d-euros-pour-sa-gestion-des-donnees-de-productivite-des-salaries.N2171902

Wadsworth E. and Walters D. (2019) Safety and health at the heart of the future of work: Building on 100 years of experience, ILO. https://www.ilo.org/safework/events/safeday/WCMS_687610/lang--en/index.htm

Wang D., Chen J., ZhaoD., Dai F., Zheng C. and Wu X. (2017) Monitoring workers' attention and vigilance in construction activities through a wireless and wearable electroencephalography system, Automation in Construction, 82, 122–137. https://doi.org/10.1016/j.autcon.2017.02.001

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

Case law

— *Antović and Mirković v. Montenegro* (European Court of Human Rights 28 February 2018). https://hudoc.echr.coe.int/fre?i=001-178904
— Article 29 Data Protection Working Party (2017). Opinion 2/2017 on Data Processing at Work
— *Bărbulescu v. Romania*, Application no. 61496/08 (European Court of Human Rights 2017). https://hudoc.echr.coe.int/fre?i=001-177082
— *Copland v. United Kingdom* (European Court of Human Rights 3 April 2007).
— *Florindo de Almeida Vasconcelos Gramaxo v. Portugal* (European Court of Human Rights 13 December 2022). https://hudoc.echr.coe.int/fre?i=002-13935
— *Halford v. UK* (European Court of Human Rights 25 June 1997).
— *Handyside v United Kingdom*, 5493/72 (European Court of Human Rights 1976).
— *Köpke v. Germany* (European Court of Human Rights 5 October 2010).
— *Libert v. France* (European Court of Human Rights 22 February 2018). https://hudoc.echr.coe.int/rus/#{%22itemid%22:[%22001-181273%22]}
— *López Ribalda and Others v. Spain* (European Court of Human Rights 17 October 2019). https://hudoc.echr.coe.int/fre?i=002-12630

All links were checked on 14.02.2024.

Cite this chapter: Ponce Del Castillo A. and Molè M. (2024) Worker monitoring vs worker surveillance: the need for a legal differentiation, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 14
Affective computing at work: rationales for regulating emotion attribution and manipulation

Frank Pasquale

1. Introduction

Affective computing attempts to read, simulate, predict and stimulate human emotion with software. Advocates for affective computing in the workplace claim that it can improve efficiency, identify better and worse work styles and indicate how engaged employees are. While these are understandable aspirations, affective computing also raises many policy concerns, especially when it enters workplaces already highly influenced by algorithmic management.[1]

Affective computing has become a popular computational and psychological research programme. Teams are now programming robots, chatbots and animations to appear to express sadness, empathy, curiosity and much more. Automated face analysis is translating countless images of human expressions into standardised code that elicits responses from machines.

However, as affective computing is adopted in the workplace, it will increasingly judge us and try to manipulate us. At least four concerns about algorithmic management via affective computing are described below: misrecognition; privacy invasion; modulation; and alienation.

2. The four major concerns

2.1 Misrecognition

First, we should not assume that affective computing technologies will work as planned. At the most basic level, they may misrecognise people and attribute one person's action to another. Even when they can consistently recognise persons and faces, the machines may fail. Psychology researchers have demonstrated that faces and expressions do not necessarily map neatly to particular traits and emotions, let alone to the broader mental states evoked in engagement or aggression detection. As Lisa Barrett and her colleagues report: 'instances of the same emotion category are neither reliably expressed through nor perceived from a common set of facial movements' (Barrett et al. 2019: 3), and thus the communicative capacities of the face are limited. The dangers of misinterpretation are clear and present in efforts to quantify engagement via face analysis.

1. For background on the regulation of algorithmic management, see Ajunwa (2023); Aloisi and De Stefano (2022); Levy (2022); and Rogers (2023).

Bias is endemic, too, and may be exacerbated by affective computing. For example, Lauren Rhue has found that 'black men's facial expressions are scored with emotions associated with threatening behaviours more often than white men, even when they are smiling' (Rhue 2019). Her work has also revealed that 'AIs display racial disparities in their emotional scores and are more likely to assign negative emotion to black men's faces' (Rhue 2018: 6).

Sampling problems are also likely to be rife. If a database of exemplary emotional engagement among service workers is developed from observation of a particular subset of the population (say, a managerial corps dominated by men), the resulting AI may be far better at finding exemplary behaviour in the future in that subset (men) than in others. Robotic human resources departments of the future may be 'machine learning' from data distorted by a discrimination-ridden past.

2.2 Privacy violations

Law professor Jennifer Bard (2021) has raised several important questions about the potential privacy violations of emotion-oriented AI. Depending on the way in which emotions are detected or attributed, Bard argues, they may be akin to thoughts, beliefs or biological aspects of the body. Having one's thoughts read, or having thoughts attributed to oneself, raises deep privacy concerns. Mental and physical health data are among the most protected, sensitive data in many privacy regimes. Similarly, freedom of belief includes the right to keep one's beliefs to oneself. Reputational integrity may depend on the right to avoid being attributed beliefs without consent, especially if that attribution of belief is socially constructed as indubitable, or is done in a secretive manner. Bard's biometric angle on privacy is a particularly promising one for regulation, as several jurisdictions already regulate the collection, sale and use of biometric data.

2.3 Modulation

Bard has also presciently foreseen concerns about the excessive control of people by emotion machines. If an affective computing system is too effective at creating feelings of shame or merit, ostracism or pleasure, it may exercise undue influence. The most obvious current examples of modulation come from the political and commercial realms, where advertisements precisely keyed to a certain emotional register are aimed at creating affective resonances between consumers and brands or between voters and politicians.

Workplace modulation is likely to be subtler, a consequence of gamification and efforts to cultivate forms of pseudo-friendship between workers and apps that constantly bill themselves as trying to be helpful, cheer the worker on, and so on. Apps already prod workers to devote more time to tasks via varied interfaces (Calo and Rosenblat 2017; Dubal 2023). Affective computing may make present forms of gamification more influential.

More forceful interventions may also be supported via affective computing. The manipulation of inchoate emotions (common in alexithymic individuals) toward fear or anxiety may even further empower surveillance campaigns to undermine unionisation and related efforts. As Richard Bales and Katherine Stone powerfully conclude:

> AI and electronic monitoring produce an invisible electronic web that threatens to invade worker privacy, deter unionization, enable subtle forms of employer blackballing, exacerbate employment discrimination, render unions ineffective, and obliterate the protections of the labor laws.' (Bales and Stone 2020: 1)

Thus precision stimulation of emotions raises special concerns in contexts of conflict between labor and capital.

2.4 Alienation

The 'app-as-friend' framing pursued by firms is bound to be rejected by many. But this distancing from affective computing's failed performances of assistance may raise its own difficulties. Alienation is a dispiriting sense of meaninglessness and powerlessness which can easily curdle into withdrawal or depression. If pervasively unsuccessful attempts at automated emotional connection result in what Rahel Jaeggi (2014) calls a 'relation of relationlessness', the stage is set for a radical degradation of the workplace from a site of some forms of education, solidarity and assistance into a single, more computationally administered environment. The most successful authoritarian populists of our time have exploited widespread alienation, channelling the resulting discontent into destructive political programmes.

3. Conclusion

The critiques above are interrelated and suggest the ideal-typical problems that arise when affective computing either fails to work or works too well:

	Fails to work	Works too well
Emotion attribution	Misrecognition	Privacy invasion
Emotion manipulation	Alienation	Modulation

This is a dispiriting matrix, but recognising the problems posed by affective computing in the workplace is a first step toward addressing them. Precise characterisation matters. Many affective computing programs are billed as 'emotion recognition', akin to facial recognition, suggesting that the identification of mental states is as indisputable and straightforward as identifying a person. However, 'emotion attribution' is a more precise term than 'emotion recognition', underscoring the contestable nature of the 'knowledge' produced by affective computing. Similarly, 'emotion manipulation' is

a term that properly applies a hermeneutics of suspicion in a scenario inadequately described as mere 'emotional stimulation'.[2]

Regulation of affective computing must be multi-layered. The same underlying data and computing systems may power judgments in many sectors of economy and society. These data and algorithms should be subject to extensive transparency requirements. The deployment and application of affective computing in particular contexts should be tightly monitored and regulated. For example, states may forbid employers from using emotion attribution in hiring contexts, or set strict conditions meant to address the concerns raised above.

In far too many workplaces, affective computing's corporate deployments will be less about providing services to employees than about shaping them in manipulative ways. Preserving the privacy and autonomy of our emotional lives should take priority over a misguided and manipulative quest for emotion machines. The procurement of such systems should be contingent on worker consultation and negotiation. Moreover, to avoid potential 'races to the bottom' in terms of the usage and adoption of the surveillance inherent in many forms of affective computing, certain non-waivable rules for its deployment should be enforced. Respect for intellectual and emotional autonomy needs to be at the core of affective computing law and policy.

References

Ajunwa I. (2023) The quantified worker: Law and technology in the modern workplace, Cambridge University Press. https://doi.org/10.1017/9781316888681

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

Bales R.A. and Stone K.V. (2020) The invisible web at work: Artificial intelligence and electronic surveillance in the workplace, Berkeley Journal of Employment and Labor Law, 41 (1), 1–60. https://doi.org/10.15779/Z380000085

Bard J.S. (2021) Developing legal framework for regulating emotion AI, Boston University Journal of Science and Technology Law, 27 (2), 271–311. https://www.bu.edu/jostl/archives/volume-27-2-summer-2021/

Barrett L.F., Adolphs R., Marsella S., Martinez A.M. and Pollak S.D. (2019) Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements, Psychological Science in the Public Interest, 20 (1), 1–68. https://doi.org/10.1177/1529100619832930

Calo R. and Rosenblat A. (2017) The taking economy: Uber, information, and power, 117 (6), Columbia Law Review. https://columbialawreview.org/content/the-taking-economy-uber-information-and-power/

Dow Schüll N. (2014) Addiction by design: Machine gambling in Las Vegas, Princeton University Press.

2. Stimulation also has problematic resonances with behaviourism. While a profitable theory of mind for commercial actors to deploy, it is only one very limited way of understanding mental life. On the role of behaviouristic models in commercial life, see Dow Schüll (2014).

Dubal V. (2023) On algorithmic wage discrimination, UC San Francisco Research Paper. http://dx.doi.org/10.2139/ssrn.4331080
Jaeggi R. (2014) Alienation, Columbia University Press. https://doi.org/10.7312/jaeg15198
Levy K. (2022) Data driven: Truckers, technology, and the new workplace surveillance, Princeton University Press.
Rhue L. (2018) Racial influence on automated perceptions of emotions. http://dx.doi.org/10.2139/ssrn.3281765
Rhue L. (2019) Emotion-reading tech fails the racial bias test, The Conversation, 3 January 2019. https://theconversation.com/emotion-reading-tech-fails-the-racial-bias-test-108404
Rogers B. (2023) Data and democracy at work: Advanced information technologies, labor law, and the new working class, MIT Press. https://doi.org/10.7551/mitpress/11253.001.0001

All links were checked on 19.02.2024.

Cite this chapter: Pasquale F. (2024) Affective computing at work: rationales for regulating emotion attribution and manipulation, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Part 5

Labour perspectives

Chapter 15
AI for good work

Frank Pot

1. Introduction. Facing transitions

We are in the middle of major discussions about the opportunities and threats of artificial intelligence. Besides economic benefits and prosperity we want good jobs. Research indicates that AI does not automatically lead to good jobs nor to the disappearance of bad jobs: the outcome depends on organisational design and management regimes on the one hand and employee participation in decision-making on the other. This is the struggle for organisational control.

The concept of good jobs means more than wages and permanent contracts; it's also about work content and labour relations. Theory-based criteria and design approaches for good work are available, and policies ought to be following the European Pillar of Social Rights Action Plan, in which the European Commission encourages national authorities and the social partners to foster workplace innovation.

Europe is engaged in a digital transition, increasingly connected to the green transition. The ambition is that, in addition to economic and climate goals, these transitions also generate good jobs and that no-one is left behind. This is neatly formulated by the European Commission in its statement on Industry 5.0:

> Industry 5.0 is characterised by going beyond producing goods and services for profit. It shifts the focus from the shareholder value to stakeholder value and reinforces the role and the contribution of industry to society. It places the wellbeing of the worker at the centre of the production process and uses new technologies to provide prosperity beyond jobs and growth while respecting the production limits of the planet. (European Commission 2021a)

However, the market mechanism does not provide good jobs by itself. Rodrik and Sabel (2019) describe a 'massive market failure' to create 'a good jobs economy', one example being that the number of workers with monotonous repetitive tasks did not decrease between 2005 and 2015.

Table 1 **Does your job involve short repetitive tasks of less than 1 minute? Does your job involve short repetitive tasks of less than 10 minutes?**

European Working Conditions Surveys	No short repetitive tasks	Between 1 and 10 minutes	Less than 1 minute	Total
2005	54.4%	25.3%	20.2%	100.0%
2010	51.2%	25.5%	23.4%	100.0%
2015	53.8%	24.6%	21.6%	100.0%

Source: Eurofound, *European Working Conditions Surveys* (in Pot 2022).

Of course, some routine tasks have been replaced by automation, robots or AI, but German research shows that new repetitive tasks have emerged in their place (Ittermann and Virgillito 2019; Lager 2019; Lager et al. 2021). One example is the expansion of the number of warehouses and new technologies such as headphones (audio picking) and Google Glass (vision picking) that lead to higher productivity but also shorter tasks and task intensification; another is Amazon Mechanical Turk that offers workers the freedom to complete very short menial tasks such as recognising and labelling images, paid as little as $0.01 each. Ironically these tasks are called 'human intelligence tasks' because machines cannot do them. The data produced as a result of this activity is necessary to feed, or train, AI systems. The estimated number of workers involved in this in 2013 was 580,000 (Kuek et al. 2015: 19); because these online workers are invisible, their activities are sometimes called 'ghost work' (Casilli 2016).

To manage the consequences of the digital transition, we need to have a good understanding of the technical and organisational alternatives and the balance of power involved in organisational design.

2. Balance of power and organisational control

The employment relationship is not just the legal link between employers and employees regarding work or services carried out in return for remuneration and the presence of reciprocal rights and obligations between the employee and the employer. Marx has extensively theoretically explained that technology and organisation play an important role in the struggle over the combination of working time (hours, minutes, breaks) and the intensity of work (effort per hour) in relation to labour productivity and pay (Marx 1887: Chapter 15).

We see this reflected in F.W. Taylor's design theory called 'scientific management' which was supposed to lead to the optimum performance of man and machine, maximising prosperity for the employer and the workers. The new organisation was characterised by a separation of mental and manual labour, the introduction of a managerial system and the splitting of tasks alongside further mechanisation and piecerate wages (Taylor 1911). For this purpose, time studies were used and, later, also motion studies (Gilbreth and Gilbreth 1917) in what we would now call a data-driven approach. In this way, every worker could perform and earn to the maximum without excessive effort.

Taylor recognised the struggle for organisational control. As a result of scientific management, workers' practice of 'going slow' (to prevent an increase in the pace and a reduction in the rate) would no longer be possible and trade unions would become superfluous (Taylor 1911). Taylor only saw opportunities. However, the theory failed in practice to fulfil all the promises. Negative consequences included de-skilling and intensification, as well as risks to health and safety. Trade unions became important, while employers also recognised the usefulness of collective agreements and governments introduced laws on labour and social protection.

Organisational control on the part of management can take different forms: 'command and control', or 'participation and trust'. A control regime that seems to be somewhere in between is what Doorewaard calls 'management by seduction', a hegemonic form of control embedded in the rules and structure of modern factories and offices which, broadly speaking, are accepted by all the parties. In a self-evident way, they bring about a social practice in which an unequal chance of realising interests and/or wishes arises and is maintained (Doorewaard 1989). Informal behaviours can differ as well, for example respect or intimidation. Management, based on algorithms without human intervention, can also include automatic decisions about ratings, rewards and penalties, as we know from the Uber app, in a new form of social domination (Nicklich and Pfeiffer 2023).

Organisational control on the part of workers can also take different forms: task autonomy and skill discretion, autonomous teams and shopfloor consultancy, co-determination and collective bargaining or collective action such as strikes. Informal behaviour can either reflect a desire to follow the rules or to try to avoid them, to be proactive or to go slow, and sometimes sabotage. In the particular context of AI, this boils down to the question of how to fool the algorithm. One example is the 'timed collective logouts by couriers in the twenty-first century that are mirroring the stopping of machines in the twentieth century' (Vandaele 2021: 227). Chase Thiel and colleagues (2023) theorise that monitoring paradoxically creates the conditions for more (not less) deviance by diminishing employees' sense of agency, thereby facilitating moral disengagement via the displacement of responsibility.

Although the social context has changed considerably, the struggle for organisational control is still ongoing. Perhaps the application of AI will mark the beginning of a new phase of this struggle. The scope for action is vast: Katherine Kellogg and colleagues (2020) point out that employers can use algorithms to direct workers by restricting and recommending; evaluate them by recording and rating; and discipline them by rewarding and replacing.

3. Technological determinism or organisational choice

It is often thought that the appearance of jobs and tasks is determined by technology and by economic factors (efficiency, productivity). However, how work is organised also appears to depend on the chosen management style. This has recently been substantiated with research from the United States. Management practices have at least

as much impact on productivity as new technology (R&D and IT) (Bloom et al. 2019), but they are very different and such differences are difficult to explain. Management styles that are not economically optimal often lead a tough existence. The leadership can opt either for 'command and control' or 'participation and trust', and that choice is not primarily determined by technology or economics. In the organisational sciences, this relativisation of technological and economic determinism has led to the use of the term 'organisational choice'. In theory, this also gives room for employees to have a say in the organisation of work and technology. For example, research around 1990 showed that the robotisation of arc welding can lead to task splitting as well as task integration (Benders 1993). An even simpler example from the present day is that of the Koninklijke Gazelle (Royal Gazelle) bicycle factory in Dieren (in the Netherlands) which has organised assembly work in such a way that workers perform tasks of a maximum of 90 seconds, whereas at the Koga bicycle factory in Heerenveen an operator assembles the entire bicycle. In principle, both factories have the same technology at their disposal. Gazelle claims 'world class manufacturing' but has designed tasks that are an affront to human dignity and that do not comply with European and Dutch legislation on monotonous and timebound work (Pot 2016).

The same argument holds for the determination of skills. Steven Dhondt and his team investigated changes in technology, work organisation and skills over time. The results show technological change to have small effects on changing skills use in contrast to the larger effects stemming from changes in the organisation of work (Dhondt et al. 2022). This conclusion has also been drawn by David Autor and colleagues:

> To make use of the strengths and limitations of machine learning, organizations will need to redesign workflow and rethink the division of tasks between workers and machines, akin to what occurred as Amazon deployed robotics in its warehouses. The resulting changes in work design will alter the nature of many jobs, in some cases profoundly. But the implications for specific skill groups are as yet uncertain and will in part depend on managerial and organizational choices, not on technologies alone. (Autor et al. 2019: 32)

4. Research on AI: mixed outcomes

Focusing on AI, the same conclusions (about organisational control and organisational choice) can be drawn. Daron Acemoglu and Pascual Restrepo (2019) point out that artificial intelligence is now mainly used to automate labour, resulting in unemployment and little or no improvement in productivity; whereas it is also possible to use AI to create new highly productive forms of labour with a decent quality of work, which would be better for people and for the economy.

Empirical research confirms that the application of AI can have different effects on job quality. Danish research shows that AI may enhance or augment skills through, for example, the increased use of high-performance work practices, or it may raise the constraints on the pace of work and reduce employee autonomy, understood as the exercise of control over one's work methods and pace (Holm and Lorenz 2022). From

11 case studies across Europe on combined automation and AI systems, Eva Heinold and her team (2023) find, in most cases, work that is less dirty, dull and dangerous in terms of job content while embodying more creative, challenging and cognitive tasks.

According to Alex Wood (2021), the existing evidence suggests that algorithmic management may accelerate and expand precarious fissured employment relations (via outsourcing, franchising, temporary work agencies, labour brokers and digital labour platforms). It may also worsen working conditions by increasing standardisation and by reducing opportunities for discretion and intrinsic skill use. Evidence from platform work and logistics highlights the danger of algorithmic management in intensifying work effort, creating new sources of algorithmic insecurity but also fuelling workplace resistance. Indeed, there may be both positive and negative outcomes for workers, depending on management regime (Kellogg et al. 2020; Poba-Nzaou et al. 2021). Raquel Kessinger (2021) demonstrates how managers in a digital marketing agency softened the edges of algorithmic evaluation by engaging in relational work with employees who were subject to algorithmic recording, in the process reducing worker stress and encouraging learning. She calls this management regime 'orchestrating friendship' reflecting, as described in a previous paragraph, a hegemonic form of control or 'management by seduction'.

Pierre Bérastégui (2021) argues that algorithmic management leads to high job standardisation due to more predictive patterns in the delivery of work and permanent digital surveillance. Platforms are the primary beneficiaries of such practices as they are able to exercise greater control over the terms of the exchange. Platform workers, on the other hand, are left with very little discretion or latitude in the way they perform their duties. This entails, among other things, psychosocial risks. The case of Amazon shows that permanent surveillance not only controls the performance of workers but also their behaviours by countering their attempts at organisational control and curtailing their trade union activities (UNI Global Union 2021). Furthermore, we know that many recruitment algorithms unintentionally discriminate against particular groups (Burt 2020).

