AI for Fair Work Report

November 2022 - GPAI Tokyo Summit

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was developed by Experts and Specialists involved in the Global Partnership on Artificial Intelligence's project on the AI for Fair Work. The report reflects the personal opinions of the GPAI Experts and Specialists involved and does not necessarily reflect the views of the Experts' organisations, GPAI, or GPAI Members. GPAI is a separate entity from the OECD and accordingly, the opinions expressed and arguments employed therein do not reflect the views of the OECD or its Members.

Acknowledgements

This report was developed in the context of the AI for Fair Work project, with the steering of the project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Future of Work Working Group. The GPAI Future of Work Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Mark Graham[*], Oxford Internet Institute; UK
Anne-Marie Imafidon[*], Stemettes; Institute for the Future of Work; UK

Project Advisory Group:
Janine Berg[*], International Labour Organization; Switzerland
Nicolas Blanc[*], CFE CGC; France
Stijn Broecke[**]; OECD
Callum Cant[‡], Oxford Internet Institute; UK
Matthew Cole[†], Oxford Internet Institute; UK
Jenny Grensman[*], Sveriges ingenjörer/The Swedish Association of Graduate Engineers; Sweden
Michela Milano[*], Centro Interdipartimentale Alma Mater Research Institute for Human-Centered Artificial Intelligence; The University of Bologna; Italy
Johan Moesgaard Andersen[*], Danish Metal-Workers Union; Denmark
Oliver Suchy[*], Department on Digital Workplace and Workplace Reporting; The German Trade Union Confederation; Germany
Lucia Velasco[*], School of Transnational Governance. European University Institute (EUI); Spain

The report was written by: **Callum Cant**[‡], Oxford Internet Institute; **Mark Graham**[*], Oxford Internet Institute; **Funda Ustek Spilda**[‡], Oxford Internet Institute; **Matthew Cole**[†], Oxford Internet Institute and **David Brand**[‡], Oxford Internet Institute.

GPAI is also grateful to the following individuals for reviewing: **Janine Berg**[*], International Labour Organization and **Saiph Savage**[*], Tecnológico de Monterrey; Berkman Klein Center for Internet & Society, Harvard University. Finally, GPAI would like to acknowledge the efforts of colleagues at the Centre of Expertise (CofE) of Paris. We are particularly grateful for the support of Laetitia Cuignet, Isabelle Herlin, Catherine Pacherie-Simeral and Edouard Havis from the CofE of Paris, and for the dedication of the Working Group Co-Chairs Uday B. Desai and Matthias Peissner.

* Expert of GPAI's Future of Work Working Group
** Observer at GPAI's Future of Work Working Group
† Invited Specialist
‡ Contracted parties by the CofE of Paris

Citation

GPAI 2022. AI for Fair Work: AI for Fair Work Report, November 2022, Global Partnership on AI.

Executive summary

Artificial Intelligence (AI) is transforming the way we live. This change is neither good nor bad; nor is it neutral.[1] Instead, the ongoing development of AI systems presents us with both opportunities and threats that will emerge in different ways in different contexts. The collective challenge ahead of us is to ensure that our adoption of this new technology maximises the upsides and minimises the downsides. There has been much debate about what ethical principles should guide our collective response to AI in order to achieve this goal. However, the existing academic and policy discourses have largely failed to address an area in which AI is already having a transformative impact: the workplace. Deployments of AI systems in the labour process are affecting an ever-expanding number of people, but most extant ethical frameworks are inadequate to address how we should work with and alongside AI systems. This report presents a set of ten Fair Work principles for AI that address this specific issue, developed through extensive tripartite consultation. They are:

1. **Guarantee fair work:** Ongoing changes in work caused by the introduction of AI systems have the potential to disrupt the labour market, but internationally agreed minimum rights and standards remain a precondition of fair AI.

2. **Build fair production networks**: AI system development and deployment relies on global networks of human labour, hardware production, and infrastructure. Organisations seeking to implement fair AI in the workplace must therefore look beyond the immediate production process to the networks of production that enabled it and use their procurement power to achieve fairness across the network.

3. **Promote explainability**: Workers have a right to understand how the use of AI impacts their work and working conditions. Organisations must respect this right and provide detailed, understandable resources to allow workers to exercise it.

4. **Strive for equity**: AI systems have been found to reproduce and scale up patterns of social discrimination. The costs associated with embedding negative consequences for marginalised groups into workplace technology are extremely high. As a result, AI systems must be (re)designed, built, and deployed in a way that actively seeks to eliminate sources of discrimination. Processes such as audits and impact assessments should be integrated into the AI system lifecycle to allow for ongoing scrutiny.

5. **Make fair decisions**: the automation of decision making can lead to reductions in accountability and fairness. But building in human oversight into a decision making loop doesn't solve this problem. Instead, the subjects of those decisions need to be empowered to challenge them, and a renewed emphasis should be placed on the liability of those stakeholders who direct the development and deployment of AI systems in the workplace.