Another point of contention is the use of AI for increasing occupational safety, some examples of which are known as predictive-based safety, with applications growing in terms of detection and warnings in workplaces and of the use of big data in accidentology and epidemiology. For example, facial recognition may be used to check whether workers are wearing the correct safety equipment. But even then, it has been observed that this can lead to the assessment and disciplining of employees, resulting in workplace stress and mental health problems (Moore and Starren 2019; Zoomer et al. 2022; INRS 2023). It turns out to be difficult to experience the advantages of predictive-based safety without the disadvantages of digital control.

Quite a large body of research shows the potential for negative effects in the course of which it could almost be forgotten that AI may also bring about positive innovations in products, services and processes. The benefits for doctors, teachers and judges are also evident where AI supports them to work in a more precise and better informed way. At the same time, recent research shows that there are significant impacts and risks to the

teaching/educating profession such as the solving of tasks by students through various AI-based applications like ChatGPT (Ghita and Stan 2022). Teachers' professional organisations are already complaining of the effects on teachers of the increasingly opaque use of AI (Onderwijsraad 2022).

According to the OECD, artificial intelligence has made significant progress in areas like information ordering, memorisation, perceptual speed and deductive reasoning – all of which are related to non-routine, cognitive tasks. As a result, the occupations that have been most exposed to advances of AI are mostly those in which computer use is high, such as in highly skilled, white collar areas including amongst business professionals, managers, science and engineering professionals, and legal, social and cultural professionals (OECD 2021). The latest variants of AI are generative pre-trained transformer (GPT) models the introduction of which may see approximately 80% of the US workforce having at least 10% of their work tasks affected while around 19% of workers may see an impact on at least 50% of their tasks (Eloundou et al. 2023). The influence of these models spans all wage levels, with higher-income jobs potentially facing greater exposure. Further research will be necessary to explore the broader implications of GPT advances, including their potential to augment or displace human labour as well as their various impacts on job quality, inequality, skills development and numerous other outcomes.

5. Criteria for good work

If we want high quality jobs based on the 'human-in-control' principle, what kind of criteria can be used? Above all, they should refer to the objective characteristics of work tasks. Subjective measurements (job satisfaction, meaningful work, etc.) are important but not sufficiently so to ensure decent work that is compliant with the law. After all, we know that how people evaluate their work partly reflects their socioeconomic position, their work history and the opportunities they see, or do not see, in the future (Both-Nwabuwe et al. 2017). Furthermore, the criteria for job quality should be distinguished from the consequences of job quality such as learning, stress, wellbeing and innovative behaviour.

In debates on transition and good work, the emphasis is on terms of employment (wages, contracts) and occupational safety and health; work content and labour relations receive rather less attention. That is why the focus in the criteria set out in Box 1 below is mainly on the latter issues.

These criteria are drawn from legislation as well as scientific theories and research. Theories about the quality of work tasks are: job demands-control-support (Karasek and Theorell 1990); the job demands-resources model (Bakker and Demerouti 2007); action regulation theory on complete jobs (Hacker 1986, 2003); conditions for wellbeing at work (Pot et al. 1994; Pot 2017); and self-determination theory (Deci et al. 2017). A number have already been included in guidelines on psychosocial risks, for example the Psychosocial Risk Management Excellence Framework (EU PRIMA-EF) in which national institutes, as well as the International Labour Office and the World Health

Organisation, have been involved (Leka and Cox 2008). Job content criteria are also covered in ISO 45003 'Occupational health and safety management' (2021). Many of the criteria mentioned in these guidelines are regular items in surveys such as the European Working Conditions Survey.

Box 1 There is good work if

Terms of employment:

- the contract offers job security
- the work provides a living wage
- the pay system is transparent and fair
- workers have decision-making authority regarding working times and taking leave and holidays
- workers have the opportunity to receive extra training and education

Job content:

- the job consists not only of executive tasks but also of preparation and support tasks. If that is the case it is called a 'complete job' (supporting tasks could be maintenance or quality control)
- difficult and easy tasks are balanced in the job
- there is autonomy regarding work pace, the order of work and the way of working
- the work is not monotonous or repetitive
- enough and timely information and feedback is given about one's own (team) work
- the support of colleagues and line management can be asked for easily
- workers have insight into the algorithms used

Working environment:

- preventive measures and – where necessary – protective measures have been implemented so that workers may work safely and in a healthy way
- the workplace of individual workers is not isolated and there are opportunities for contact

Internal labour relations:

- enough and timely information is given about the strategy and the results of the entire organisation
- workers in shopfloor consultation can participate in decisions regarding (new) processes and the division of tasks and targets – 'organisational tasks'
- there is legal employee representation
- measures have been taken to prevent bullying, sexual harassment, discrimination and violence from colleagues/customers/clients
- the treatment is respectful
- there is no 'real-time' (digital) control of performance and movements
- agreements have been made about the collection and protection of worker data (GDPR)
- workers do not have to respond to messages outside working hours (there is a right to disconnect)

6. Beyond policies and regulations

Of course, new legislation on AI and labour law reform is necessary and several initiatives at European and national level are underway (Ponce Del Castillo and Naranjo 2022). However, for organisational control and organisational choice, hard regulation can be supportive but it is neither sufficient nor particularly effective. In some situations, the joint actions of the social partners and governments provide better opportunities including, for instance, in national research and implementation programmes on workplace development, employee-driven innovation and innovative work organisation (Alasoini 2016; Oeij et al. 2017; Pot et al. 2023). The European Commission refers to this area in the European Pillar of Social Rights Action Plan:

> Social dialogue, information, consultation and participation of workers and their representatives at different levels (including company and sectoral level) play an important role in shaping economic transitions and fostering workplace innovation, in particular with a view to the ongoing twin transitions (digital and green) and the changes in the world of work. (European Commission 2021b: 16)

The European Workplace Innovation Network (EUWIN) (2021) describes workplace innovation as new and combined interventions in work organisation, human resource management, labour relations and supportive technologies. The term embodies a participatory process of innovation which leads to workplace practices that are empowering and which sustain continuing learning, reflection and innovation. This approach applies the good work criteria and leads to higher labour productivity and a stronger innovative capacity within the organisation. Recent empirical support can be found in the European Company Survey, based on interviews with managers, which shows that companies with high job quality and high employee involvement have the best scores on employee wellbeing as well as the best organisational performance (Eurofound and Cedefop 2020).

Another way of moving forward is agreements between the Social Partners on how to tackle the digital and green transitions. One example is the Joint Declaration on Artificial Intelligence of the Telecom Social Dialogue Committee of UNI Europa ICTS and ETNO (2020). Both parties favour a 'human-in-control' approach to AI, meaning that humans should remain in control. They also firmly support respect for human rights as a cornerstone value in the use of all AI technology. AI and other emerging technologies should indeed not hinder individual wellbeing but help build a sustainable and inclusive society. Another example is the European Social Partners Framework Agreement on Digitalisation (2020) which also covers work organisation, work content and skills, working conditions and work relations. These agreements reflect a positive approach to the struggle for organisational control.

Collective bargaining is certainly a promising way to regulate the labour market and terms of employment in those sectors where AI has become important (Vandaele 2021; Lamannis 2023). Where collective agreements can be reached, they must be applied in organisations through co-determination and direct participation. This presupposes a management regime based on participation and trust and an awareness of organisational choice.

Where the conditions for collective bargaining do not yet exist, two factors are considered to be key to understanding the mobilisation processes of, for instance, Amazon Mechanical Turk workers and food delivery couriers: the development of specific communities where these workers could meet and share similar concerns; and particular traditions of political activism on which they could draw to organise their collective action. These communities help build a sense of solidarity and identification while the local traditions provide political scripts and resources as well as the self-confidence needed to transform solidarity into action. Both factors together are facilitating the emergence of a new kind of 'associational power', as an alternative to traditional trade unions (Cini 2023).

Bernd Waas (2022: 202) concludes in his working paper on AI and labour law that 'It can be said that the idea of co-determination has not only lost none of its importance, but that securing sufficient co-determination in the era of AI and Big Data seems more urgent than ever.' For example, in Germany, recent amendments in the Betriebsverfassungsgesetz (Works Constitution Act) were accepted in 2022 in which co-determination on AI systems has been added.

7. Conclusion: the continued relevance of 'human-in-control'

The direct participation of workers in the processes of technological and organisational innovation is even more important for designing good work on the 'human-in-control' principle. Workplace innovation provides such an approach and so do several others: quick response manufacturing (Suri 2010), sociotechnical systems design (Mohr and Van Amelsvoort 2016), relational coordination (Hoffer Gittell 2016) and human-centred design (Parker and Grote 2019, 2022). However, not all approaches that promise good work can be trusted. For example, 'lean' has many variants, not all of which turn out to be good for the quality of work (Huo and Boxall 2018). A critical attitude remains necessary to continue the focus on 'human-in-control' and on placing 'the wellbeing of the worker at the centre of the production process'. Ultimately, however, the central concern around the implementation of AI systems in the workplace is the establishment and development of democracy at work.

References

Acemoglu D. and Restrepo P. (2019) The wrong kind of AI? Artificial intelligence and the future of labor demand, Working Paper 25682, National Bureau of Economic Research. https://doi.org/10.3386/w25682

Alasoini T. (2016) Workplace development programmes as institutional entrepreneurs: Why they produce change and why they do not, Doctoral Dissertation 12/2016, Aalto University.

Autor D., Mindell D.A. and Reynolds E.B. (2019) The work of the future: Shaping technology and institutions, Fall 2019 Report, Massachusetts Institute of Technology.

Bakker A.B. and Demerouti E. (2007) The job demands-resources model: State of the art, Journal of Managerial Psychology, 22 (3), 309–328. https://doi.org/10.1108/02683940710733115

Benders J. (1993) Jobs around automated machines, Dissertation, Radboud University Nijmegen. http://hdl.handle.net/2066/146345
Bérastégui P. (2021) Exposure to psychosocial risk factors in the gig economy: A systematic review, Report 2021.01, ETUI. https://www.etui.org/publications/exposure-psychosocial-risk-factors-gig-economy
Bloom N., Brynjolfsson E., Foster L., Jarmin R., Patnaik M., Saporta-Eksten I. and Van Reenen J. (2019) What drives differences in management practices?, American Economic Review, 109 (5), 1648–1683. https://doi.org/10.1257/aer.20170491
Both-Nwabuwe J.M.C., Dijkstra M.T.M. and Beersma B. (2017) Sweeping the floor or putting a man on the moon: How to define and measure meaningful work, Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01658
Burt A. (2020) How to fight discrimination in AI?, Harvard Business Review, 28 August 2020. https://hbr.org/2020/08/how-to-fight-discrimination-in-ai
Casilli A. (2016) Is there a global digital labor culture? Marginalization of work, Global Inequalities, and Coloniality, Paper presented at the 2nd symposium of the Project for Advanced Research in Global Communication (PARGC), April 2016, Philadelphia, United States. https://shs.hal.science/halshs-01387649
Cini L. (2023) Resisting algorithmic control: Understanding the rise and variety of platform worker mobilisations, New Technology, Work and Employment, 38 (1), 125–144. https://doi.org/10.1111/ntwe.12257
Deci E.L., Olafsen A.H. and Ryan R.M. (2017) Self-determination theory in work organizations: The state of a science, Annual Review of Organizational Psychology and Organizational Behavior, 4, 19–43. https://doi.org/10.1146/annurev-orgpsych-032516-113108
Dhondt S., Kraan K.O. and Bal M. (2022) Organisation, technological change and skills use over time: A longitudinal study on linked employee surveys, New Technology, Work and Employment, 37 (3), 343–362. https://doi.org/10.1111/ntwe.12227
Doorewaard H. (1989) De vanzelfsprekende macht van het management: Een verkennend onderzoek naar hegemoniale aspecten van de macht van het management bij, Van Gorcum.
Eloundou T., Manning S, Mishkin P. and Rock D. (2023) GPTs are GPTs: An early look at the labor market impact potential of large language models, Cornell University. https://doi.org/10.48550/arXiv.2303.10130
Eurofound and Cedefop (2020) European Company Survey 2019: Workplace practices unlocking employee potential, Publications Office of the European Union. http://eurofound.link/ef20001
European Commission (2021a) Industry 5.0: Human-centric, sustainable and resilient, Publications Office of the European Union. https://op.europa.eu/en/publication-detail/-/publication/aed3280d-70fe-11eb-9ac9-01aa75ed71a1
European Commission (2021b) The European Pillar of Social Rights Action Plan, Publications Office of the European Union. https://op.europa.eu/webpub/empl/european-pillar-of-social-rights/downloads/KE0921008ENN.pdf
EUWIN (2022) Workplace innovation: Europe's Competitive Edge. A manifesto for enhanced performance and working lives, European Journal of Workplace Innovation, 7 (1), 132–141. https://www.workplaceinnovation.org/kennis/workplace-innovation-europes-competitive-edge/
Federal Ministry of Justice (2022) Works constitution act. https://www.gesetze-im-internet.de/englisch_betrvg/

Ghita I.A. and Stan A. (2022) The dilemma of teaching in the digital era: Artificial intelligence. Risks and challenges for education, Jus et Civitas, 9 (2), 56–62.
Gilbreth F.B. and Gilbreth L.M. (1917) Applied motion study: A collection of papers on the efficient method to industrial preparedness, Sturgis & Walton.
Hacker W. (1986) Complete vs. incomplete working tasks: A concept and its verification, in Debus G. and Schroiff H.-W. (eds.) The psychology of work organization, North-Holland Publishers, 23–36.
Hacker W. (2003) Action regulation theory: a practical tool for the design of modern work, European Journal of Work and Organizational Psychology, 12 (2), 105–130. https://doi.org/10.1080/13594320344000075
Heinold E., Rosen P.H. and Wischniewski S. (2023) Advanced robotic automation: Comparative case study report, Publications Office of the European Union. https://doi.org/10.2802/15784
Hoffer Gittell J. (2016) Transforming relationships for high performance: The power of relational coordination, Stanford Business Books.
Holm J.R. and Lorenz E. (2022) The impact of artificial intelligence on skills at work in Denmark, New Technology, Work and Employment, 37 (1), 79–101. https://doi.org/10.1111/ntwe.12215
Huo M-L. and Boxall P. (2018) Are all aspects of lean production bad for workers? An analysis of how problem-solving demands affect employee well-being, Human Resource Management Journal, 28 (4), 569–584. https://doi.org/10.1111/1748-8583.12204
INRS (2023) Artificial intelligence in the service of occupational safety and health: Challenges and prospects for 2035, Institut National de Recherche et de Sécurité.
ISO (2021) ISO 45003: Occupational health and safety management — Psychological health and safety at work — Guidelines for managing psychosocial risks. https://www.iso.org/standard/64283.html
Ittermann P. and Virgillito A. (2019) Einfacharbeit und Digitalisierung im Spiegel der Statistik, in Hirsch-Kreinsen H., Ittermann P. and Falkenberg J. (eds.) Szenarien digitalisierter Einfacharbeit, Nomos, 69–86.
Karasek R. and Theorell T. (1990) Healthy work: Stress, productivity, and the reconstruction of working life, Basic Books.
Kellogg K.C., Valentine M.A. and Christin A. (2020) Algorithms at work: The new contested terrain of control, Academy of Management Annals, 14 (1), 366–410. https://doi.org/10.5465/annals.2018.0174
Kessinger R. (2021) Orchestrating friendship within a firm: Softening the edges of algorithmic evaluation, Master thesis, Massachusetts Institute of Technology. https://hdl.handle.net/1721.1/139385
Kuek S.C., Paradi-Guilford C.M., Fayomi T., Imaizumi S. and Ipeirotis P. (2015) The global opportunity in online outsourcing, The World Bank. http://documents.worldbank.org/curated/en/138371468000900555/The-global-opportunity-in-online-outsourcing
Lager H. (2019) Anpassungsfähigkeit in Zeiten der Digitalisierung: Zur Bedeutung von Empowerment und innovativer Arbeitsorganisation, Springer.
Lager H., Virgillito A. and Buchberger T.-P. (2021) Digitalization of logistics work: Ergonomic improvements versus work intensification, in Klumpp M. and Ruiner C. (eds.) Digital supply chains and the human factor, Springer, 33–53.
Lamannis M. (2023) Collective bargaining in the platform economy: a mapping exercise of existing initiatives, Report 2023.02, ETUI. https://www.etui.org/publications/collective-bargaining-platform-economy

Leka S. and Cox T. (eds.) (2008) The European framework for psychosocial risk management: PRIMA-EF, Institute of Work, Health and Organisations. http://www.prima-ef.org/uploads/1/1/0/2/11022736/prima-ef_ebook.pdf

Marx K. (1887) Capital, Volume 1: The process of production of capital. https://archive.org/details/CapitalVolume1

Mohr B.J. and Van Amelsvoort P. (eds.) (2016) Co-creating humane and innovative organizations: Evolutions in the practice of socio-technical system design, Global STS-D Network Press.

Moore P.V. and Starren A. (2019) OSH and the future of work: Benefits and risks of AI tools in workplaces, Discussion Paper, EU-OSHA. https://osha.europa.eu/en/publications/osh-and-future-work-benefits-and-risks-artificial-intelligence-tools-workplaces

Nicklich M. and Pfeiffer S. (2023) Digitalisation and self-perpetuation: Dynamics, drivers and temporalities of the transformation of working worlds, Work Organisation, Labour and Globalisation, 17 (1), 7–11. https://doi.org/10.13169/workorgalaboglob.17.1.0007

OECD (2021) Artificial intelligence and employment: New evidence from occupations most exposed to AI, Policy Brief December 2021. https://www.oecd.org/future-of-work/reports-and-data/AI-Employment-brief-2021.pdf

Oeij P.R.A., Rus D. and Pot F.D. (eds.) (2017) Workplace innovation: Theory, research and practice, Springer.

Onderwijsraad (2022) Inzet van intelligente technologie: Een verkenning. https://www.onderwijsraad.nl/publicaties/adviezen/2022/09/28/inzet-van-intelligente-technologie

Parker S.K. and Grote G. (2019) Automation, algorithms, and beyond: Why work design matters more than ever in a digital world, Applied Psychology, 71 (4), 1171–1204. https://doi.org/10.1111/apps.12241

Parker S.K. and Grote G. (2022) More than 'more than ever': Revisiting a work design and sociotechnical perspective on digital technologies, Applied Psychology, 71 (4), 1215–1223. https://doi.org/10.1111/apps.12425

Poba-Nzaou P., Galani M., Uwizeyemungu S. and Ceric A. (2021) The impacts of artificial intelligence (AI) on jobs: An industry perspective, Strategic HR Review, 20 (2), 60–65. https://doi.org/10.1108/SHR-01-2021-0003

Ponce Del Castillo A. and Naranjo D. (2022) Regulating algorithmic management: An assessment of the EC's draft directive on improving working conditions in platform work, Policy Brief 2022.08, ETUI. https://bit.ly/48DtKJz

Pot F.D. (2016) Gazelle, is dit wat we willen?, Zeggenschap, 27 (1), 8–11. https://www.kennisbanksocialeinnovatie.nl/kennis/gazelle-is-dit-wat-we-willen/

Pot F.D. (2017) Workplace innovation and wellbeing at work, in Oeij P.R.A., Rus D. and Pot F.D. (eds.) Workplace innovation: Theory, research and practice, Springer, 95–110.

Pot F.D. (2022) Monotonous and repetitive work – some people are more unequal than others, in Abrahamsson K. and Ennals R. (eds.) Sustainable work in Europe: Concepts, conditions, challenges, Peter Lang, 77–96.

Pot F.D., Peeters M.H.H., Vaas F. and Dhondt S. (1994) Assessment of stress risks and learning opportunities in the work organisation, European Work and Organisational Psychologist, 4 (1), 21–37. https://doi.org/10.1080/09602009408408604

Pot F.D., Alasoini T., Totterdill P. and Zettel C. (2023) Towards research-based policy and practice of workplace innovation in Europe, in Oeij P.R.A., Dhondt S. and McMurray A.J. (eds.) A research agenda for workplace innovation: The challenge of disruptive transitions, Edward Elgar Publishing, 255–271.

Rodrik D. and Sabel C.F. (2019) Building a good jobs economy, Working Paper November 2019, Harvard University. http://tinyurl.com/ybgpblpp

Suri R. (2010) It's about time: The competitive advantage of quick response manufacturing, Productivity Press.

Taylor F.W. (1911) Principles of scientific management, Harper & brothers.

Thiel C.E., Bonner J., Bush J.T., Welsh D.T. and Garud N. (2023) Stripped of agency: The paradoxical effect of employee monitoring on deviance, Journal of Management, 49 (2), 709–740. https://doi.org/10.1177/01492063211053224

UNI Global Union (2021) The Amazon panopticon: A guide for workers, organizers and policymakers. https://uniglobalunion.org/report/the-amazon-panopticon/

Vandaele K. (2021) Collective resistance and organisational creativity amongst Europe's platform workers: A new power in the labour movement?, in Haidar J. and Keune M. (eds.) Work and labour relations in global platform capitalism, Edward Elgar, 206–235.

Waas B. (2022) Artificial intelligence and labour law, Working paper 17, Hugo Sinzheimer Institute for Labour and Social Security law. https://www.hugo-sinzheimer-institut.de/faust-detail.htm?sync_id=HBS-008498

Wood A.J. (2021) Algorithmic management: Consequences for work organisation and working conditions, JRC Working Papers series on Labour, Education and Technology 2021/07, European Commission. https://joint-research-centre.ec.europa.eu/system/files/2021-05/jrc124874.pdf

Zoomer T., Van der Beek D., Van Gulijk C. and Kwantes J-H. (2022) Algoritmisch management en arbeidsveiligheid: het doel heiligt niet alle, Tijdschrift voor toegepaste arbowetenschap, 35 (2), 54–61. http://resolver.tudelft.nl/uuid:c4b0d751-3f8f-44df-a8af-3d979318821d

All links were checked on 20.02.2024.

Cite this chapter: Pot F. (2024) AI for good work, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 16
Social dialogue as a form of bottom-up governance for AI: the experience in France

Odile Chagny and Nicolas Blanc

1. Introduction: the need for a shift in perspective

'Social dialogue' refers to all the negotiations, consultations and exchanges that take place between employers and workers in a company or in a sector at local, national, European or international level. As Alain Supiot reminds us (Supiot 2001), this triple right to representation, action and collective bargaining has been the real driving force behind labour law. Yet digital transformation, and especially the development of artificial intelligence systems, raises a host of challenges in terms of the need to adapt social dialogue to take account of the specific features of transformations at work, to reflect on the place of social dialogue in relation to other modes of regulation and to develop exploratory approaches to new dialogue practices.

When we look at social dialogue issues related to the implementation of AI systems, the current period is crucial in several respects due to a combination of several factors.

First, the concept of AI in the workplace is now sufficiently established to consider the necessity to adapt social dialogue to the specific dimensions of AI systems compared to other digital technologies (i.e. taking into account the different steps in AI system value creation and the issues of acceptability, transparency, appropriation, etc.). Essential contributions to this debate have been provided by experts such as Dr Christina Colclough, founder of The Why Not Lab, international organisations such as the Trade Union Advisory Committee (TUAC) to the Organisation for Economic Co-operation and Development (OECD), the European Trade Union Institute (ETUI) and NGOs such as Future of Society, AlgorithmWatch, etc.[1]

In the French context, the major trade union organisations have all produced guidelines on the issue: the Ethics & Digital HR charter of Confédération Française de l'Encadrement CGC (CFE-CGC; the French management union) (CFE-CGC 2018); the 'robolution' guide of Confédération Générale du Travail des Ingés Cadres Techs (Ugict-CGT; CGT General Union of Engineers, Executives and Technicians) (UGICT-CGT 2020); the guide to AI at work of Confédération française démocratique du travail Cadres (CFDT Cadres; French Democratic Confederation of Labour managerial union) (Salis-Madinier 2022); the report on AI and human resources sponsored by Confédération Générale du Travail – Force Ouvrière (FO; General Confederation of

1. The various contributions to the ETUI AI Talks reflect this diversity, as did the ETUI seminar 'Rethinking Labour Law in the Digitalisation Era' https://elw-network.eu/wp-content/uploads/2020/10/European-Labour-Law-Conference-2020-Programme.pdf

Labour – Workers' Power) (Geuze 2022); and most recently, the guide to the AI Act released by Force Ouvrière Cadres (FO Cadres; the Confederal Union of Executives and Engineers – Workers' Power (FO Cadres 2023).

Second, we are gradually entering a phase where new AI systems, which we can describe as 'eco-systemic', are gaining in importance and spreading throughout the economy. These systems differ in many respects from the 'expert AI systems' of the previous generation. They modify the relational structure of the economy, displace and even 'reinvent' value chains and foster new innovation modalities.

This is taking place in the context of an unprecedented legislative framework being established at European level to impose requirements to tackle the risks associated with AI and to define rules for accessing, sharing and creating value with data. The impacts of these new systems are far from being fully clear and understood. It is not uncommon, for example, to see actors involved in the process of introducing eco-systemic AI systems in their business mentioning that they are 'gambling' in relation to their effects.[2] When value creation is complex to anticipate, the economic calculus is difficult to implement, especially on an ex ante basis. Most approaches implemented to assess the potential economic impact of AI systems are in many ways inappropriate as they focus on standard metrics, especially productivity. An illustrative example of this is provided by the AI Act assessment studies.[3] However, when economic calculation becomes uncertain, governance issues become strategic, while defining the rules for sharing value among shareholders becomes a particularly crucial step. Moreover, the latest generations of AI are based on highly assertive learning processes, making it virtually impossible to look inside the machine (a neural network can have, for example, 200 to 280 billion parameters) and raise the crucial questions regarding transparency and explainability.

All these developments encourage us to go beyond the concept stage. Clarifying the impacts and entering the operational phase of social and technological dialogue is essential in implementing more than a 'proof of concept' approach. The transition is what is obviously at stake, but the pace of change is far from rapid. This is obvious in the French context. At Pôle Emploi, the French public employment service, for example, it took six months in 2019 for a debate to take place within the framework of the central works council on an AI project ('Intelligence Emploi') which aimed to test an algorithm enabling advisers to respond more quickly to emails. Six out of the seven unions refused to take part in the vote. Confédération française démocratique du travail (CFDT; the French Democratic Confederation of Labour) publicly denounced these delays (CFDT 2019a) and demanded to monitor the tests within the framework of the social dialogue in order to measure their impact on working conditions, jobs and the service provided. The union also called for the adoption of an ethical charter and criticised the ability of the management to measure ex ante productivity gains (CFDT 2019a, 2019b). The

2. An example of serendipity in the effects of AI can be found in the Sopra Steria use case: https://youtu.be/Az2T251__MY?feature=shared

3. This is notably the case of the impact assessment accompanying the AI Act, produced with the support of consultancy firms.

ethical charter was adopted in April 2022 (Pôle emploi 2022), no fewer than three years after the announcement of the deployment of the AI project within the company.

2. AI systems in the workplace

In April 2022, La Poste announced the introduction of its first AI systems. Scheduling management tools based on AI will enable network operation managers to be offered 'turnkey' scheduling scenarios which they can then modify or validate. For the time being, however, the presentation of these tools has continued to elude the social dialogue and consultation bodies. In reaction, the trade unions, especially Force Ouvrière (FO; Workers' Power), asked for the adoption of an 'agreement on method' (FO Com 2022)[4] regarding the deployment of AI in the enterprise.

Workers' representatives are increasingly raising their voice to assert their information and consultation rights in the case of the introduction of AI systems. But they face major difficulties and are very often being left out of the AI decision-making process in terms of the technological aspects, the criteria used, the data collected and, even more so, in the nature and role of the algorithms. This observation is clear from the feedback received by several working groups recently set up in France to consider ways of strengthening 'technological' social dialogue in the context of the introduction of AI.[5]

However, this right to information and consultation has recently been potentially reinforced in the field of artificial intelligence. A recent decision of the Pontoise Court of Justice[6] concluded that companies should accept that workers' representatives (the social and economic committee) must be consulted and have recourse to an expert when new technology (in this case an AI system) is introduced, even if it has no identified impact on working conditions, a provision provided for by Articles L 2312-8 and L2315-94 of the labour code.[7]

Despite not yet being in wide usage, this right to information is not particularly well adapted to the context of artificial intelligence, largely because it does not take into account the specific temporality of AI systems and that these systems are not 'finished' when they are introduced, with the consequence that social dialogue about them has

---

4. This agreement was negotiated between an employer or employer representatives and one or more trade union organisations in order to define in advance the method of negotiation.

5. This is notably the case for the DIALIA project launched in 2023, coordinated by Institut de recherches économiques et sociales (IRES) in partnership with four trade unions (CFE-CGC, CFDT, Ugict-CGT and FO Cadres) and co-financed by Agence nationale pour l'amélioration des conditions de travail (ANACT; the French Agency for the Improvement of Working Conditions). This aims to contribute to the deployment of a shared methodological framework giving effect to the 2020 European Framework Agreement on Digitalisation (AI dimension) and which brings together a community of 80 participants, most of them members of trade union organisations.

6. TJ Pontoise, 15 April 2022, no. 22/00134.

7. This provision was introduced by the Ordinance of 2017 reforming the use of expert assistance by employee representative bodies. In the relevant article in the former labour code (Article L. 2323-29), recourse to an expert in the case of the introduction of new technologies was possible only if skills, remuneration, training or working conditions were affected: https://www.legifrance.gouv.fr/loda/id/LEGIARTI000035608975/2017-09-24/.

to be ongoing. This is the objective of the European SeCoIa Deal project,[8] launched in 2021 and co-financed by the European Commission. This project brings together around forty participants mainly from France and Italy, coordinated by CFE-CGC,[9] to explore the first avenues to design a 'new' social dialogue convening all stakeholders (providers, service providers, customers, companies, platforms) concerned with the transformation induced by the development of AI in order to reflect and promote bottom-up governance, in particular given the forthcoming European AI Act. CFE-CGC decided to initiate this project following the observation that the regulatory models emerging for AI are built mainly on a top-down basis and take little account of the real impacts on jobs and workplace organisations.

3. Top-down AI governance

For several years now, we have been experiencing in Europe the implementation of an AI governance system based on three forms of regulation.

The first is strong regulation by law, with the AI Act being the model proposed by Europe to be deployed in a uniform manner in all EU countries, like GDPR.

The second type of regulation is based on standards, with the aim of standardising the tools of the market – the International Organization for Standardization (ISO) at the international level and those of the European standardisation organisations at European level – so as to codify the AI Act in the form of complementary rules.

The last is based on soft law. Self-regulation can be proposed in the form of a charter, manifesto or ethics committee in companies as a complement to legal regulation.

The most important issue in this governance model is that the AI Act integrates these so-called soft law notions in the obligations imposed in respect of AI systems concerning the workplace. Although the text considers these systems as high-risk – which can be considered as a victory – compliance can be self-assessed, and the developers of these systems could, at least in terms of the changes proposed by the Council and the European Parliament, decide themselves if they believe the system to be high-risk (AlgorithmWatch 2023). The audits that can be conducted by the authorities are not particularly explicit: the developers of these AI systems may be asked for substantial evidence, but 'substantial' is not specified and so is open to interpretation (Bertuzzi 2023a). There are few or no safeguards for these systems which are having a major impact on the lives of employees, for instance when they are hired or when their performance is evaluated.

Another issue in the AI Act is the lack of obligations on the deployers of these AI systems. Here again there are no safeguards at the level of companies and only the instructions for proper functioning will guarantee the appropriate use of these tools.

8. https://secoiadeal.eu/

9. The project was piloted by CFE-CGC in collaboration with its partners (IRES, Astrées, CIDA and U2P).

More generally, there is also a patent risk concerning the guarantee of respect for our fundamental rights. An assessment of this general guarantee is mandatory, whatever the risk level of an AI system, but this is far from being delivered under a risk-based approach. In the initial Commission proposal, the AI Act defines in detail only the essential requirement. For example, Article 10 on data governance does not define the kind of biases that could be mitigated, leaving the responsibility of definition to the standardisation organisations. An application even identified as low risk, and therefore with few obligations, can affect the mental or physical integrity of employees or be discriminatory. In any case, the EC' proposal contained no obligation to carry out a fundamental rights and algorithm impact assessment ('FRAIA') (OECD 2023) of high-risk applications and to propose corrective measures. Consequently, 118 civil society organisations, including AlgorithmWatch and European Digital Rights (EDRi) put out a joint statement calling for an Artificial Intelligence Act which puts fundamental rights first (AlgorithmWatch 2021). Substantial improvements regarding fundamental rights were introduced by the European Parliament in its position adopted in May 2023, imposing on those deploying a high-risk system in the EU the obligation to carry out a fundamental rights impact assessment, including consultation with the competent authorities and relevant stakeholders (Article 29) (European Parliament 2023). However, the final outcome of the trilogue is far from certain, with the Spanish presidency having proposed to remove the fundamental rights impact assessment obligations and the mandatory consultation with relevant stakeholders (Bertuzzi 2023b).

The standardisation aspects of regulation are covered elsewhere in this volume (see Giorgi) and so, turning attention next to soft law, the third type of AI governance, it is clear that the number of charters in place is far from negligible (for example, 85 were identified by researchers between 2014 and 2019 (Jobin et al. 2019). However, even though they may be useful, they cannot replace the strong regulation proposed by the AI Act. Furthermore, the charters are never discussed with stakeholders such as employees, customers and beneficiaries; they are imposed on them and often remain at the level of broad generalities as to the guarantees provided in practice. Moreover, ethics committees are often opaque because they are not open to the same stakeholders, while they can also serve as a pseudo-scientific guarantee for AI systems that are not particularly transparent.

This lack of upstream discussion with stakeholders leads to the charters having sizable heterogeneity. This is particularly the case in France, where several companies have recently adopted ethical charters on AI, including Crédit Agricole (Crédit Agricole 2017), Thalès in January 2019 (Thalès 2021), Orange in April 2020, MAIF (MAIF 2021), Banque de France (Banque de France 2021) and Pôle Emploi (Pôle emploi 2022). However, it is scarcely possible to compare the charter of Pôle Emploi, which deals with fairness, non-discrimination, transparency, security and environmental impact issues, with that of MAIF, which emphasises the mastery of technology in the service of people, or that of Crédit Agricole, which focuses solely on its customers and their data and on the improvement of the services offered to them.

4. Social dialogue: bottom-up AI governance in companies

What role can social dialogue play in regulating AI? In the following paragraphs, we mainly draw on the conclusions of the European SeCoIa Deal project (SeCoIa Deal 2023), dedicated to sharing knowledge and experience and to the joint building of operational tools regarding this specific issue.

Social dialogue, as a form of social regulation, must of course be considered in conjunction with other levels of regulation.

As far as data is concerned, it is essential to rely on standards and labels, to have the opportunity to approve 'ethical' charters and to integrate the need to respect fundamental rights into the AI regulation (and therefore into the obligations that will fall on the developers of AI systems). The French experience with ethical AI charters (see above) suggests that there is a need to harmonise these in order to put forward the interests of employees but also those of customers. The doctrine proposed for public administration in August 2022 by Conseil d'État (French Council of State) (Conseil d'État 2022) provides concrete leads for all actors in this direction.

Levers can also be pulled where the potential offered by Article 88 GDPR can be seized, opening the way for collective bargaining on adjustments to data protection regimes in workplace relationships. Union representatives have the possibility, via GDPR, to check that there is no fully automated profiling (Article 22), i.e. that there is no processing of an employee's personal data to analyse and predict his or her behaviour, such as determining his or her performance at work.

Nevertheless, only via negotiation in companies will it be possible to initiate discussions on all the important issues related to the implementation of these systems: acceptability, transparency, explainability, appropriation, bias, robustness and organisational risks. Raising awareness of data processing and developing a 'data' culture is, in this regard, essential.

In the context of an AI Act at European level that will essentially proceed via self-regulation (by developers) and via the responsibilities that fall on deployers, employee representatives could be the 'first-level regulators'[10] capable of ensuring that the obligations set out in the future regulation of AI system providers and users (employers) will be met.

In the context of the SeCoIa Deal project, several operational tools have been co-elaborated in order to enable representatives to exert this role at company level:

- AI register system. Under the draft AI Act, where high-risk AI systems are used for business purposes, deployers have obligations in terms of using notices, keeping a log and carrying out data protection impact analysis. These obligations do not apply to systems that are not considered high-risk. Introducing a tool to

10. An expression developed by participants in the SeCoIa Deal project in order to insist on the importance of worker representatives with regard to the development of AI in a work context.

track AI systems installed in the company would be of use. Based on the principle of GDPR and its record of processing activities, a register would be set up to monitor the AI systems used in a company.

- Review clause in the framework of the cycle of use of AI-based tools at work. In order to ensure relevance and guarantee confidence in the tool and in the purpose of its use, it seems useful to imagine 'permanent', 'long-term' social dialogue on the AI-based tools used within a company. This dialogue would be based, among other things, on support for a review clause allowing the formalisation of a series of meetings between the actors, known to all in advance in principle and purpose, which will be held when the predetermined conditions are met. This clause may be included in a contract, in a collective agreement or in a declaration by the head of a company, or in a charter, resulting in a legally binding commitment on its part.
- Corporate AI ethics committee. The work of the SeCoIA Deal project has highlighted the need for 'first level control' in the workplace where AI systems are introduced and used. The creation of an ethics committee involving employee representatives, in conjunction with the creation of an AI ethics officer and record-keeping, is likely to strengthen deployers' obligations and the consideration of the evolving nature of AI systems.

Each of these three 'innovative' proposals elaborated in the framework of the SeCoIa Deal project have been integrated in the final roadmap of Conseil National de la Refondation Numérique, an initiative launched at the end of 2022 by the French government which brings together citizens, social partners and representatives from associations, businesses, research and government in order to identify solutions to contemporary issues, including digital transitions at work (Conseil national du numérique 2023).

5. Conclusion

In conclusion, bottom-up governance of AI, as exemplified in the French context, is the only way of deploying 'trustworthy AI' benefiting all stakeholders and, above all, employees. As the latest ILO study shows (Gmyrek et al. 2023), the impact on managerial jobs is going to be significant and it is therefore up to the unions to write the rules for tomorrow's technological social dialogue.

References

AlgorithmWatch (2021) Civil society calls on the EU to put fundamental rights first in the AI Act, 30 November 2021. https://algorithmwatch.org/en/eu-artificial-intelligence-act-for-fundamental-rights/

AlgorithmWatch (2023) EU legislators must close dangerous loophole in AI Act – Statement with 118 organizations, 7 September 2023. https://algorithmwatch.org/en/eu-must-close-loophole-ai-act/

Banque de France (2021) La Banque de France signe la charte internationale pour une IA inclusive, Communiqué de Presse, 11 June 2021. https://www.banque-france.fr/fr/communiques-de-presse/

Bertuzzi L. (2023a) MEPs advance on AI conformity assessment for high-risk uses, Euractiv, 08 February 2023. https://www.euractiv.com/section/artificial-intelligence/news/meps-advance-on-ai-conformity-assessment-for-high-risk-uses/

Bertuzzi L. (2023b) EU policymakers prepare to close first aspects of AI regulation, Euractiv, 11 July 2023. https://www.euractiv.com/section/artificial-intelligence/news/eu-policymakers-prepare-to-close-first-aspects-of-ai-regulation/?utm_source=substack&utm_medium=email

CFDT (2019a) Intelligence artificielle à Pôle emploi… la CFDT s'engage pour les salariés !, 07 August 2019. https://pste.cfdt.fr/portail/pste/nos-secteurs-professionnels/emploi/pole-emploi/-emploi-intelligence-artificielle-a-pole-emploi-la-cfdt-s-engage-pour-les-salaries-srv1_1005117

CFDT (2019b) Intelligence artificielle à Pôle emploi, 16 July 2019. http://cfecgc-metiersdelemploi.fr/2019/07/intelligence-artificielle-a-pole-emploi-cce-du-16-juillet-2019.html

CFE-CGC (2018) Charte éthique et numérique. https://bit.ly/48tQnjw

Conseil d'État (2022) S'engager dans l'intelligence artificielle pour un meilleur service public, 30 August 2022. https://www.conseil-etat.fr/actualites/s-engager-dans-l-intelligence-artificielle-pour-un-meilleur-service-public

Conseil national du numérique (2023) CNR – Les transitions numériques au travail. https://cnnumerique.fr/nos-travaux/cnr-les-transitions-numeriques-au-travail

Crédit Agricole (2017) Charte d'utilisation des données clients : le groupe Crédit Agricole prend une position forte avec la formalisation d'une charte. https://www.credit-agricole.com/responsable-et-engage/notre-strategie-rse-etre-acteur-d-une-societe-durable/blocs-masterpage-rse/charte-d-utilisation-des-donnees-clients-le-groupe-credit-agricole-prend-une-position-forte-avec-la-formalisation-d-une-charte

European Parliament (2023) Parliament's negotiating position on the artificial intelligence act, 07 June 2023. https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2023)747926

FO Cadres (2023) Le règlement européen sur l'intelligence artificielle 'IA Act' : un texte à s'approprier syndicalement, Analyse et Prospective, 8. https://foc.media.fo-cadres.fr/Analyse_et_prospective_n_8_IA_Act_2023_2716_5b4d256abe.pdf

FO Com (2022) Déploiement des premiers outils d'intelligence artificielle : FO demande un accord de méthode, Tracts cadres, 19 May 2022. http://www.focom-laposte.fr/outils-intelligence-artificielle-fo-demande-accord/

Geuze F. (2022) Intelligence artificielle, algorithmes et ressources humaines : un nouvel enjeu syndical, Institut de recherches économiques et sociales. http://www.ires.fr/index.php/etudes-recherches-ouvrages/etudes-des-organisations-syndicales/item/6517-intelligence-artificielle-algorithmes-et-ressources-humaines-un-nouvel-enjeu-syndical

Gmyrek P., Berg J. and Bescond D. (2023) Generative AI and jobs: A global analysis of potential effects on job quantity and quality, Working Paper 96, ILO. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---inst/documents/publication/wcms_890761.pdf

Jobin A., Ienca M. and Vayena E. (2019) The global landscape of AI ethics guidelines, Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

MAIF (2021) Charte numérique MAIF. https://entreprise.maif.fr/files/live/sites/entreprise-Maif/files/pdf/nos-chartes/charte-numerique-maif.pdf
OECD (2023) Fundamental Rights and Algorithms Impact Assessment (FRAIA), Catalogue of Tools and Metrics for Trustworthy AI, 5 April 2023. https://oecd.ai/en/catalogue/tools/fundamental-rights-and-algorithms-impact-assessment-%28fraia%29
Pôle emploi (2022) Charte de Pôle emploi pour une intelligence artificielle éthique. https://www.francetravail.org/accueil/communiques/pole-emploi-se-dote-dune-charte-pour-une-utilisation-ethique-de-lintelligence-artificielle.html?type=article
Salis-Madinier F. (2022) Le guide de l'intelligence artificielle au travail, Eyrolles.
Supiot A. (2001) Revisiter les droits d'action collective, Droit Social, 7/8, 687–704.
Thalès (2021) Charte éthique du numérique. https://www.thalesgroup.com/fr/global/responsabilite-dentreprise/gouvernance/thales-sengage-numerique-responsable
Ugict-CGT (2020) Intelligence artificielle et algorithme : pour quelle robolution. https://ugictcgt.fr/guide-ia/

All links were checked on 26.02.2024.

Cite this chapter: Chagny O. and Blanc N. (2024) Social dialogue as a form of bottom-up governance for AI: the experience in France, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 17
Collective bargaining and AI in Italy

Luciana Guaglianone

1. Introduction: the Italian production model and its impact on collective bargaining

The industrial structure in Italy is extremely fragmented with 99.5% of companies having fewer than 50 employees (according to Istat data). The adoption of AI-based production technologies is limited by the reduced investment capacities of companies, particularly of the smallest firms. Financial difficulties are not the only ones holding back the introduction of artificial intelligence systems in companies; poor managerial skills and organisational flexibility, which also characterise them, are further obstacles (Bannò et al. 2021).

The results of research carried out by the Artificial Intelligence Observatory of Politecnico di Milano confirm this. In 2021, 59% of large companies launched artificial intelligence projects compared to just 6% of small and medium-sized enterprises. Financial and organisational difficulties determine the types of systems installed: in 35% of cases, the introduction of technologies concerned chatbots and virtual assistants, while in 32% it was intelligent data processing (via algorithms to analyse and mine information from data).

The low presence of artificial intelligence systems and the nature of the solutions adopted condition both the number and the content of collective agreements, which are few and rather repetitive concerning the terms (Guaglianone 2021). The number of agreements rises if we consider jobs connected to the use of platforms where the most significant contractual issues concern, in Italy as elsewhere, the work of couriers (Cruz 2022).

1.1 The industrial relations system and model as a key to understanding bargaining trends in digital and automated work

The Italian industrial relations system is highly anomalous. Although collective bargaining is the prevailing method of action and the backbone of industrial relations, it is voluntary, private in nature and not regulated by law. However, both trade unions and collective bargaining are recognised by Italian law as fundamental, and the legislature has intervened and continues to intervene to promote their role. The historical interrelationship that exists between statutory law and collective bargaining applies to both branch and company levels of collective bargaining.

Meanwhile, the model of Italian industrial relations is, according to the classical classification, conflictual with weak participatory rights that mainly take the form of information rights. More participatory bargaining content, such as the right to consultation, is present in many collective agreements although most reproduce the legal obligations, merely indicating a time limit within which consultation procedures must be considered to have been completed (Guaglianone 2017).

2. Technological innovation and collective agreements – a brief overview

Digitalisation and automation have partially changed the attitude of all trade unions towards modes of participation which are now seen as fundamental to define and implement digitalisation[1] (Patto per la fabbrica 2018). This change in attitude has led to a broadening of the areas of bargaining, reflecting a real interest in sharing knowledge (e.g. the health of the company's finances; technological and organisational innovation plans; production quality).

For example, the duty to extend information rights also to the subject of technological innovation which, by way of interpretation, could already be inferred from the text of Article 5 of Legislative Decree no. 25 of 6 February 2007 (transposing Directive 2002/14/EC, which established a general framework for informing and consulting employees), in Contratto Collettivo Nazionale Lavoro (CCNL; branch collective agreement) per i lavoratori addetti all'industria metalmeccanica privata e alla installazione di impianti (for workers employed in the private engineering and plant installation industry), signed on 5 February 2021, also includes codetermination as well as the duty to provide, in any case, written reasons in the event that companies do not accept the trade unions' proposals.

On the whole, we cannot speak of a change in the model of the industrial relations system but what is innovative lies in the awareness, which has grown on the part of all trade unions, of the need to consider themselves not only antagonistic actors but also partners in the transformation of the enterprise. In concrete terms, this new conviction has meant that many branch collective agreements have introduced joint observatories to monitor and analyse digital transformation processes (e.g. CCNL Gas e Acqua, 30 July 2022; CCNL Chimica-Farmaceutica, 1 July 2022; CCNL per i lavoratori addetti all'industria metalmeccanica privata e alla installazione di impianti, 5 February 2021; CCNL Credito, 19 December 2019).

3. Case studies – couriers and a variety of collective agreements

The interest of both trade unions and lawmakers in Italy has, when it comes to platform work, mainly focused on couriers who make up about 12% of platform workers.

1. Accordo interconfederale Attuazione del patto per la fabbrica, 12/12/2018. https://olympus.uniurb.it/index.php?option=com_content&view=article&id=19549:patto2018&catid=233&Itemid=139

Evidence of this interest can be seen in a growing number of collective agreements – currently three – that regulate this form of work (CCNL Logistica e Trasporti (Logistics and Transport); CCNL UGL/Assodelivery supplementary sector agreement; Just Eat/Takeaway supplementary company agreement).

In order to reconstruct the bargaining context, we must first outline the legal framework within which, due to the interrelationship of the legal and contractual provisions, collective agreements are concluded. Article 2 of Legislative Decree no. 81/15 (as amended by Art. 47ff. of Law no. 128 of 2 November 2019), while defining work organised through platforms, including digital platforms, as 'hetero-organised' work,[2] extends the protections specific to paid employment to this form of work. The same law (Art. 2 para. 2(a)), however, authorises branch collective agreements – entered into by trade unions that are comparatively more representative at sector level – to derogate from it in the event that the social partners consider the sector to have particular production and organisational needs affecting pay and conditions. In 2019, this law was amended (Article 47ff. of Legislative Decree No. 81/15) with specific reference to couriers. The new law limits the social partners' power to derogate as regards the conditions of the employment relationship; however, it does entrust them with the task of defining the criteria for setting the overall remuneration, respecting the prohibition on piecerates but taking into account the manner in which the service is performed and the organisation of work.

The possibilities left open by the new legal provision have led to two collective agreements (see sub-sections below). It should be noted that neither of the collective agreements make the slightest reference to rights linked to the digitalisation of work. The core of the Logistica e Trasporti branch agreement and of the CGIL-CISL-UIL/Just Eat supplementary company agreement is the regulation of work based on the Protocol implementing Art. 47ff. Legislative Decree 81/2015; the UGL/Assodelivery sector agreement aims to define the employment relationship and to regulate some of its areas.

3.1 Protocol implementing Art. 47ff. Legislative Decree 81/2015

In November 2019, the Italian Parliament adopted a law which, while defining couriers as hetero-organised workers, delegates the setting of pay to collective bargaining (Art. 47ff. Law no. 128 of 2 November 2019). The following year, the trade unions Confederazione Generale Italiana del Lavoro (CGIL; Italian General Confederation of Labour), Confederazione Italiana Sindacati Lavoratori (CISL; Italian Confederation of Trade Unions and Unione Italiana del Lavoro (UIL; Italian Labour Union) and the Logistica e Trasporti employer associations signed a Protocol implementing Art. 47ff. Legislative Decree 81/2015. This Protocol commits the social partners to apply to couriers not only the pay set out in the Logistica e Trasporti branch collective agreement,

2. The term introduced by Article 2 to define quasi-subordinate workers, i.e. those who exclusively supply personal labour in favour of an employer who organises the work with reference to the time and place at which it occurs.

as supplemented by the Protocol of 18 July 2018 signed by the same parties, but all the contractual provisions contained in this agreement.

The inclusion of digital platform workers as hetero-organised workers has left untouched the issue of information rights and participation rights. The text of the new branch collective agreement (CCNL Logistica e Trasporti 2021) contains only the commitment to include the challenges of technological and digital innovation and structural changes in future negotiations.

3.2 The UGL/Assodelivery supplementary sector agreement

Assodelivery (an employer association including Deliveroo, FoodToGo, Glovo, SocialFood, Uber Eats and Just Eat as affiliates) was not one of the signatories of the Protocol implementing Art. 47ff. Legislative Decree 81/2015 (see Section 3.1). As a result, a large proportion of couriers (and therefore the large majority of those involved in platform-based food delivery) were not covered by it. The gap was filled by a collective agreement signed in September 2020 between Assodelivery and the Couriers' Union of Unione Generale del Lavoro (UGL), the first of its kind to regulate the employment relationship of platform-based food delivery couriers. This agreement, which made use of the regulatory provisions that allow collective bargaining to derogate from the law in the presence of the specific needs of the sector (Art. 2 of Legislative Decree 81/15, as amended by Art. 47ff. of Law no. 128 of 2 November 2019), defines couriers as self-employed workers while extending to them certain essential protections specific to employees. Again, however, nothing was said in the agreement about the right of trade unions to be informed of the algorithmic management of work, nor is the newly established joint committee involved in this issue.

The validity of this agreement has, however, been called into question by several rulings[3] that have declared it invalid due to UGL's lack of representativeness (Martelloni 2020).

3.3 The Just Eat/Takeaway agreement

In November 2020, Just Eat left Assodelivery and, in March 2021, signed a company agreement with CGIL, CSIL and UIL. Just Eat's goal was to experiment with its own work organisation model which defined the service as paid employment but, unlike in other countries (such as, for example, Spain), the terms of the collective agreement state that trade unions are not involved in discussions on algorithmic management and only individual rights are protected.

3. In its ruling of 30 June 2021 the Bologna Tribunal held that the collective agreement signed by Assodelivery and UGL Couriers Union to be unlawful since it was signed by a union which was not representative at branch level, as required by articles 2 and 47 of Legislative Decree no. 81/2015. The ruling was upheld in Bologna Court, decision no. 1332/21 of 12 January 2021.

The Logistica e Trasporti branch collective agreement, with appropriate adaptations, once again provides the legal basis for the employment relationship of couriers (Barbieri 2021; Forlivesi 2021).

4. Case studies: AI-based technologies in the Wind and TIM agreements in the telecommunications sector

Between 2020 and 2021, two telecommunications companies (Wind and TIM) deployed software known as Afiniti Advanced Routing which pairs customers with call centre agents using artificial intelligence. The implementation of this technology was preceded by the signing of collective agreements with the CGIL, CISL, UIL and UGL. The interesting part of the texts of the agreements (both have the same content) concerns the description of how the Afiniti system works. The data collected by the software, both through matching calling customers and telephone operators as well as those generated by agents' activities, are anonymised: each agent is assigned a code which is different from the usual code and of which only the system is aware. The system acquires the data independently through a predefined route that cannot be modified and, therefore, no reports can be generated that correlate the work done to the performance of individual workers since the software is designed to process data for commercial purposes only.

The key issue for the unions, however, was the fear that the system could indeed covertly monitor agents' work; the agreement therefore intervenes on this aspect, citing the provisions contained in Article 4 of Law no. 300/70 (Statuto dei lavoratori; Workers' Statute) as a limit to the legitimacy of the operation of the software and as the legal basis of the agreement itself. Consequently, Afiniti Advanced Routing software may only be used for organisational and production needs, and for work safety and the protection of company assets; it may not be used to monitor the workplace secretly and neither may its use entail any negative consequences for the management of labour relations. This prohibition, demonstrating the strong interest that trade unions have in the protection of individual rights, is reinforced by repeated references in the collective agreements to the provisions of the Privacy Code (Legislative Decree no. 196/2003, transposing EU Regulation on data protection 2016/679) which prohibit covert monitoring by technological means.

As far as collective rights are concerned, the desire to negotiate the management of the software with trade unions can only be clearly understood if all the agreement clauses dealing with this subject are read in conjunction with each other. At first reading, trade union participation seems to be limited only to annual bargaining rounds aimed at monitoring the effects of the introduction of the system. In reality, as is clear from the subsequent contractual provisions, trade unions also have the right to propose improvements, which the companies undertake to assess, as well as the duty, in the event that the unions identify critical issues, to discuss jointly how to overcome them. Finally, unions have the right to terminate the agreement in the event that the critical issues raised prove to be unresolvable, not only at company level, but also following the joint assessments that are to take place subsequently (at territorial and branch levels).

5. Conclusions

The introduction of digital systems, especially software with content generation, prediction, recommendation and decision-making capabilities that influence the contexts with which those systems interact, would suggest there is a need to rethink decision-making models, extending them to forms of governance that include civil society (Guaglianone 2020) and the social partners. However, the situation at present, even at European level, sees a tension between the regulatory models proposed by lawmakers and the expectations expressed by the social partners. The contrast between the social partners' aspirations for participation, contained in the Framework Agreement on Digitalisation (Rota 2020), and the text of the AI Act (COM (2021) 206 final) should be read in these terms. Whereas the latter classifies as high-risk those artificial intelligence systems that generate outputs such as content, predictions, recommendations and decisions related to labour relations, it does not envisage any role for trade unions (Ponce Del Castillo 2021).

As far as the situation in Italy is concerned, at least with regard to the social partners' right to codetermine the changes brought about by new technologies, this split does not concern the tension between the regulatory models proposed by lawmakers (e.g. legal regulations) and trade union will, but is instead created by the same model chosen and proposed by collective organisations (Patto per la fabbrica). We are therefore faced with a gap between participatory techniques that are imagined and actual bargaining practice. Only one of the collective bargaining agreements examined (see Section 2) includes a duty to consult on AI systems, even if the subject is rather generically indicated with the expression 'technological innovation plans' (in the metalworking and engineering branch collective agreement); while in the other collective bargaining agreements participation takes the form of the establishment of joint observatories. This form of participation is extremely bland, especially given the nature of the issues which would require, at the very least, information procedures, if not consultation (Ponce del Castillo 2021).

One exception is bargaining related, and not confined to a specific sector,[4] which is that concerning the deployment of AI-based software (see Section 4), part of a particularly participatory element of industrial relations characterised either by the signing of protocols (CCNL in the electricity sector, 15 January 2021) or by greater awareness of digitalisation (in the telecommunications sector). In this case, the bargaining subject is the monitoring of the use of an algorithm which, by combining artificial intelligence and big data in real time, predicts the behaviour of people contacting a call centre in order to match them with like-minded call centre agents (Carchidi 2022).

Bargaining related to the work of couriers may be assessed differently. Driven by the need to find a non-judicial form of protection (a path followed by many couriers themselves)[5] (Bellavista 2022; Razzolini 2020), and under pressure as a result of the

4. In January 2022 ENEL (energy sector) signed a collective agreement with CGIL, CISL, UIL and UGL regulating the use of the Afiniti Advanced Routing system. The text is identical to the TIM and Wind agreements.

5. See, among others, Palermo Court decision no. 3570/20 of 24 November 2020 and Turin Court decision of 18 November 2021.

intense media attention, the government issued a law that, according to the traditional model of intervention, weaves together legal provisions and references to collective bargaining.

In all these cases, however, collective bargaining has not gone beyond the bargaining of standard protections; that is, simple protections that do not take into account the peculiarity of the work model entailed by labour platforms or the means by which that work is carried out. In other words, bargaining concerns the form and the conditions of work but does not seek to guide or control the digital mechanisms that drive it. In short, collective bargaining remains an active instrument in regulating work, but the main scope of bargaining content is related to protecting 'traditional' and basic individual rights. What is being negotiated is the effects that the organisation of work performed using digital platforms have had, but no demand is being made for control over these, i.e. for joint management of the decisions being made.

6. Prospects

An interesting intertwining of tradition and innovation is what could/should be produced by the legislation contained in Legislative Decree no. 104 of 27 June 2022 (implementing Directive 2019/1152 on transparent and predictable working conditions in the European Union). Article 1a requires the employer to inform:

> the employee of the use of automated decision-making or monitoring systems intended to provide indications relevant to the recruitment or assignment, management or termination of the employment relationship, task allocation as well as indications affecting the monitoring, evaluation, performance and fulfilment of the contractual obligations of employees.

Paragraph 2 of the same Article models the duty on the particular type of work and, in an analytical manner, indicates all the points that the information must both touch on and seek to make workers aware of, and then comprehend, in terms of the purposes, aims and limits of automated systems. The functioning of the system, the parameters used to train it, the repercussions (if any) on the systems themselves, the control measures that are taken and the correction processes carried out, or that can be carried out by human personnel, are all subjects that must be brought to the knowledge of the worker.

The interest of these provisions to industrial relations scholars is twofold. First, the text of the directive does not make any specific reference to information rights linked to work that uses automated monitoring or decision-making systems. The Italian legislator, therefore, has chosen to broaden the scope and subject matter of the requirement to inform. Second, ownership of the right to information is also assigned (under paragraph 6 of Art. 1a) to trade unions. More specifically, in accordance with branch practice (see recital no. 49), it is the company-level representatives or, if these are not present, the territorial ones (of those trade unions that are comparatively more representative at branch level) that own the right. Furthermore, this right would seem to be reinforced by the possibility of requesting further information, if the information given is deemed

insufficient, as well as by the employer's duty to comply with this request within a period of 30 days (Iodice 2023).

The additions made by Legislative Decree no. 104/22 to the text of the directive – even though no application of it has yet come to light – certainly constitute an extension of information rights. They fit, however, into the model of weak participation typical of Italian industrial relations since they lack any reference not only to the duty to negotiate (which is never present in the Italian legal system) but also to consult, which could have been indicated as a precondition to decisions being made about the organisational measures to be introduced. What has already been stated regarding the split between the abstract tension towards a more intense participation model and the weight of tradition therefore continues to be a hindrance also in this specific case (Donini and Ingrao 2022).

References

Bannò M., Filippi E. and Trento S. (2021) Risks of automation of occupations: An estimate for Italy, Stato e mercato, 3/2021, 325–350. https://doi.org/10.1425/103268

Bellavista A. (2022) Riders e subordinazione: a proposito di una recente sentenza, Lavoro Diritti Europa, 2/2022, 2–12 https://www.lavorodirittieuropa.it/images/BellavistaRider2022.pdf

Carchidi D. (2022) Afiniti, un caso riuscito di contrattazione dell'algoritmo. https://www.slc-cgil.it/notizie-tlc-ed-emittenza/3791-afiniti-un-caso-riuscito-di-contrattazione-dell-algoritmo.html

Donini A. and Ingrao A. (2022) Algoritmi e lavoro, Labour Law Community. https://www.labourlawcommunity.org/ricerca/algoritmi-e-lavoro/

Guaglianone L. (2017) Los derechos de información y consulta en Italia: entre normas de ley y disposiciones contractuales, in Villalon J.C., Menéndez Calvo M.R. and Nogueira Guastavino M. (eds.) Representación y representatividad colectiva en las relaciones laborales, Bomarzo, 561–575.

Guaglianone L. (2020) Industria 4.0 y modelo partecipativo: ¿dialogo social vs dialogo civil? Las repercusiones sobre el sistema de relaciones industriales italiano, Temas laborales, 152, 97–113.

Guaglianone L. (2021) Brecha de género, nuevas tecnologías y trabajo digital: enfoque sobre Italia, in Rodríguez Fernández M.L. (ed.) Tecnología y trabajo: el impacto de la revolución digital en los derechos laborales y la protección social, Aranzadi, 81–104.

Iodice D. (2023) Il d.lgs. n. 104/2022 nella prospettiva del diritto sindacale: quale futuro per le relazioni industriali?, Working Paper 1/2023, Adapt University Press, 4–20 https://www.bollettinoadapt.it/il-d-lgs-n-104-2022-nella-prospettiva-del-diritto-sindacale-quale-futuro-per-le-relazioni-industriali/

Martelloni F. (2020) CCNL Assodelivery - UGL: una buca sulla strada dei diritti dei rider, Questione giustizia. https://www.questionegiustizia.it/articolo/ccnl-assodelivery-ugl-una-buca-sulla-strada-dei-diritti-dei-rider

Ponce del Castillo A. (2021) The AI regulation: Entering an AI regulatory winter? Why an ad hoc directive on AI in employment is required, Policy Brief 2021.07, ETUI. https://www.etui.org/publications/ai-regulation-entering-ai-regulatory-winter

Razzolini O. (2020) I confini tra subordinazione, collaborazioni eterorganizzate e lavoro autonomo coordinato: una rilettura, Diritto delle relazioni industriali, 360 (2), 346–376.
Rodríguez Fernández M.L. (2023) La participación de las personas trabajadoras en la gobernanza de la transición digital: las experiencias de la Unión Europea y de España, Revista de Derecho Social, (101), 107–140.
Rota A. (2020) Sull'accordo quadro europeo in tema di digitalizzazione del lavoro, Labour and Law Issues, 6 (2), 23–48. https://doi.org/10.6092/issn.2421-2695/12042
Villalón J.C. (2022) Digital work: A new task for social dialogue, in do Rosario Palma Ramalho M., Carvalho C. and Vicente J.N. (eds.) Trabalho na era digital: que direito? - Work in a digital era: Legal challenges, Estudios APODIT 9, AAFDL Editora, 517–544.

All links were checked on 26.02.2024.

Cite this chapter: Guaglianone L. (2024) Collective bargaining and AI in Italy, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 18
Collective bargaining and AI in Spain

María Luz Rodríguez Fernández

1. Introduction: the regulatory context for collective bargaining related to AI

1.1 The 'Riders' Law'

Law 12/2021 of 28 September, whereby Estatuto de los Trabajadores (ET; Workers' Rights Statute) was amended to guarantee the labour rights of people engaged in distribution and delivery through digital platforms, better known as the 'Riders' Law', is the epitome of Spanish legislation with respect to how the digital transition and its associated forms of employment are handled.

From the very outset, the Law's Explanatory Memorandum clearly highlights the interaction between the regulation of work on app-based delivery platforms and the advance of technology (coverage of the Law in this particular area relates specifically to 'persons who provide remunerated services consisting in the delivery or distribution of any consumer product or good'). The most significant points of this interaction are emphasised by the Law's insistence on the compatibility between technological progress and the protection of workers' rights and, at the same time, by its coverage also of the consequences of algorithmic management, in which area the Law extends to all kinds of companies and not just delivery platforms. Law 12/2021 therefore considers that making technological advance compatible with worker protection is: 'the formula (…) for ensuring that the positive effects from the technological revolution are distributed equitably and that this revolution is of benefit to the advancement of society'. In conjunction with this, the social partners and the legislature must pay special attention to algorithmic management, not only due 'to the changes that are being introduced to the management of business services and activities (and) to all aspects of working conditions', but also because 'such alterations are taking place aside of the traditional scheme of participation by the employees of a company'. Ultimately, the relationship between the digital transition and its impact on the world of work is what supports and gives overall meaning to the Law:

> … one of the other reflections shared at the social dialogue round table [is that] we cannot ignore the impact of new technologies in the labour environment or the need for labour legislation to take into account these repercussions, not only on the collective and individual rights of workers but also on competition between companies.

This shared understanding of the advance of technology and its consequences for work could be the reason why agreements have been reached as a result of social dialogue in Spain on the regulation of subjects as controversial as digital platform work and algorithmic management, unlike in other parts of the European Union (EU) and other countries which have taken initiatives on these issues.[1] And it is not insignificant that this all took place during the social dialogue on tackling the Covid-19 pandemic considering that, as the Explanatory Memorandum of Law 12/2021 states, 'despite the enormous difficulties represented by tackling this challenge, especially the technical difficulties, social dialogue has allowed our country to be a pioneer in advancing on this subject, which it does together with a diagnosis and a shared solution by the social partners.'

Consequently, the regulation of work on app-based delivery platforms or, rather, the presumption that work on app-based delivery platforms is employed work and not engaged in on a freelance basis, has been the product of social dialogue and of governance by the social partners in the digital transition. There is no official data but, according to the available estimates, Spain is the EU country with the highest percentage of platform workers (Urzí Brancati et al. 2020: 14-15) and it seems that the 'platformisation' of the Spanish economy shows no signs of stopping. That explains the particular relevance of why the social partners in Spain have taken charge of governing this phenomenon. But what is most notable, for the purposes of this chapter, is that Law 12/2021 also regulates algorithmic transparency and that, regarding this issue, there has also been agreement through the social dialogue.

In accordance with the revised Article 64(4)(d) ET, worker representatives have the right to be 'informed by the company about the parameters, rules and instructions that form the basis of the algorithms or artificial intelligence system that affect decision-making that could have an impact on working conditions and on access to and keeping a job, including profiling.' This algorithmic transparency rule opens up a whole realm of possibilities for the participation of workers in areas that were previously part of corporate prerogative and that, therefore, were essentially beyond their reach.

In truth, this rule affects the 'soft' part of the possible participation of worker representatives given that Article 64(4)(d) only makes mention of the right 'to be informed', meaning that a company should inform worker representatives of the main aspects of algorithmic decisions. And it is also true that this right is imprecisely set up regarding the form, time and actual content of the information: what those 'parameters, rules and instructions that form the basis of the algorithms or artificial intelligence system' actually consist of is something that, at the very least, is foreign to the traditional jargon of labour relations. Finally, it is likewise true that this right to information about algorithmic decisions is accompanied neither by the right of employees (other than worker representatives) to be informed about the algorithmic decisions that affect them (although this is already guaranteed by the application of Articles 13(2)(f), 15(1)(h) and 22 of the General Data Protection Regulation) nor by an assessment of the decisions adopted as a result of the application of algorithms in order to check if they are biased or causing discrimination (Ginés i Fabrellas 2021).

1. These initiatives can be consulted in the summary provided by the International Labour Office (ILO 2022: 31-33).

Despite previous, well-founded objections in the specialist literature, the reality is that the recognition of this right opens up an unprecedented check on corporate prerogative. Reviewing the ET and other labour laws shows that the lawfulness of any business decisions that are made can be subject to control, but always after the fact. Furthermore, up to now, the reasons upon which such decisions were based were a kind of 'black box' rooted in corporate prerogative. Now, if the business decisions in question affect working conditions, access to employment or keeping a job, including profiling, and those decisions have been adopted using algorithms or AI systems, then, as required by Article 64(4)(d) ET, worker representatives have the right to know the ultimate reasons for them, not just whether they comply with the law. This right to algorithmic transparency opens up the black box to a certain extent, allowing worker representatives to see inside and, most importantly, to utilise what they see to guarantee that the business decisions in question do not cause bias or differentiated treatment without justification. This cannot be done regarding business decisions that are made without using algorithms (there is no way to look inside the heads of people who make the corresponding decisions), which will therefore create differences regarding the control of business decisions by worker representatives depending on whether or not such decisions are automated (Rodríguez Cardo 2022: 167).

This is what is meant by knowing 'the parameters, rules and instructions' that form the basis of algorithmic decisions. It does not mean access to the source code of an algorithm, which could even be protected by industrial confidentiality, while knowledge of that code would require worker representatives to have programming-related skills that they often do not possess. However, the algorithmic transparency referenced in Article 64(4)(d) ET does, first of all, allow having the knowledge that a company, through the use of algorithms, is making decisions that affect 'working conditions and on access to and keeping a job, including profiling.'

It should be clarified that the use of an algorithm does not relieve a company of its accountability related to such decisions. An algorithm is nothing more than a tool used by a company to make those decisions, but it does not substitute a company's will or power. Some expressions that are very much in vogue, such as 'your boss is an algorithm' (Aloisi and De Stefano 2022), could lead to confusion in this regard. The 'boss' continues to be the company that is making the decisions that affect its employees, while the algorithm is nothing more than a software tool it uses as part of the process.

Algorithmic transparency encompasses the parameters, rules and instructions or, in other words, the knowledge of an algorithm's operating logic, characteristics and consequences. More specifically, worker representatives have the right to know the following: (a) the variables used by an algorithm and whether or not they include personal data; (b) the weight of each variable in the decisions that are made; and (c) the rules and instructions (programming rules) used by the algorithm (Galdón Clavell et al. 2022: 13).

Worker representatives should have this information at a timely moment. All the aforementioned is useless if worker representatives are kept on the sidelines and only learn it after algorithmic decisions have already being made. If this happens, then the

purpose of algorithmic transparency cannot be met. It is not about knowing for the sake of knowing; rather, it is about avoiding the certain risks caused by algorithmic decisions. Algorithms have substantial potential to perpetuate discrimination in terms of the selection of the data with which they are fed and/or through the programming rules that are followed (Bernal Santamaría 2020). Therefore, only by knowing, in advance, how those data are selected and processed can such discrimination be prevented (Gómez Gordillo 2021: 179). Moreover, this would be in accordance with the mandate of Article 64(6) ET which requires that information be provided 'at a time, in a manner and with the content [...] such that worker representatives can examine it adequately', and in accordance with the content of Article 64(5)(f) ET which gives worker representatives the right to issue a report prior to the execution of decisions by a company regarding 'the implementation and review of work organisation and control systems' if they are operated using algorithms.

1.2 The Charter of Digital Rights

On 14 July 2021, the Charter of Digital Rights was presented in Spain, laying out citizens' fundamental digital rights. Regarding the use of algorithms and AI systems in the workplace, the Charter sets down several provisions. Section XIX.6, which establishes the guarantees related to the use of algorithms, states that:

> ... the development and use of algorithms and any other equivalent procedures in the work environment will require an impact assessment related to data protection. The analysis thereof will include the risks related to the ethical principles and rights pertaining to artificial intelligence contained in this Charter, and it will particularly include the gender perspective and the prohibition of any discrimination, both direct and indirect (...).

The ethical principles and rights pertaining to artificial intelligence are defined in Section XXV.1 which states that 'artificial intelligence must ensure a human-centric vision and the inalienable dignity of people; it will pursue the common good and it will comply with the do no harm principle.' In addition to this are the provisions of Section XXV.2.b, which sets out that 'during the development and lifecycle of artificial intelligence systems (...), the conditions of transparency, auditability, explainability, traceability, human supervision and governance will be established.'

As can be seen, the provisions of the Charter of Digital Rights go beyond the provisions of Law 12/2021 with respect to algorithmic transparency. The Charter of Digital Rights requires there to be an 'impact assessment' of the use of algorithms which must take into account the gender perspective which is not included in Article 64(4)(d) ET. At the same time, the impact assessment must likewise take into account the ethical principles and rights related to artificial intelligence that are established in the Charter including, among others, some that are as crucial as the human-centric vision and the 'do no harm' principle, as well as the auditability of algorithms. This is also not present in Article 64(4)(d) ET; if it were, it would serve to define more accurately the duty of algorithmic transparency contained in that Article.

The problem, however, is that the Charter's provisions lack legal effect. The Charter itself acknowledges this point:

> Prior Considerations: the nature of the Charter is not regulatory, rather its objective is not only to acknowledge the very recent application and interpretation challenges represented by the adaptation of rights to the digital environment, but also to suggest principles and policies that refer to this new context.

The only thing that the Charter seeks to do is describe a scenario that facilitates the adoption of public policies whose purpose is to protect fundamental rights from the onslaughts they face from the progress of digitalisation (De la Sierra 2022: 49-50). Without such public policies, the Charter has no legal effectiveness whatsoever. As such, in accordance with the provisions of Section XXVIII of the Charter, the government has to adopt 'the appropriate measures, within the scope of its jurisdiction, to guarantee the effectiveness of this Charter'. Even so, the symbolic value of the rights included in the Charter should be noted, as well as the possibility that they could serve as guiding principles for a better interpretation of existing rights and institutions, particularly the right to algorithmic transparency.

2. Collective bargaining experiences related to AI in Spain

While the subject of AI is not covered in the first collective agreements in the world related to platform workers (Rodríguez Fernández 2022), two pioneering experiences can be identified in collective bargaining in Spain related to the consequences of AI at work. The number of collective agreements that include this subject can only increase, however. Currently, just 9.6% of Spanish companies use AI (INE 2023), but in Spain's 2025 Digital Plan, the government has set the goal of 25% of companies using AI by that year. As the number of companies that use AI increase, the number of collective agreements that deal with its regulation will also certainly rise.

2.1 Sectoral negotiation on AI: collective agreement in the banking sector

On 29 January 2021, the XXIV Convenio colectivo del sector de la banca (24th collective agreement of the banking sector) was signed between Asociación Española de la Banca (the Spanish Banking Association) and the unions Comisiones Obreras (CCOO; Workers' Commissions), Unión General de Trabajadores (UGT; General Union of Workers) and Federación de Banca de FINE (FINE Banking Federation).[2] This collective agreement runs from 1 January 2019 to 31 December 2023 (the agreement entered into force retroactively, a fairly frequent practice in Spain).

2. Accessible as registered in Agencia Estatal Boletín Oficial del Estado (BOE) at: https://www.boe.es/eli/es/res/2021/03/17/(1).

In general, this collective agreement acknowledges that collective bargaining must play a leading role in the digital transformation processes of companies. Article 79 of the agreement states as follows:

> Given that the digital transformation is an element of company restructuring, with potential effects not only on employment but also on job characteristics and working conditions, the parties acknowledge that collective bargaining, due to its very nature and functions, is the instrument for facilitating adequate and fair governance of the impact of the digital transformation of (companies) on employment in the sector.

Note that the use of collective bargaining as a tool for tackling the consequences of the digital transformation of the banking sector is considered to be the formula for 'governing' such consequences 'fairly', meaning by taking into account the necessary balance between the interests of the company and those of workers, which do not always coincide.

Under this guiding principle, the collective agreement first regulates the framework within which teleworking in the sector takes place (Article 27). Subsequently, Article 80 regulates the digital rights of employees, setting down: (a) the right to digital disconnection; (b) the right to privacy with regard to the use of digital devices owned by the company; (c) the right to privacy regarding the use of video surveillance, sound recording and geolocation devices; (d) the right to digital education, which encompasses actions to eradicate digital gaps; and (e) a 'right regarding artificial intelligence'. This latter right has two facets. From the perspective of employees, they have the right not to be the object of decisions based 'solely and exclusively on automated variables' and not to be discriminated against by decisions that might be based exclusively on algorithms. In both cases, an employee could request the intervention of a human. From the perspective of worker representatives, there is a right to receive information about the use of algorithms or AI which, as required by Article 64(4)(d) ET, includes not only the data that are fed to algorithms and their operating logic, but also an assessment of the outcomes in order to see if algorithmic decisions are resulting in discrimination.

The preceding is in line with what was provided for in the 2020 European Social Partners Framework Agreement on Digitalisation which, regarding AI, set down that in 'situations where AI systems are used in human-resource procedures (…) transparency needs to be safeguarded through the provision of information. In addition, an affected worker can make a request for human intervention and/or contest the decision along with testing of the AI outcomes.' But the collective agreement of the banking sector goes even further and beyond the provisions of Article 64(4)(d) ET given that, together with information about data and about the operating logic of an algorithm, it requires an impact assessment of the decisions adopted through the use of AI as a means of preventing possible bias.

CCOO, UGT and FINE acknowledge that, after initiating and then defining a legal framework for exercising the right of algorithmic transparency, there has been barely any progress as regards implementation. The reason is that the substantial rises in

the cost of living that occurred shortly after the agreement was signed have diverted the priority of unions towards calling for wage increases in accordance with the rise in inflation, thereby placing algorithmic transparency and digital rights on the back burner.

Drawing on its experiences with the banking sector agreement, CCOO has been able to make general progress on this subject through two routes.[3] The first refers to the drafting of a procedure for requesting information from companies about algorithms and AI, and then a subsequent one for reporting to Inspección de Trabajo y Seguridad Social (ITSS; the Labour and Social Security Inspectorate) any breach of the duty to provide that information. There is some confusion among companies about what algorithmic transparency means, in the sense that the majority of enterprises regard that they do not have to volunteer information about the algorithms they use; rather that the information must be requested by worker representatives. No such understanding is inferred from the wording of Article 64(4)(d) ET, however – and rather the opposite: companies must provide that information without having to be prompted.

The CCOO procedure[4] includes the information that worker representatives in a company should request regarding algorithmic decisions. In particular, this information includes the following: (a) 'the preliminary design specifications, including the criteria and parameters and/or variables that have been determined for managing data'; (b) 'the final technical specifications (the pseudocode) that explain what the artificial intelligence system-algorithm does'; and (c) 'the impact assessment regarding data protection and if there is a risk related to the rights and freedoms of people, as well as the periodic assessments that are conducted regarding outcomes'. In cases when a company refuses to provide the described information, despite a request having been made, the second protocol concerns the reporting of that situation to the ITSS, causing the latter to contact the company and insist on compliance with its obligations regarding algorithmic transparency.

The second route through which CCOO is seeking to advance is via the training of company employees who are engaged in programming or the acquisition of algorithms. With this in view, it is seeking to include a standard clause in all collective agreements warning of the risks of using algorithms and how these can be dealt with:

> Any person who programs or acquires algorithms must receive training by a company to gain proper knowledge of the risks of partiality and discrimination, as well as training on adopting possible measures for reducing those risks. Algorithms must be periodically audited by independent third parties, chosen together between the company and unions, to verify that those algorithms are not

3. I would like to thank Raúl Olmos and Raquel Boto, of the CCOO Secretary's Office for Union Action and Employment, for the information they have provided in the writing of this section.

4. The history behind this is that, in October 2022, CCOO of Catalonia requested that the Glovo delivery platform provide information about the algorithms and/or AI system it used, and subsequently provided worker representatives with a procedure for requesting information on this subject from all companies (https://www.ccoo.cat/noticies/ccoo-de-catalunya-exigeix-coneixer-com-funcionen-els-algoritmes-de-glovo/). As of January 2024, however, Glovo had not delivered the information requested.

subject to partiality or discriminatory outcomes. The results of these audits will be made available to all persons who are affected by algorithmic decisions, including union representation.

For now, this clause has not been included in any collective agreement, but it will be called for in the next rounds of negotiations.

2.2 Company-level bargaining related to AI: agreement between Just Eat and CCOO and UGT

On 17 December 2021, a collective agreement was signed between Takeway Express (Just Eat) and CCOO and UGT. The term of the agreement began on the date it was signed and extended to 31 December 2023. This delivery platform, unlike others, has followed a strategy of acknowledging the existence of an employment relationship between couriers and the platform, and has negotiated collective agreements with traditional unions in various European countries (Denmark, Germany, Italy and Spain) (Hadwiger 2022).

As with the collective agreement in the banking sector, the Just Eat agreement contains a chapter dedicated to the digital rights of employees. Article 67 regulates data protection (specifically, the principles of data minimisation, limitation of purpose, and transparency and accuracy in processing), while Article 68 regulates all other digital rights: (a) the right to disconnection; (b) the right to privacy in the use of digital devices owned by the company; (c) the right to privacy in the use of video surveillance and sound recording devices in the workplace; (d) the right to privacy in the use of geolocation systems, which includes employees' right to know the characteristics of these systems and information about how they can exercise rights of access, rectification, restriction of processing and erasure; (e) the right to information about algorithms and AI; and (f) the right to information about digital work tools, especially the use of chatbots or humans for responding to communications between employees and the platform.

Regarding the right to information about algorithms and AI, the agreement includes provisions that go far beyond what is required by Article 64(4)(d) ET. First of all, the agreement specifies that the platform must provide worker representatives with 'the relevant information used by the algorithm and/or artificial intelligence systems' for organising delivery activity, such as the type of contract that employees have, the number of hours during which they have provided their services or the days off that they have taken. Second, the data that cannot be used by the platform in its algorithm is also specified: 'the company will guarantee that (…) data that could give rise to a violation of fundamental rights must not be considered, such as (…) the sex or nationality of employees.' The agreement thereby determines not only the data to which worker representatives must have access – those used by the algorithm for organising its delivery activity – but also those that cannot be used by the algorithm because they could give rise to discrimination against employees.

In order to be able to learn about and manage information related to algorithms and AI, the agreement provides for the creation of a joint committee, called the 'algorithm committee'. This is composed of two people representing Just Eat and two others representing each of CCOO and UGT. In addition to receiving the information required by the agreement, the representatives of both unions may request that the person responsible for supervising the algorithm and/or AI system appears before the committee.

Finally, the agreement sets out two restrictions on the information related to algorithms and AI to which worker representatives have a right. The first is that the platform will not provide information 'that is protected by regulations in force'. Protection could come from legislation on data protection (related to employee data) or legislation on industrial secrecy (related to the algorithm and/or AI system). The second restriction refers to the confidentiality with which worker representatives must treat the information they receive. Worker representatives may not use information related to algorithms and AI 'for purposes other than those that were the reason for handing it over or [use that information] for functions that exceed their scope of competency'.

Both provisions are in line with what was already established under Article 65 ET, paragraph 4 of which states that a company will not be bound to provide worker representatives with information related to 'industrial, financial or commercial secrets whose disclosure could (…) hinder the functioning of the company [or] cause serious harm to its financial stability'. Similarly, Article 65(3) ET specifies that 'no document delivered by the company may be used beyond the strict scope thereof or for purposes other than those that were the reason for handing it over'. The collective agreement between Just Eat and CCOO and UGT thus substantially recalls the already existing legal obligations.

More than two years have passed since this collective agreement was signed but, as of January 2024, it had yet to be implemented.[5] All the players involved indicate, however, that the algorithm committee will 'soon' begin its tasks and that the outcomes of its actions, as well as the legal disputes to which it will almost certainly give rise, will be able to be analysed.

For now, however, what does exist is a certainly sophisticated legal framework on algorithmic transparency at a platform delivery company: a legal framework that was created through collective bargaining and that could very well serve as an example or guide for other collective agreements.

5. The slowness in getting the algorithm committee running was due to the election processes of worker representatives in the company having not yet ended. Until those results were known, the committee could not begin functioning. I would like to thank Rubén Ranz of UGT for the information he provided in writing this section.

3. Conclusions

Spain has pioneered legislation on algorithmic transparency in companies. Resulting from a social dialogue agreement, Article 64(4)(d) ET makes it mandatory for companies to provide information to worker representatives about the 'parameters, rules and instructions' that form the basis of the algorithms or artificial intelligence systems that are used for making decisions that 'have an impact on working conditions and on access to and keeping a job, including profiling'. The provisions of this legislation have been supplemented by those of the Charter of Digital Rights, which include not only the requirement to assess the impact of decisions adopted through algorithms or AI systems, but which also outline respect for the principles of auditability, a human-centric approach and the 'do no harm' principle. However, the provisions of the Charter of Digital Rights have no legal effectiveness; they are merely recommendations and not obligations for companies regarding how they handle algorithms and AI.

Spain also has collective bargaining experiences related to algorithmic transparency: the collective agreement in the banking sector; and the collective agreement between Just Eat and CCOO and UGT. That one of the collective agreements was negotiated for a sector and the other agreement at company level tells us that collective bargaining on algorithmic transparency can occur at both levels of negotiation (sector and company). Furthermore, both these experiences in collective bargaining on algorithmic transparency have been conducted by traditional unions: CCOO and UGT. This proves that, in collective bargaining, traditional unions are able to include content related to the digital transformation of both the economy and companies, and that they can represent new groups of workers arising out of the heat of the technological revolution. Finally, even though both these collective bargaining experiences have defined the rules of the game for companies in the exercise of algorithmic transparency, the implementation of the agreed provisions is taking longer than expected.

References

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

Bernal Santamaría F. (2020) Big data: gestión de recursos humanos y el derecho de información de los representantes de los trabajadores, Cuadernos de Derecho Transnacional, 12 (2), 136–159. https://doi.org/10.20318/cdt.2020.5605

De la Sierra S. (2022) Una introducción a la carta de derechos digitales, in Cotino Hueso L. (ed.) La carta de derechos digitales, Tirant Lo Blanch.

Hadwiger F. (2022) Realizing the opportunities of the platform economy through freedom of association and collective bargaining, Working Paper 80, ILO. https://doi.org/10.54394/VARD7939

Galdón Clavell et al. (2022) Información algorítmica en el ámbito laboral: guía práctica y herramientas sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral, Ministry of Labour and Social Economy. https://www.mites.gob.es/ficheros/ministerio/inicio_destacados/Guia_Algoritmos_ES.pdf

Ginés i Fabrellas A. (2021) El derecho a conocer el algoritmo: una oportunidad perdida de la 'Ley Rider', IUSLabor, 2, 1–5. https://raco.cat/index.php/IUSLabor/article/view/389840
Gómez Gordillo R. (2021) Algoritmos y derechos de información de las personas trabajadoras, Temas Laborales, 158, 161–182.
ILO (2022) Decent work in the platform economy, Meeting of experts on decent work in the platform economy. https://www.ilo.org/global/topics/non-standard-employment/whatsnew/WCMS_855048/lang--en/index.htm
INE (2023) Encuesta sobre el uso de TIC y del comercio electrónico en las empresas, Instituto Nacional de Estadística. https://www.ine.es/prensa/tic_e_2022_2023.pdf
Rodríguez Cardo I.A. (2022) Gestión laboral algorítmica y poder de dirección: ¿hacia una participación de los trabajadores más intensa?, Revista Jurídica de Asturias, 45, 157–172. https://reunido.uniovi.es/index.php/RJA/article/view/18991
Rodríguez Fernández M.L. (2021) Collective bargaining for platform workers: Who does the bargaining and what are the issues in collective agreements, E-Journal of International and Comparative Labour Studies, 11 (1), 61–82.
Urzí Brancati M.C., Pesole A. and Fernández Macías E. (2020) New evidence on platform workers in Europe, Publications Office of the European Union. https://doi.org/10.2760/459278

All links were checked on 27.02.2024.

Cite this chapter: Rodríguez Fernández M.L. (2024) Collective bargaining and AI in Spain, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 19
Union influence over algorithmic systems: evidence from Sweden

German Bender

1. Introduction

Algorithmic systems[1] are no longer a futuristic scenario that we only encounter in science fiction. Most of us interact with them on a daily basis and they now affect workers in virtually all sectors of European labour markets. Although algorithmic systems have rapidly become ubiquitous, they are still a rather new phenomenon. This may explain why there have as yet been relatively few cases of collective bargaining concerning these systems and even fewer that have been studied in the research literature or which have informed union strategies.[2]

This chapter discusses some aspects of union influence over algorithmic systems in the Swedish context, drawing on empirical cases and recent scholarly work. The case descriptions and introduction to the Swedish context are necessarily brief, but it is hoped that they shed some light on how unions are handling the consequences of digital technologies in the labour market.

2. Evidence from Sweden

2.1 The empirical context

It is well-known that Sweden has a highly developed industrial relations system. One aspect of this is structural and organisational, with strong unions, a union membership rate of about 70 per cent (74 per cent among white collar workers and 62 per cent among blue collar ones), collective bargaining coverage of around 90 per cent and strong employer associations (Kjellberg 2021a, 2021b).

Another aspect is the institutional setting in which the social partners operate which aims at minimising state involvement and creating a level playing field for labour and capital. This is achieved through legislation and the framework agreements between union and employer associations at central and sectoral levels which support local level collective bargaining. Importantly, the Swedish industrial relations system permits the

1. The term algorithmic systems is used here to denote digital and often dynamic (adaptive, changing, self-correcting) systems that use hardware, algorithms and large datasets to produce outcomes in specific contexts (Bender 2021; Seaver 2019). Systems used for digital automation and artificial intelligence are included in this category.
2. Due to the novel character of this research, three of the four case studies used in this chapter have not previously been published (for the other, see Selberg (2023)).

social partners to derogate from legal provisions, but only if they can mutually agree on other conditions through collective bargaining. This allocates considerable bargaining power and discretion to unions and employers, while providing a legal framework that upholds workers' rights and voice mechanisms.

For the purposes of this volume, it should also be noted that Sweden has a long-standing tradition of worker and union influence over the use of technology in working life (Ehn and Sandberg 1983; Sandberg et al. 1992). Within this sociotechnical tradition, technological innovations are seen by both employers and unions as highly contingent on the social context in which they are used and they are therefore deeply intertwined with social factors (e.g. organisational, economic and cultural ones) (Trist and Bamforth 1951). Hence, rather than focusing on the technology itself, the social partners are mostly concerned with how technology is used and the consequences it may have.

2.2 Case studies 1 and 2: collective bargaining in the platform economy

Perhaps the most publicly known example of collective bargaining on algorithmic systems is the agreement signed in February 2021 by the food delivery platform Foodora and Transportarbetareförbundet (Transport; the Swedish Transport Workers' Union). In fact, the negotiations resulted in two agreements running from 1 April 2021 to 30 April 2023: one at sectoral level (which already applied to certain companies with employees in this sector); and one especially tailored company agreement for Foodora. A precondition that made it possible to reach a collective agreement is that Foodora recognises most of its workers as employees, contrary to many other platform companies (especially in food delivery) who regard workers as self-employed (Ilsøe and Söderqvist 2022).[3]

A legal analysis of the agreements conducted by Niklas Selberg, Associate Professor in Private Law at Lund University, was presented at the International Labour and Employment Relations Association (ILERA) Conference in 2022. Selberg characterises the agreements as 'an episode of trade union renewal', but also notes that 'a number of topics relevant to the gig economy or platform work' have been left out, e.g. the duration of fixed-term employment, algorithmic management, data protection and portability, privacy and surveillance (Selberg 2021: 12). However, it is arguable that some of these issues are indirectly addressed because a wide range of issues, such as remuneration, working hours and scheduling, and health and safety, are covered. These issues are affected by the algorithmic system that allocates, directs and surveils workers, and calculates labour costs and delivery prices. Hence, the system must be adapted to the provisions of the collective agreement. In other words, direct bargaining on data and algorithms is not needed for unions to influence the digital technology that Foodora

3. There are unions in Sweden that organise self-employed workers (e.g. the white collar Unionen, Sweden's largest union), and could therefore potentially organise that category of platform workers as well, but Transport does not.

uses (c.f. De Stefano 2019). Furthermore, some of these issues are covered by national and EU labour legislation and by other EU regulations.[4]

According to Selberg, the agreement shows that the platform economy is indeed compatible with Sweden's highly regulated and organised industrial relations system that relies on regulation mechanisms like sectoral and local collective agreements supported by labour law. The Foodora-Transport deal can be seen as a first step to incorporate platform-based food delivery into the Swedish labour market model. While it certainly leaves scope for improvement, the agreement has arguably laid the foundation for better working conditions for Foodora's employees. Selberg finds that the health and safety regulations in the agreement are actually even better than the legal provisions in Arbetsmiljölagen (the Swedish Work Environment Act). While the agreement leaves out some important provisions (e.g. overtime, holidays and redundancy), other parts regulate workers' remuneration, working hours and thus may have consequences for Foodora's algorithmic system. Selberg concludes: 'On the one hand labour and employment law maintains its relevance for the future and for the platform economy, on the other hand the crucial element of this field of law in protecting the weaker party to a work contract risks being watered down and diluted.'

So far, the main union strategy in the Swedish platform economy has been to organise workers and pressure digital platforms into signing collective agreements. This is of course only possible if the platform companies acknowledge that their workers are employed, not independent contractors. In the case of Foodora, which employs most of their couriers, this strategy seemed to work for Transport after more than a year of organising, media pressure and negotiations. However, it has since surfaced that some of Foodora's couriers are not employed at the company itself but by another company named Pay Salary, owned by the Spanish temporary work agency Jobandtalent (Arbetsvärlden 2022). This led to legal action by Transport when the contract of one of its members was terminated when he demanded permanent employment after working at the company for two years. However, the Swedish Labour Court recently decided against the union's demands, siding with Foodora's argument that the courier was employed at Pay Salary (Arbetsdomstolen 2022). This is a rare case of legal action in the Swedish platform economy, and left Transport considering whether it should pressure Pay Salary into signing the collective agreement for temporary work.

In an earlier case study, Söderqvist and Bernhardtz (2019) interview representatives of three platform companies that have signed white collar and blue collar collective agreements pertaining to temporary work agencies.[5] A similarity between all these agreements and the one with Foodora is that they do not regulate data or algorithmic technologies directly but contain provisions (e.g. on wages, working hours and working

4. Selberg mentions the EU directive on transparent and predictable working conditions, the directives on fixed-term work and part-time work, and the directive on health and safety for workers with fixed-term contracts and in temporary employment relationships, as well as the General Data Protection Regulation and the European Convention on Human Rights.

5. his study shows that the Foodora agreement was not the first collective agreement in the Swedish platform economy. Apart from the three examples cited in the paper, another example of an early agreement in the platform economy was that between the ride-sharing company Bzzt and the unions Transport and Unionen. However, the Foodora-Transport deal was the first specifically tailored 'platform agreement'.

conditions) that indirectly affect their design and implementation as well as the business models they serve. An additional similarity is that all three companies, as well as Foodora, highlight the public relations benefits of signing a collective agreement, branding the companies as a 'fair option' for both employees and clients (Söderqvist and Bernhardtz 2019: 5).

In terms of digital technology itself, an interesting aspect raised by one platform company is the possibility of digitally implementing collective agreements in their algorithms to be able to update any changes or new provisions automatically. The firm is already automating functions in its algorithms to ensure compliance with collective agreements and, in the words of one respondent: 'The dream would be to download code for the different agreements' (Söderqvist and Bernhardtz 2019: 7). As this firm has signed separate agreements for blue and white collar workers, the authors suggest that this type of automated digitalisation might improve regulatory compliance and lower transaction costs for both firms and trade unions.

2.3 Case study 3: collective bargaining on digital automation

A rare case of collective bargaining on algorithmic systems is, however, studied in a research paper that the author is working on with Fredrik Söderqvist, PhD candidate at Blekinge Technical University. The paper is based on 26 interviews conducted with local union and employer representatives at the mining company Boliden in northern Sweden, with this primary data being complemented with secondary data from bargaining protocols, collective agreements, company documents and other written material. Specifically, the paper studies how digital automation and the digitalisation of mining operations is codetermined by the social partners at local level. This includes the installation of a wifi-based location system in the extended tunnel grid of underground mines one kilometre below ground; the implementation of semi-automated vehicles during night-time; and the remote operation of heavy machinery. It is important to stress that automation and digitalisation is viewed by both the firm and the local unions (one blue collar and three white collar) as necessary to increase productivity and safety.

The overall findings are that, similar to the conclusions reached by Selberg about the Foodora deal, unions and employers are able to codetermine or bargain on algorithmic systems in the workplace to some extent, and largely do so by regulating their use conditions and effects. For instance, unions in Boliden do not examine the source code of automated vehicles, or the data used to develop their algorithms, but instead strive to safeguard that the vehicles are used in ways that are less detrimental to workers. As leverage in these negotiations, unions can use national labour law which obliges employers to inform and negotiate with unions before making organisational or technological changes that affect employees. Furthermore, various kinds of EU regulations provide unions and workers with influence over employer initiatives regarding digital technology (see Footnote 2 above).

In one of the cases, the firm wanted to use semi-automated vehicles during the night in order to increase productivity. This would only be possible where the unions agreed

to derogate from the legislation and from the sectoral agreements regulating working hours, thereby providing unions with leverage over how this type of automation would be implemented. Another interesting case is the wifi-based location system, which allows the employer to track each individual worker in the mines. In order to implement this digital system, the unions demanded (successfully) that all personal data be anonymous; that only specifically designated supervisors be allowed to identify each tag (an identification number) with its correspondent worker; and that de-anonymisation could only be permitted in the case of an emergency (e.g. an accident or a fire). Moreover, data generated by the system can only be used for specific security-related purposes and not to calculate – for example – workers' productivity, working patterns or remuneration.

2.4 Case study 4: workers' voice and artificial intelligence

Although not a case study in the strict sense, a recent conference paper by Samuel Engblom (2021), legal scholar and then Policy Director at the Swedish Confederation of Professional Employees (Sweden's largest white collar union confederation), presents interesting insights into worker influence over algorithmic technologies for unions and labour scholars alike. The purpose of the paper is to 'is to explore how the introduction of algorithms in working life affects the possibilities of workers' representatives on the workplace level to effectively exercise the rights given to them through legislation and collective bargaining agreements' (Engblom 2021: 1). Drawing on interviews[6] with white collar union members in the public and private services sector, the author outlines a framework for assessing the extent to which mechanisms for worker voice (e.g. labour law and collective agreements) provide employees with influence over algorithmic systems. The paper tentatively proposes three questions that can be posed in this context (Engblom 2021: 5):

1. How well does the mechanism cover all situations where digital tools have an impact on working conditions or the work environment? An important issue here is whether the introduction or modification of these technologies triggers the voice mechanisms. In the Boliden case, the employer was obliged by Swedish labour law to inform and negotiate with local unions before implementing automation technologies. Further, the use of data for non-specified purposes would have triggered regulations agreed by the social partners through collective bargaining.
2. What is the ability of the mechanism to handle the dynamic development of digital tools and their subsequent updates? A key challenge here is that many systems make piecemeal incremental changes (sometimes as inconspicuous software updates) which, eventually, can have significant effects on workers. It is not clear if and when digital updates or plug-ins can or should be subject to workers' voice mechanisms.

6. Interviews were conducted with four full-time union representatives in a government agency, a region (an administrative entity in Sweden, which mainly governs healthcare and public transport) and a multinational bank. The bank interview was complemented with a group discussion with Nordic union representatives.

3. To what extent does effective worker voice require technical expertise and transparency? One concern here is that many union representatives do not have the technical expertise to assess the hardware or software of algorithmic systems. However, the need for this type of insight and influence may be overstated. Instead, Engblom's respondents suggest that narrow technological aspects should not take focus away from the core issue, which is the impact of technology on workers. As expressed by one of the union representatives: 'I'm a practical person. I don't care what is going on inside the black box as long as it does what I want it to do, it works and is easy to use.'

Algorithmic systems are increasingly based on artificial intelligence, e.g. machine learning and other dynamic digital technologies, that adapt and change in ways that are difficult to predict or understand even for the companies that develop or implement them. This makes it even more difficult for union representatives to assess if or when changes are significant enough to require re-negotiation or other forms of worker involvement or regulation.

The rationale for the three questions proposed in Engblom's paper is to develop a tentative framework to handle novel and unforeseen challenges for worker voice mechanisms. This type of pre-emptive and cooperative bargaining strategy is typical of Swedish unions. Legal action directly concerning algorithms is rare; the only case of which the author is aware is a complaint filed in 2020 to Justitieombudsmannen (JO; the Parliamentary Ombudsman) by Akademikerförbundet (SSR; the Union for Professionals), arguing that the municipality of Trelleborg should disclose an algorithm used to automate decisions on social benefits. The union argued that the algorithm, or its specification, should be public information available to all citizens, especially those affected by a municipal decision based on this technology.

3. Discussion and conclusions

This chapter offers a brief overview of the current actions and strategies that Swedish unions have taken to address the challenges posed by algorithmic systems in working life. There are a number of conclusions we can draw from this overview.

First, the Swedish labour movement has (so far) largely eschewed legal action against employers. Instead, the main strategy has been to push for collective bargaining and other forms of worker voice mechanisms.

Second, union actions and strategies have not primarily been aimed at influencing the technology itself, for example by negotiating on the source code. Instead, they have focused on what might be called sociotechnical bargaining, attempting to influence the contexts and the consequences of algorithmic systems.

A third conclusion is that, so far, Swedish labour law and mechanisms for worker voice have mostly been able to accommodate algorithmic systems and business models, allowing for some extent of union influence over these technologies. However, as noted

by Selberg (2021: 11), 'being included in labour and employment law does not preclude a very precarious situation in the labour market. In the Swedish context it is possible to shift a lot of the enterprise risk onto an employee.' We should also note that the problem on the misclassification of workers as self-employed still poses a problem for the Swedish labour market model (Ilsøe and Söderqvist 2022), albeit affecting a small part of the workforce (Bender 2022; Sabanova and Badoi 2022).

In summary, labour law is a necessary but insufficient safeguard in the modern labour market, increasingly permeated as it is by dynamic algorithmic systems. Strengthening union influence will be essential to prevent the further erosion of vital societal interests such as wages, working conditions and employment security.

References

Arbetsdomstolen (2022) AD 2022 nr 45, Mål nr A 154/21, Swedish Labour Court. https://www.arbetsdomstolen.se/sv/meddelade-domar/arkiverade-domar/2022/2022-11-16-ad-2022-nr-45/

Arbetsvärlden (2022) AD-dom om Foodora på tvärs med EU:s linje om gigjobb, arbetsvarlden.se, 21 November 2022. https://www.arbetsvarlden.se/ad-dom-om-foodora-pa-tvars-med-eus-linje-om-gigjobb/

Bender G. (2021) Algorithmic control, Social Europe, 12 February 2021. https://socialeurope.eu/algorithmic-control

Bender G. (2022) Online platforms and platform work – Sweden. Mapping platform economy, Friedrich-Ebert-Stiftung. https://nordics.fes.de/e/factsheets-online-platforms-and-platform-work

De Stefano V. (2019) 'Negotiating the algorithm': Automation, artificial intelligence, and labor protection, Comparative Labor Law and Policy Journal, 41 (1), 15–47. http://dx.doi.org/10.2139/ssrn.3178233

Ehn P. and Sandberg Å. (1983) Local union influence on technology and work organization: Some results from the DEMOS Project, Arbetslivscentrum.

Engblom S. (2021) Algorithms, trade unions, and effective workers' voice, Paper presented at The 19th ILERA World Congress, Lund University, 21-24 June 2021, International Labour and Employment Relations Association.

Ilsøe A. and Söderqvist C.F. (2022) Will there be a Nordic model in the platform economy? Evasive and integrative platform strategies in Denmark and Sweden, Regulation and Governance, 17 (3), 608–626. https://doi.org/10.1111/rego.12465

Kjellberg A. (2021a) Den svenska modellen 2020: pandemi och nytt huvudavtal, Arena Idé. https://arenaide.se/rapporter/svenska-modellen-2022/

Kjellberg A. (2021b) Sweden: Collective bargaining under the industry norm, in Müller T., Vandaele K. and Waddington J. (eds.) Collective bargaining in Europe: Towards an endgame, ETUI, 583–603. https://www.etui.org/publications/books/collective-bargaining-in-europe-towards-an-endgame-volume-i-ii-iii-and-iv

Sabanova I. and Badoi D. (2022) Online platforms and platform work: The complex European landscape. Mapping platform economy, Friedrich-Ebert-Stiftung. https://futureofwork.fes.de/our-projects/mapping-platform-economy/report

Sandberg Å. et al. (1992) Technological change and co-determination in Sweden, Temple University Press.
Seaver N. (2019) Knowing algorithms, in Vertesi J. and Ribes D. (eds.) DigitalSTS: A field guide for science & technology studies, Princeton University Press.
Selberg N. (2021) Autonomous regulation of labour in the gig economy: Collective bargaining for food delivery workers in Sweden, Paper presented at The 19th ILERA World Congress, Lund University, 21-24 June 2021, International Labour and Employment Relations Association.
Selberg N. (2023) Autonomous regulation of work in the gig economy: The first collective bargaining agreement for riders in Sweden, European Labour Law Journal, 14 (4), 609–627. https://doi.org/10.1177/20319525231178980
Söderqvist F. and Bernhardtz V. (2019) Labor platforms with unions: Discussing the law and economics of a Swedish collective bargaining framework used to regulate gig work, Working Paper 2019:57, Swedish Entrepreneurship Forum. https://entreprenorskapsforum.se/wp-content/uploads/2019/03/WP_57-1.pdf
Trist E.L. and Bamforth K.W. (1951) Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system, Human Relations, 4 (1), 3–38. https://doi.org/10.1177/001872675100400101

All links were checked on 27.02.2024.

Cite this chapter: Bender G. (2024) Union influence over algorithmic systems: evidence from Sweden, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

Chapter 20
AI systems, risks and working conditions

Vincent Mandinaud and Aída Ponce del Castillo

1. Introduction

Technological change is at the core of work. Workers have been exposed to industrial automation in many forms and for many decades. Today, they are able to work in physical and digital workplaces, as well as in hybrid ones, and are exposed to AI systems in their many implementation modalities. Their exposure to and interaction with this technology are diverse and can be experienced before the beginning of the employment relationship; that is, during the recruitment and selection process, including through automated interviews and tests. Once the employment relationship has started, the exposure and interaction continue, whether it be with tools or applications such as virtual assistants and chatbots, software for task allocation, robotic machines, drones, computer vision devices, embedded algorithms, etc. Workers can work with AI systems, developing or maintaining AI; they can be managed by AI; and they interact with AI in many other ways (Lane et al. 2023).

This chapter reflects on the conditions under which artificial intelligence systems might not harm or worsen working conditions but instead contribute to enhancing and improving them. How can they be deployed and controlled to ensure this? And how can regulation, collective bargaining and other mechanisms contribute? In reviewing these questions, the aim is to seek to shed some light on the design, development, introduction and use of AI systems in the workplace.

The hypothesis is that the improvement in working conditions brought about by AI does not depend solely on the qualities of these technical systems, even if they embody by design a certain vision of society and work. It also, and perhaps above all, depends on the ability of social systems to supervise these tools and their uses so as to put them at the service of a job well done and in good conditions. In other words, the impact of AI systems on working conditions depends on the quality of the composition of the technical chain, as well as on the ability to act in the interests of human integrity, health, safety, performance and democracy in the workplace.

2. The ambivalence of technologies and work organisations

Like digital technologies in general, AI systems are ambivalent about working conditions. As Stiegler (2015) points out, following Derrida (1972), they constitute a pharmakon: they allow poisons to become remedies and remedies to become poisons depending

on the situations and uses to which they are put. They can also serve as scapegoats.[1] In other words, they are fundamentally promising and simultaneously threatening; they can strengthen the factors that cause deteriorating working conditions alongside producing fresh approaches to managing the hazards of that decline. They may aggravate conventional occupational or organisational risks but, in trying to alleviate or sidestep these, they may create novel hazards of their own (Verkindt 2020).

For example, by facilitating the processing of large masses of data, AI opens up interesting prospects in accidentology and epidemiology. Solutions for monitoring work environments also open up prospects for detection, warning and sustainable prevention of workplace risks. The development of teleoperations and collaborative robotics does help reduce or eliminate certain types of exposure. On the other hand, by positioning themselves at the heart of the organisation, these technologies can relegate human work to the background. Their use can lead to a focus on the risks they are able to detect, leaving aside those that escape them because of atypical situations or specific organisational dynamics. Alertness tools can generate psychosocial risks and lead to the individualisation of occupational health and safety issues (INRS 2022). They can also expose workers to risks relating to the non-respect of fundamental rights, through the use of automated data processing, to prevent this or that traditional occupational risk.

These ambivalences are also reflected in workers' perception of AI. An OECD survey in the manufacturing and finance sectors of seven countries asked workers who use AI systems whether they felt it had improved or worsened their performance, enjoyment, mental health and wellbeing, and physical health and safety, as well as how fairly they felt management treated them. Workers in the finance sector reported that AI had improved their performance (79%), enjoyment (63%) and mental health (54%), either by a little or by a lot. Workers in manufacturing reported the following corresponding figures: 80%, 63% and 55%, respectively. When asked how fairly their manager or supervisor treated them, workers in the finance sector (45%) and in the manufacturing sector (43%) stated that AI had improved fairness in management. The authors highlight that these findings suggest that AI, when utilised correctly, can contribute to higher productivity and better job quality. However, the report also states that the impact on performance and working conditions depends on how workers interact with AI systems. It further points out that workers' confidence in AI depends on their degree of training, information and consultation (Lane et al. 2023).

3. An AI-related risk mapping

In light of the rapid and continuing progress of AI, top AI researchers across the world have proposed urgent priorities for AI risk control and governance, arguing that, if managed carefully and distributed fairly, advanced AI systems could help humanity (Bengio et al. 2023). However, the downside is that, alongside their capabilities, come large-scale risks that society is not on track to handle well. The workplace can be an example of how those risks might materialise in a given context.

1. An explanation (in French) of the term pharmakon: https://arsindustrialis.org/pharmakon

As AI may affect all industries and occupations (OECD 2023), as well as routine and non-routine tasks, understanding its full impact on working conditions is a complex task. This section provides a 360-degree view of the possible risks that AI can produce at work. To define a comprehensive approach, these have been clustered by dimensions: risks related to organisational operations; risks related to work organisation itself; and risks related to the human dimension. The risks are not exhaustive, can overlap in time and, according to the specific context, migrate from one dimension to another.

To prevent, mitigate or eliminate risks adequately, it is key to identify the hazards arising from AI systems that can result in harm, the vulnerabilities intrinsic to an AI system and the sector-specific characteristics of a given workplace. The risk dimensions being mapped here are transversal to any organisation, but a layered analysis is helpful to a better understanding of their impacts. More specific analysis from a sectoral point of view is also needed. Knowing that risk assessment is pivotal in governing AI at work, and that new European regulations will be implemented in this regard, this section can serve as a guide to risk assessment related to AI systems at work, taking into account that the AI supply chain is complex and non-transparent with other actors, such as the developers or providers of AI systems, cloud providers and third parties, perhaps needing to be involved.

The first dimension relates to organisational operations. The AI market is not settled yet: there are many types of AI (symbolic, generative, narrow and probably general, etc.) and the adoption of AI tools is concentrated among large companies, on the one side, and 'young' firms with relatively high productivity on the other (Calvino and Fontanelli 2023). In their operations, companies adopt AI systems mainly for automation, decision-making support and to reduce staff costs, beyond concerns about the return on investment and strategic vision. The major challenges that arise are related to cybersecurity, security breaches and intrusions; privacy; data management; computational resources; and scalability. However, they also relate to meaningful automated decision-making and explainability (Bérubé et al. 2021; Dvorack et al. 2023; Shaw et al. 2019) as well as third party risks (Buehler et al. 2021).

The second dimension relates to work organisation. When attempting to modernise work and companies, the implementation of AI systems can also pose a challenge to work organisation. Work organisation is understood as the coordination and control of work, the division of work into tasks, the bundling of tasks into jobs and assignments, the interdependence between workers, and how work is coordinated and controlled to fulfil the organisation's goals (Eurofound 2023). However disruptive it may appear, the use of AI systems can refine, consolidate and complement existing managerial models such as Taylorism or Toyotaism, and processes such as productivism or extractivism, enabling companies to govern the day-to-day lives of workers in a way that is unprecedented in history (Ferraras 2023). Research by Paola Tubaro, Antonio A. Casilli and colleagues has identified and studied the impact of these systems on the working conditions of 'vulnerable' workers (Tubaro et al. 2022; Tubaro et al. 2020). The implementation of AI can radicalise power issues within the company (Ferreras 2023) and have an impact on the relationship of subordination between employer and workers (Aloisi and De Stefano 2022). By using automated decision-making and monitoring systems, usually known as

algorithmic management,[2] employers can technically govern the daily lives of workers (Ferreras 2023), while the outcomes can be biased or lead to discrimination. Chatbots can be used to communicate with workers, replacing basic human and personal interactions and rendering human communication fragmented and ineffective.

AI systems can also influence working conditions through work intensification. Generative AI[3] is used to write documents more quickly and allow workplaces to become increasingly multimodal. AI systems can expose workers to new forms of surveillance enabled by the collection and exploitation of individual and collective data in the name of performance or, sometimes, in the name of occupational health and safety or security (Ponce del Castillo and Molè, this volume). They can expose workers to their inability to define, organise and carry out quality work, infuse tacit knowledge and make informed decisions. Workers who are most affected by the implementation of AI systems may belong to the most vulnerable groups with lower power and agency (Curtis et al. 2023). And they can produce or intensify discriminatory practices (Pasquale 2015) and create further inequalities.

When working with AI systems integrated in machines, robots or cobots, besides being exposed to the risks derived from joint human-robot activities such as technical design constraints, sensing and zoning, situational awareness in relation to safety risks, malfunctions and program changes that may lead to physical injuries (Jansen et al. 2018), workers can be exposed to other risk factors such as sensory degradation or other environmental factors.

AI systems can have an impact not only on jobs, but also qualifications, skills and identities (Benhamou 2022). The skills needed to work alongside AI vary from sector to sector. Already, generative AI can perform many tasks that previously required 'social intelligence' (Frey and Osborne 2023). For automation, Acemoglu and Restrepo (2018) argue that the skills needed would be a mix of numeracy, communications and problem-solving skills. For AI and generative AI, what mix of skills do workers need to identify deep fakes, unreliable content and wrong or even malicious recommendations?

Moreover, outside the employment relationship, since AI systems can be used in the recruitment process to advertise vacancies, select candidates, filter applications and evaluate candidates (European Commission 2021: Annex III.4), the act of bringing people into a company can increasingly be left to automated decisions, rendering it a purely technical process.

---

2. Automated monitoring and decision-making system, or algorithmic management, is defined here as automated or semi-automated computing processes that perform one or more of the following functions: workforce planning and work task allocation; dynamic piecerate pay setting per task; controlling workers by monitoring, steering, surveiling or rating their work and the time they need to perform specific tasks, nudging their behaviour; measuring actual worker performance against the predicted time and/or effort required to complete a task, and providing recommendations on how to improve worker performance; and penalising workers, for example through the termination or suspension of their accounts. Metrics might include estimated time, customer ratings or a worker's rating of customers (Ponce del Castillo and Naranjo 2022).

3. Following García-Peñalvo and Vázquez-Ingelmo (2023: 14) 'the general public commonly uses the term "Generative AI" to refer to the creation of tangible content (such as images, text, code, models, audio, etc.) via AI-powered tools. However, the AI research community primarily discusses generative applications focusing on the models used, without explicitly categorizing their work under the term "Generative AI"'.

The third dimension relates to the human aspect. AI systems can be deployed at a detailed level in an organisation and can make real-time 'decisions' about workers, plan and allocate tasks to them or discipline them. Working conditions could deteriorate if the tools provided reinforce strategic and organisational orientations that endanger the physical and mental integrity of workers. Robust research shows that the implementation of new technology can be associated with job strain and an increased pace of work (Jansen et al. 2018) which can result in musculoskeletal disorders (Cippelletti et al. 2023) or poor psychological health. It can create stressors – technostress – including work overload, role ambiguity and job insecurity (Atanasoff and Venable 2017; Stadin et al. 2016; Stamate et al. 2021). Researchers from Stanford University have delved into the psychological and psychosocial impacts of AI and suggest that it can contribute to a degradation of workers' autonomy and control, but also to their demoralisation and discontent (Luxton and Watson 2023) or simple boredom (Jansen et al. 2018).

Additionally, the working environment can influence the mental and physical condition of workers, with risks that go beyond working life into the personal sphere. Workers are at risk of being monitored and profiled on the basis of their behaviour, reputation, physiology, biometrics and even their 'emotions'. (For an in-depth analysis of worker monitoring and surveillance, see Ponce del Castillo and Molè, this volume.)

Preventing and managing AI-related risks in the workplace is a good opportunity to rethink the relationship between humans and technology and to avoid AI systems posing societal-scale risks: acceleration of the existing inequalities and social injustice; erosion of social stability; and a weakening of our shared understanding of reality. As leading AI scholars expressed in their joint paper for the OECD (Bengio et al. 2023), 'without sufficient caution, we may irreversibly lose control of AI systems, rendering human intervention ineffective'. The question is how to deploy AI systems in the workplace in ways that ensure social intelligence prevails and that current working conditions improve?

4. The context: purpose and conditions of use in heterogeneous workplaces

The impacts of AI systems in the workplace certainly depend on how they are conceived and developed, how they incorporate values and representations, and how they are put into practice in a specific workplace. Equally important, they also depend on how workplaces absorb them, on the context of their use, on how they are introduced and used, and on how their various components frame and regulate their design, use and effects. The issue is not only the 'black box', understood as the systems, products or services that use the computer models created by training data representing the context in which they will be used (Galanos and Stewart, this volume). The role played by institutional contexts, organisational and management models and business practices in the possible improvement or deterioration of working conditions along the entire value chain, from production to use, must also be recognised.

The objective of some AI systems is to mimic human behaviour, although they are not yet able to understand context and, hence, to reason like a human. If we take one of the traditional and general definitions, the Larousse dictionary states that such systems are a 'set of theories and techniques used to create machines capable of simulating human intelligence'. For the European Parliament, AI systems are technical systems capable of perceiving their environment, managing these perceptions, solving problems and taking actions to achieve a specific goal (European Parliament 2023). However, as research on the definition has evolved, academic literature reminds us that AI does not really replicate human behaviour, even if it gives the impression of doing so (Galanos and Stewart, this volume). Antonio A. Casilli explains that we are dealing with systems that are maintained and fed by crowdworkers organised into large geographical areas – mainly from the Global South – or, rather, into language areas (Casilli, this volume).

From a natural resource perspective, Kate Crawford (2022) explains that AI systems are alloys of minerals, sweat, tears, data, classifications and prejudices. Harry Collins (2019), on the other hand, uses the term 'artifictional intelligence' to mean that AI systems cannot (for the time being, or perhaps ever) reproduce human language in action since the latter is bound up within specific contexts and thus tacit sociocultural backgrounds of meaning.

Indeed, the technological black boxes (Pasquale 2015) that make up AI systems (especially those resulting from machine learning, deep learning and even generative AI) are often highlighted and criticised for their opacity and the power, orientations and discriminations they conceal (Masure 2019). Incidentally, if there is something specific about these new technological black boxes, it is that they are not simply opaque to the average person (as is the case with most conventional digital tools), but also to experts in the field.

However, we have to acknowledge that one black box can hide or reveal another. The reality of work and of the organisations in which AI systems are implemented also constitutes a black box. First and foremost, this is so for the designers and suppliers of these technologies who claim to be able to improve performance, and sometimes even working conditions, in organisations without these technologies being made discussable, intelligible and adaptable by all the constituent parts of the organisation and without being able to guarantee the explainability of the choices made by AI systems.

The reality of the work carried out in companies is too often overshadowed by a certain functional vision which does not correspond to that of the people who ultimately carry out the tasks or make decisions at various levels of an organisation. And even if they were better taken into account, this could continue to obscure the reality of work up and down the value chain, as well as its impact on human health and ecosystems. The technological black box reveals the black box of the organisational model and its decision-making processes which, in turn, hides the reality of work and masks the voice of workers and their representatives regarding the tools, organisation, working conditions and quality of their work (Sennet 2000). This, for its part, covers up the reality of ecosystem degradation and disguises the contributions of the non-humans

who nonetheless create the conditions for human life (Latour 2015) and human work (Friedmann 1975).

With artificial intelligence or in its absence, one cannot improve working conditions against workers or without them. If AI is to improve working conditions, then workers and their representatives must have a say in the design, development, introduction, testing, evaluation, deployment and monitoring of AI in the workplace as well as in the strategic direction of the organisational models in which they are embedded and with which they share their intelligence and labour. In summary, workers need to have an active role in the lifecycle of the AI systems to which they are exposed. This is one of the aspects discussed in the agreement on the digital transformation of companies signed by the European Social Partners in June 2020, the Autonomous Framework Agreement on Digitalisation.[4] This Agreement must contribute to ensuring that the activities that underpin and/or are supported by the development of AI do not become part of a process of deporting the problems of working conditions to the other side of the world, or of 'technological zombification' (Monnin et al. 2021); that is, technological development that contributes to the construction of unsustainable infrastructures and the degradation of the world we live in and the world off which we live (Charbonnier 2020).

5. The place and role of workers' representatives in the regulation of AI systems

When deploying AI systems at work, accountability, transparency and explainability for the employment-related decisions supported by AI are essential prerequisites (OECD 2023). To ensure that workers benefit and that their working conditions are improved when AI systems are in place, a clear and efficient regulatory framework is needed.

It is important to re-emphasise here the importance of the Framework Agreement on Digitalisation in that this has paved the way and offers points of support for various forms of implementation in the various Member States of the European Union. However, three years after it was signed, this Agreement has not yet resulted in national legislation, regulation or agreements that translate its letter and spirit into concrete actions in companies and value chains in the European Union so that the voices of workers and their representatives are heard in defence of decent working conditions and the sharing of value. It does have to be acknowledged here that workers' representatives have not, in general, recently been in a strong position to reach such agreements in a workplace setting, allowing them to regulate AI, although good examples do exist (see Rodríguez Fernández, this volume).

4. To check the state of implementation of the European Social Partner Framework Agreement on Digitalisation, see: What's happening already at national level and how to support social partners to implement? 27 April 2021. https://resourcecentre.etuc.org/european-social-partner-framework-agreement-digitalisation-whats-happening-already-national-level

The European Commission, through its legislative programme, appears to be a powerful player in this respect even though its AI Act is not specifically designed to provide the guarantees on AI needed in the workplace; indeed, it is rather more concerned with AI on the European product market. Moreover, there is little room for employee representatives to participate and it is clear that legislation will therefore be necessary in the future when it comes to workplace applications of AI.

On the other hand, civil society actors seem to have only limited ability to challenge or contribute to regulation, whether they are human rights or civil liberties associations (such as Access Now, Amnesty Tech, European Digital Rights Network (EDRi), among others), or indeed trade unions, even though their work can be remarkable in many respects. 2023 was a year when we saw a number of leading figures in the AI industry call for the regulation of technological development, with some even arguing for a moratorium, as a means of preventing AI from endangering our societies and possibly humanity itself.[5] This communications exercise shows how the champions of innovation in the field are also making the effort to establish themselves as champions of regulation,[6] even if it means taking up all the bandwidth, or at least trying to do so. In doing this, they have developed a discourse on an ethical AI, even if built on the underpinnings of a dystopian arena, which would render these new captains of industry, in association with renowned scientists, leading actors responsible for the reliability of the programs they intend to make openly available.

Mapping the different regulatory arenas for AI (Benbouzid et al. 2022) shows that regulation has become a competitive field in which workers' representatives are struggling to make themselves heard.

The polarisation is no longer between the proponents of innovation and the proponents of regulation. Innovation players have taken over the regulatory arena, to some extent dispossessing workers' representatives of their preferred space. The landscape of AI regulation, understood as the social control of AI, as set out by the authors in this volume, reveals different regulatory regimes but, above all, a certain transformation of the rules of the game and the balance of power.

That said, the role of workers' representatives in preserving if not improving working conditions, recognised as such in national legislation or practice regardless of whether as trade union representatives or as elected officials, as stipulated in ILO Convention 135, remains important. There is an abundance of academic literature on the role of

5. See the open letter 'Pause Giant AI Experiments: An Open Letter', published on 22 March 2023 by the Future of Life Institute https://futureoflife.org/open-letter/pause-giant-ai-experiments/

6. Some examples of the various interviews and press releases of AI developers related to the governance and regulation of AI are as follows: Shariatmadari D. (2023) '"I hope I'm wrong": the co-founder of DeepMind on how AI threatens to reshape life as we know it'. Interview, The Guardian. https://www.theguardian.com/books/2023/sep/02/i-hope-im-wrong-the-co-founder-of-deepmind-on-how-ai-threatens-to-reshape-life-as-we-know-it; Yun Chee (2023) 'Exclusive: EU's Breton to discuss AI rules with OpenAI CEO', Reuters. https://www.reuters.com/technology/eus-breton-meet-openai-ceo-san-francisco-june-eu-officials-say-2023-05-30/; Zakrzewski et al. (2023) 'Tech leaders including Musk, Zuckerberg call for government action on AI'. *The Washington Post*. 13 September. https://www.washingtonpost.com/technology/2023/09/13/senate-ai-hearing-musk-zuckerburg-schumer/

workers' representatives which should serve as a point of reference for taking a stand in this context of change. Representatives of workers are needed most of all to ensure that AI systems are designed, developed and used responsibly in the workplace. To do this, in addition to their traditional role and within the frame of reference that is labour law, they must also take on aspects relating to the use of data – especially personal data – in organisations and must rely, at the very minimum, on the GDPR given the current state of legal support.

At company level, knowing their working environment, they can use access to expert and non-expert knowledge to identify risk of harm signals, particularly where there is a multiplicity of technologies converging in work processes or in the use of AI (generative or otherwise) for certain tasks. They can also play a key role in impact assessments, whether they be occupational health and safety risk assessments, data protection impact assessments under the General Data Protection Regulation or fundamental rights impact assessments. They can thus play a key role in identifying abusive practices as they are relevant players in helping to clarify privacy and data protection rights within workplace organisation.

To achieve this, workers' representatives need to be able to train and update their frame of reference, as well as promote and participate in experiments in other ways of effectively, safely and democratically integrating AI systems into work organisations. They must also be able to train (and not just in law, but also in design, management, sociology and other disciplines) and acquire the skills to understand how AI systems work. This will allow them greater opportunity to influence them so that they can contribute constructively to the decisions, actions and projects likely to give birth to a trustworthy AI.

But it is not all about the company. Far from it. It is also in the interests of the organisations of workers' representatives to articulate in a better fashion the different scales involved: from the shopfloor to the company, from the company to the group, to the territory, to the sector, to the industry, to the national and transnational levels. Only if workers' representatives know better how to combine professional, multi-professional, organisational and institutional logics will they be able to forge new alliances to have a greater say in the regulation of AI.

6. Conclusion

AI is a game changer for the world of work in that it not only challenges the way work is organised, and how workplaces are equipped to produce goods and services, but also the way in which workplaces are structured to produce quantities and qualities of jobs and work. The design, development, deployment and use of AI systems therefore raise numerous challenges for workers' representatives. This volume focuses on issues relating to the working environment at large in the attempt to show that analysing the consequences of AI systems for working conditions is inherently complex. Moreover, these consequences are also contextual and inseparable from the ability of workers' representatives to influence the direction of technical systems and the organisational

models in which they are embedded, and which they may either act to reinforce or to weaken. Thus, in seeking to challenge the omnipotence sometimes attributed to AI systems, whether beneficial or harmful to working conditions, the contributions in this volume highlight the relative weakness of workers' representatives when it comes to regulating AI systems in work organisations and their relative relegation in the face of market forces, science and industry in particular.

We would like to conclude by discussing two aspects of AI's game-changing character for the world of work.

First, many of the difficulties encountered in the deployment of AI systems in work organisations are similar to those already encountered with other types of new technology. The difficulties are expressly similar when it comes to the design and development of such projects, and also their management. Workers and representatives are mobilised only too late, at the end of the process, and are barely involved in project orientation whether on the scale of company projects or on the scale of structuring programmes for sectors or industries. From this point of view, if nothing changes in the way AI systems are deployed, there is a strong likelihood that organisations will come up against the same pitfalls as before and that AI systems too will fail to live up to all the promises they make, both in terms of their contribution to improving organisational performance and to improving working conditions within these organisations. Paradoxically, however, the complexity of this type of technology, and the high stakes associated with it, may lead us not only to raise the bar in project management, thereafter to recognise the need to improve the skills of those who design and deploy AI systems in organisations by bringing them closer to the realities of work, but also to place a much greater value on the role of workers and their representatives in the success of transformation projects.

Second, whilst a certain number of working conditions issues associated with the deployment of AI systems falls within the scope of classic project management practices, AI systems do, however, place more specific pressure than other digital technologies on issues linked to decision-making in work organisations. Indeed, insofar as they simulate human cognitive capacities, they compete with the ability to tell, to read, to see, to recognise, to translate, to organise, etc., as well as with the ability to decide. This is why automated decisions are governed by the GDPR and why the place of working men and women is the subject of such intense attention from the social partners. Although AI systems do not replace decision-makers, they are supposed to help them to decide, and to decide better; it being understood that this 'deciding better' would in fact be more realistically read as 'deciding more justly, because more objectively and more rationally'. In this way, they help to reinforce the idea that decisions are, and/or should be, as rational as possible in order to be as fair as possible.

This perspective is perhaps debatable. Deciding, or justly deciding, is not just a cognitive process, but a rational one. Decisions, like indecisions, are sociopolitical gestures, sometimes passionate, sometimes unreasonable, with limited rationality or with plural rationalities. If AI seeks to surpass legal normativity with a social normativity that it is capable of bringing to light by exploiting vast datasets, and thus to compete with the law in saying what the right rule or the right decision should be, it should not be the case

that, in depoliticising the decision through sub-political techniques, it participates in depoliticising work, organisations and working conditions themselves. Moreover, at the same time neither should it deprive the possibility of the collective determination of the conditions that not only enable work to be carried out, but which are also ameliorated by work.

References

Acemoglu D. and Restrepo P. (2018) The race between man and machine: Implications of technology for growth, factor shares, and employment. American Economic Review, 108(6), pp. 1488-1542.

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

Atanasoff L. and Venable M.A. (2017) Technostress: Implications for adults in the workforce, The Career Development Quarterly, 65 (4), 326–338. https://doi.org/10.1002/cdq.12111

Benbouzid B., Meneceur Y. and Smuha N. (2022) Four shades of regulation of artificial intelligence: A cartography of normative and definitional conflicts, Réseaux, 232-233 (2-3), 29–64.

Bengio Y. et al. (2023) Managing AI risks in an era of rapid progress. https://managing-ai-risks.com/

Benhamou S. (2022) Les transformations du travail et de l'emploi à l'ère de l'intelligence artificielle : évaluation, illustrations et interrogations, CEPALC, Nations unies. https://hdl.handle.net/11362/48529

Bérubé M., Giannelia T. and Vial G. (2021) Barriers to the implementation of AI in organizations: Findings from a Delphi study, Proceedings of the 54th Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.805

Buehler K., Dooley R., Grennan L. and Singla A. (2021) Getting to know – and manage – your biggest AI risks, McKinsey and Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/getting-to-know-and-manage-your-biggest-ai-risks#/

Calvino F. and Fontanelli L. (2023) A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity, OECD Science, Technology and Industry Working Papers 2023/02, OECD Publishing. https://doi.org/10.1787/0fb79bb9-en

Casilli A.A. (2019) En attendant les robots : enquête sur le travail du clic, Seuil.

Charbonnier P. (2020) Abondance et liberté, une histoire environnementale des idées politiques, La découverte.

Cippelletti E., Azouaghe S., Pellier D. and Landry A. (2023) Assessing MSDs before introduction of a cobot: Psychosocial aspects and employee's subjective experience, Relations industrielles/Industrial Relations, 78 (1). https://doi.org/10.7202/1101311ar

Collins H. (2019) Sociologie méta-appliquée et intelligence artificielle, Zilsel, 5, 161–173. https://doi.org/10.3917/zil.005.0161

Cramarenco R.E., Burcă-Voicu M.I. and Dabija D.C. (2023) The impact of artificial intelligence (AI) on employees' skills and well-being in global labor markets: A systematic review, Oeconomia Copernicana, 14 (3), 731–767. https://doi.org/10.24136/oc.2023.022

Crawford K. (2022) Contre-atlas de l'intelligence artificielle, Zulma.

Curtis C., Gillespie N. and Lockey S. (2023) AI-deploying organizations are key to addressing 'perfect storm' of AI risks, AI and Ethics, 3 (1), 145–153. https://doi.org/10.1007/s43681-022-00163-7

Derrida J. (1972) La dissémination, Seuil.

Eurofound (2023) Work organisation. https://www.eurofound.europa.eu/en/topic/work-organisation

Dvorak J, Kopp T, Kinkel S and Lanza G. (2022) Explainable AI: A key driver for AI adoption, a mistaken concept, or a practically irrelevant feature?. Applications in Medicine and Manufacturing, p. 88.

European Commission (2021) 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, 21.4.2021. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

European Parliament (2023) Artificial intelligence: Threats and opportunities. https://www.europarl.europa.eu/news/en/headlines/society/20200918STO87404/artificial-intelligence-threats-and-opportunities

Ferreras I. (2023) IA : Vers une radicalisation des enjeux de pouvoir dans l'entreprise, Harvard Business Review. https://www.hbrfrance.fr/organisation/intelligence-artificielle-vers-une-radicalisation-des-enjeux-de-pouvoir-dans-lentreprise-60301

Frey CB. and Osborne M. (2023) Generative AI and the future of work: a reappraisal. The Oxford Martin Working Paper Series on the Future of Work. Forthcoming in Brown Journal of World Affairs. pp. 1–12. https://www.oxfordmartin.ox.ac.uk/downloads/academic/2023-FoW-Working-Paper-Generative-AI-and-the-Future-of-Work-A-Reappraisal-combined.pdf

Friedmann G. (1975) Où va le travail humain ?, Gallimard.

García-Peñalvo F.J. and Vázquez-Ingelmo A. (2023) What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI. https://doi.org/10.9781/ijimai.2023.07.006

INRS (2022) L'intelligence artificielle au service de la santé et de la sécurité au travail : enjeux et perspectives à l'horizon 2035. https://www.inrs.fr/media.html?refINRS=PV%2020

Jansen A., Beek D., Cremers A., Neerincx M. and Middelaar J.V. (2018) Emergent risks to workplace safety: Working in the same space as a cobot, TNO Innovation for Life.

Lane M., Williams M., and Broecke S. (2023) The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers, OECD Social, Employment and Migration Working Papers 288, OECD Publishing. https://doi.org/10.1787/ea0a0fe1-en

Latour B. (2015) Face à Gaïa : huit conférences sur le nouveau régime climatique, La Découverte.

Luxton D.D. and Watson E. (2023) Psychological and psychosocial consequences of super disruptive AI: Public health implications and recommendations, in Intersections, Reinforcements, Cascades, Proceedings of the 2023 Stanford Existential Risks Conference, 60–74. https://doi.org/10.25740/mg941vt9619

Masure A. (2019) Résister aux boîtes noires : design et intelligences artificielles, Cités, 80, 31–46. https://doi.org/10.3917/cite.080.0031

Monnin A., Bonnet E. and Landivar D. (2021) Héritage et fermeture, une écologie du démantèlement, Divergences.

OECD (2023) OECD Employment Outlook: Artificial intelligence and the labour market, OECD Publishing. https://doi.org/10.1787/08785bba-en

Pasquale F. (2015) The black box society: The secret algorithms that control money and information, Harvard University Press.

Ponce del Castillo A. and Naranjo D. (2022) Regulating algorithmic management. Policy Brief 2022.08, ETUI. https://www. etui. org/publications/regulating-algorithmicmanagement (09.02. 2023).

Sennet R. (2000) Le travail sans qualités : les conséquences humaines de la flexibilité, Albin Michel.

Shaw J., Rudzicz F., Jamieson T., and Goldfarb A. (2019) Artificial intelligence and the implementation challenge, Journal of Medical Internet Research, 21 (7), https://doi.org/10.2196/13659

Stadin M., Nordin M., Broström A., Magnusson Hanson L.L., Westerlund H. and Fransson E.I. (2016) Information and communication technology demands at work: The association with job strain, effort-reward imbalance and self-rated health in different socio-economic strata, International Archives of Occupational and Environmental Health, 89 (7), 1049–1058. https://doi.org/10.1007/s00420-016-1140-8

Stamate A.N., Sauvé G. and Denis P.L. (2021) The rise of the machines and how they impact workers' psychological health: An empirical study, Human Behavior and Emerging Technologies, 3 (5), 942–955. https://doi.org/10.1002/hbe2.315

Stiegler B. (2015) La société automatique : 1. L'avenir du travail, Fayard.

Tubaro P., Casilli A.A. and Coville M. (2020) The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence, Big Data and Society, 7 (1), 1–12. https://doi.org/10.1177/2053951720919776

Tubaro P., Coville M., Le Ludec C. and Casilli A.A. (2022) Hidden inequalities: The gendered labour of women on micro-tasking platforms, Internet Policy Review, 11 (1), 1–26. https://doi.org/10.14763/2022.1.1623

Verkindt P.-Y. (2020) Ambivalences et promesses dans le champ de la santé et de la sécurité des travailleurs, in Adam P., Tarasewics Y. and Le Friand M. (eds.) Intelligence artificielle, gestion algorithmique du personnel et droit du travail, Dalloz, 199–208.

All links were checked on 29.02.2024.

Cite this chapter: Mandinaud V. and Ponce del Castillo A. (2024) AI systems, risks and working conditions, in Ponce del Castillo (ed.) Artificial intelligence, labour and society, ETUI.

List of contributors

German Bender is Chief Analyst at the Stockholm-based think tank Arena and previously worked for more than a decade at Sweden's two largest trade union confederations, TCO and LO, as a senior research officer and speechwriter. He obtained his PhD at Stockholm School of Economics and has recently been a visiting research fellow at Harvard University's Center for Labor and a Just Economy. His dissertation concerns how the Swedish industrial relations system handles the challenges related to wage bargaining, digital automation and migration.

Nicolas Blanc is National Secretary for Economic Transition for Confédération française de l'encadrement – Confédération générale des cadres (CFE-CGC). Between 2019 and 2023, he held the position of national delegate in charge of digital issues within the Economic Department of CFE-CGC. Nicolas is an expert on AI and the future of work for the Global Partnership on AI, a specialist in the Expert Group on AI Risk and Accountability at the OECD and a member of the Innovation-Friendly Regulations Advisory Group at European expert group level.

Benedetta Brevini is Visiting Professor at the Institute for Public Knowledge, New York University and Associate Professor in the political economy of communications at the University of Sydney. She is also Senior Visiting Fellow at the LSE's Department of Media and Communications. Previously, she worked as a journalist in Milan, New York and London for CNBC, RAI and *The Guardian*. She is the author of several books, including Is *AI Good for the Planet* (Digital Futures, 2022), *Amazon: Understanding a Global Communication Giant* (Routledge, 2020) and *Public Service Broadcasting Online* (Palgrave Macmillan, 2013); and is editor of *Beyond Wikileaks* (Palgrave Macmillan, 2013), *Carbon Capitalism and Communication: Confronting Climate Crisis* (Palgrave Macmillan, 2017) and *Climate Change and the Media* (Peter Lang, 2018). She is currently working on a new volume for Polity entitled *Communication systems, technology and the climate emergency*.

Antonio A. Casilli is full Professor of Sociology at the telecommunications school (Telecom Paris) of Institut Polytechnique de Paris and a researcher at the Interdisciplinary Institute on Innovation (i3), an institute of the Centre National de la Recherche Scientifique (CNRS; French National Centre for Scientific Research). He is also an associate researcher at LACI-IIAC (Critical Interdisciplinary Anthropology Centre) of École des Hautes Études en Sciences Sociales (EHESS; School for Advanced Studies in Social Sciences). Additionally, Antonio is a faculty fellow at the Nexa Centre for Internet and Society of Politecnico di Torino.

Odile Chagny is an economist at Institut de Recherches Économiques et Sociales (IRES; Institute for Economic and Social Research). She initially focused on forecasting, foresight and public policy analysis in several French administrations and research centres. For the previous fifteen years, she was working for trade unions and workers' representatives and, since 2014, at IRES. She is the founder of the Sharers & Workers network. In 2020 she co-authored a book on 'uberisation': *Désubériser, reprendre le controle* (Éditions du Faubourg).

Hamid Ekbia is a university professor and Director of the Autonomous Systems Policy Institute at Syracuse University, where he is affiliated with the Maxwell School of Citizenship and Public Affairs and the School of Information Studies. He is interested in how AI and computing mediate the social, economic and cultural aspects of modern life. He is the (co)author of a number of books including *Artificial Dreams: The Quest for Non-Biological Intelligence* (Cambridge University Press, 2008); *Heteromation and Other Stories of Computing and Capitalism* (MIT Press, 2017); *Big Data is not a Monolith* (MIT Press, 2016) and *Universal Access and its Asymmetries The Untold Story of the Last 200 Years* (MIT Press 2022).

Vassilis Galanos is Research Fellow at the University of Edinburgh's Science, Technology and Innovation Studies Department and the Edinburgh College of Art. Vassilis teaches courses on internet, AI and society and various aspects of technological innovation, while serving as Associate Editor of the journal *Technology Analysis and Strategic Management* (Taylor & Francis). Vassilis researches and publishes on the interplay of expectations and expertise in the historical and current conceptual and regulatory development of artificial intelligence, robotics and internet technologies. Vassilis's further academic interests include cybernetics, media theory, invented religions, oriental and continental philosophy, community-led initiatives, journalism and art.

Natalia Giorgi is a project officer at the European Trade Union Confederation (ETUC). She works on EU and EFTA-funded projects aimed at strengthening the representation and participation of workers in standardisation. Natalia is responsible for digitalisation and AI, for monitoring EU policy and regulatory developments in AI and assessing them in relation to workers' rights and standardisation. She holds a bachelor's degree in political science from the Université de Montréal, a degree in international relations from Université Libre de Bruxelles and a master's degree in public administration, with a major in international management, from Ecole Nationale d'Administration Publique of Quebec.

Sandy J. J. Gould is an academic at the School of Computer Science and Informatics, Cardiff University, in the UK and a member of the School's Human-Centred Computing group. He studies people's interactions with digital technologies, especially measurement, tracking and surveillance in work contexts. This includes research in understanding new forms of work, such as crowdwork, but also in how more traditional forms of work are being changed by technology.

Luciana Guaglianone is Associate Professor in labour law at the Law Faculty, Università di Brescia, a lecturer in trade union law, a labour law official referee in ANVUR and a member of CUG Università di Brescia, Italy. She has undertaken research in various fields of labour law, focusing on an antidiscrimination approach to the area of gender, age and disability. She has been a visiting scholar for several periods of research in Spain and in the UK, and participates in various national and international research groups.

Mario Guglielmetti holds a PhD in public law from Università di Trento and an LLM from the College of Europe, Bruges. He is a legal officer at the European Data Protection Supervisor (EDPS). Prior to that, he worked as seconded national expert at the European Commission (DG Justice and Consumers, Data Protection Unit) and as legal officer at Garante per la protezione dei dati personali (the Italian data protection authority). He has also worked as a lawyer for the international law firm Clifford Chance. His contribution to this volume is written in his personal capacity.

Lukas Hondrich holds an MSc degree in cognitive-affective neuroscience from TU Dresden. He researches the requirements for human-centred AI systems and implements these as an AI engineer. His main research interest lies in participatory methods for AI systems, from formal frameworks to practical implementation.

Vincent Mandinaud is a project manager at Agence Nationale pour l'Amélioration des Conditions de Travail (ANACT), where he currently leads the Agency's thematic priority on digital and ecological transition. Vincent is a sociologist specialised in sociotechnical change, and teaches at the Graduate School of Civil, Environmental and Urban Engineering (ENTPE) and at Université Jean Moulin Lyon 3.

Michele Molè is a PhD candidate at the Faculty of Law of Rijksuniversiteit Groningen, The Netherlands. His research focuses on workplace surveillance, new technologies and worker protection in the European legal framework. Michele holds a master's degree in law from Università degli Studi di Milano. He was a trainee at ETUI in 2022.

Anne Mollen conducts research on automation, algorithms and artificial intelligence at the Institute of Communication Studies at the University of Münster and advises the civil society organisation AlgorithmWatch as a Senior Research Associate. Her research focuses on the sustainability of AI, automation and public opinion formation, algorithm-based discrimination and the use of automated decision-making systems in the workplace.

Helga Nowotny is Professor Emerita in science and technology studies, ETH Zurich, and former President of the European Research Council. She is a member of the board of trustees of Falling Walls Foundation, Berlin; and Vice-President of the Lindau Nobel Laureate Meetings. She has received honorary doctorates from, among others, the University of Oxford and the Weizmann Institute of Science in Israel.

Frank Pasquale is Professor of Law at Cornell Tech and Cornell Law School. He is an expert in the law of artificial intelligence, algorithms and machine learning. His books include *The Black Box Society: the Secret Algorithms That Control Money and Information* (Harvard University Press, 2015) and *New Laws of Robotics* (Harvard University Press, 2020), while he co-edited *The Oxford Handbook on the Ethics of Artificial Intelligence* (Oxford University Press, 2020) and *Transparent Data Mining for Big and Small Data* (Springer-Verlag, 2017). He is an Affiliate Fellow at Yale University's Information Society Project and a member of the American Law Institute.

Aída Ponce Del Castillo is a lawyer by training. She holds a master's degree in bioethics and has obtained her European doctorate in law. At the Foresight Unit of the ETUI, her research focuses on the cross-boundary field between science and emerging technologies, especially with regard to ethical, social and legal issues, with a focus on AI. Additionally, she is in charge of conducting foresight projects. She is a member of the committee for the National Convergence Plan for the Development of AI, in Belgium. At the OECD she is a member of the working parties on biotechnology, nanotechnology and converging technologies, as well as on Al governance. She was previously Head of the ETUI Health and Safety Unit and coordinator of the Workers' Interest Group at the Advisory Committee of Safety and Health to the European Commission.

Frank Pot is Emeritus Professor of Social Innovation of Work and Employment, Radboud Universiteit Nijmegen, and honorary advisor of the European Workplace Innovation Network. He is former director of TNO Work and Employment and a former professor by special appointment in work and technology, Universiteit Leiden.

María Luz Rodríguez Fernández is full Professor in labour and social security law at Universidad Castilla-La Mancha, Spain. She was mediator and head of the legal department of Servicio Interconfederal de Mediación y Arbitraje (SIMA; Interconfederal Mediation and Arbitration Service) as well as Secretary of State for Employment at the Spanish Ministry of Labour and Immigration in 2011. She was a senior specialist in labour market institutions at the International Labour Organization and drafted the report on 'Decent Work in the Platform Economy' for the Meeting of Experts held in 2022.

Teresa Rodríguez de las Heras Ballell is an associate professor of commercial law at Universidad Carlos III. de Madrid, Spain. She is a delegate of Spain at UNCITRAL for WG VI and WG IV, and an expert for UNCITRAL and UNIDROIT on digital economy projects. She is an arbitrator at Las Corte de Arbitraje de Madrid (Madrid Court of Arbitration) and Corte Española de Arbitraje (Spanish Court of Arbitration). She is also member of three EU Commission expert groups: liability and new technologies; online platform economy; and on B2B data sharing and cloud computing. She is a member of the Executive Committee and Council of the European Law Institute (ELI), and the author of the ELI guiding principles on automated decision-making in Europe.

James Stewart is a lecturer in science, technology and innovation studies at the University of Edinburgh. He works on the personal, industrial, political and social dimensions of information and communications technology, the internet and AI, spanning the domestication of computing, gender and technology, digital inclusion, studies of the internet of things, labour, telecommunications innovation, 'smart' cities and co-design. He currently works on government use of online influence infrastructures and the governance of generative AI.

Inga Ulnicane is a research fellow at the University of Birmingham, UK. She is an interdisciplinary social scientist working at the intersection of policy research, political science and social studies in science and technology. She has published extensively on topics such as the politics and policy of artificial intelligence, the governance of emerging technologies, grand societal challenges and responsible research and innovation. In addition to academic research, she has prepared commissioned reports for the European Parliament and the European Commission.

European
Trade Union Institute
Bd du Jardin Botanique, 20
1000 Brussels
Belgium
etui@etui.org
www.etui.org

Artificial intelligence, labour and society

Edited by Aída Ponce Del Castillo

The rapid expansion of artificial intelligence is unparalleled, establishing it as a ubiquitous element in workplaces and our daily lives. The era when AI was exclusively associated with robots and intricate algorithms for the technically proficient is over. This marks a significant paradigm shift, with profound implications and changes regarding the world we live in.

This book proposes an analysis of this transformation, and does so by bringing together the reflections of high-level academics and research activists from across the world. It adopts a multidisciplinary approach, incorporates a diversity of geographical and cultural points of view and focuses on the deep and often invisible implications of AI for the labour market and society as a whole. Thematically speaking, it addresses AI's legal, societal, global, environmental, technological and labour aspects.

The contributing authors show how contemporary AI is gradually reshaping society. They also highlight the need to understand the dynamics of technological convergence and remind us of the critical role played by humans, who remain present at every stage of AI's lifecycle and value chain. Prevention and precaution are portrayed as crucial components of the critical thinking approach we need to adopt towards AI.

The book also examines the power dynamics of technological evolution and governance, explores AI's relationship with the environment and tackles the crucial issue of AI's impact on the labour market, working conditions and labour standards. Its fundamental goal is to prompt readers to transcend the narrow disciplinary perspectives associated with the silos within which we tend to work, introducing a range of novel and diverse academic viewpoints on AI and its challenges.

D/2024/10.574/10
ISBN 978-2-87452-707-4