6. **Use data fairly**: the collection of large quantities of data and the concentration of its ownership may exacerbate risks for individuals and social groups, especially when shared with third parties. Limits must therefore be put on collection (i.e. data minimisation) and processes must be instituted for subjects to access and protect their data in a comprehensive and explainable format. Organisations should provide comprehensive guidelines for individuals to understand data ownership, data usage and any potential risks that result, so that they are able to question, contest, and when necessary, reject, decisions made about them.

7. **Enhance safety**: advances in algorithmic management have increased the risks of work intensification and surveillance. In this context, the right to healthy, safe working environments must be protected. Potential improvements in safety should be capitalised upon, but deployment must

[1] Melvin Kranzberg, "Technology and History: 'Kranzberg's Laws,'" *Technology and Culture* 27, no. 3 (July 1986): 544, https://doi.org/10.2307/3105385.

take place in a way which reflects the different understandings of stakeholder groups about the trade-offs involved.

8. **Create future-proof jobs:** the introduction of AI systems to workplaces can cause specific risks such as job destruction and deskilling. These risks can be reduced by treating the introduction of AI as an opportunity for workers and organisations to engage in a participatory and evolutionary redesign of work which uses the rewards of AI to increase job quality.

9. **Avoid inappropriate deployment**: organisations should proactively test AI systems to a high standard in order to avoid harms in advance, rather than iterating to address them post-deployment.

10. **Advance collective worker voice**: the risks and rewards of AI systems are understood differently by different stakeholder groups. These divergences should be proactively negotiated, rather than suppressed. Pursuing AI system implementation in a multi-stakeholder environment requires a mechanism to turn ethical principles into ethical practice through democratic participation by workers. Collective bargaining between workers and management is best suited to play this role.

Introduction

Rapid technological advances can force social change and influence the social relations that structure society. The latest so-called 'AI Spring' is having such an effect. Analysts of AI development refer to AI springs and winters to characterise the cyclical up and downswings in funding and research momentum.[2] Since 2012, the use of large datasets and advances in computing power have rapidly accelerated the performance of AI systems. The range of AI use-cases have also grown, and in combination these changes have led to a seasonal shift. The blossoming of a new AI spring over the course of the last decade has been accompanied by an expansion in the discussion of AI ethics, as the new capacities generated during this burst of development have interacted with the complex reality of human social life. As a result, a broad range of stakeholders have begun to ask not only what we *can* do with AI, but what we *should* do with AI, and *how* we should do it.

The OECD's *Recommendation on Artificial Intelligence* was adopted in 2019 as the first intergovernmental standard on AI. The GPAI was founded to advance the values laid out in this standard through multistakeholder research and applied activities. This report has been produced by the GPAI Future of Work working group to build on the OECD recommendation in the specific subdomain of work. The rationale behind this approach derives from the unevenness of much of the existing discussion of the risks and opportunities associated with AI.

Algorithm Watch maintains a crowd-sourced inventory of AI ethics which currently features 173 distinct sets of principles produced by civil society, governments, the private sector and other actors. The majority of the guidelines are non-binding recommendations with no associated capacity for enforcement. Many share common themes. In a study of 84 sets of such guidelines, researchers identified eleven overarching principles that featured prominently.[3] In order of popularity, these were: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity. Five of those values (transparency, justice and fairness, non-maleficence, responsibility and privacy) appeared in the majority of the standards studied. Other longitudinal reviews have identified six key ethical concerns: inconclusive evidence, inscrutable evidence, misguided evidence, unfair outcomes, transformative effects, and traceability, and showed

[2] Jacques Bughin and James Manyika, "The Coming AI Spring," Project Syndicate, October 14, 2019, https://www.project-syndicate.org/commentary/artificial-intelligence-spring-is-coming-by-james-manyika-and-jacques-bughin-2019-10.
[3] Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (September 2019): 389–99, https://doi.org/10.1038/s42256-019-0088-2.

how they have remained consistent over time.[4] However, this convergence has not necessarily been reflective of the full range of stakeholder opinion.

Unsurprisingly, the process of principle formation seems to have strongly reflected existing forms of social power. In the context of the workplace, this means that the interests of managers, owners and shareholders have tended to come out on top. Rather than a truly diverse discussion about ethics in the era of AI, the debate has instead placed the views of already-hegemonic actors front and centre. Research has demonstrated how this has led to the emergence of an unrepresentative consensus amongst powerful actors that excludes those most impacted by the topics under discussion.[5] Consequently, much of the debate remains superficial, self-congratulatory and out of touch with the everyday implications of AI for the people exposed to the bulk of risk.

One of the major impacts of this imbalance has been the neglect of the questions that matter most to the people on the front lines of technological change. AI systems are already being implemented in workplace contexts across the globe, often transforming the way work is being done. Labour processes like last mile delivery, warehousing, and many others have undergone massive reorganisations, and workplace functions as diverse as work allocation, productivity management, and recruitment are increasingly automated. But despite this rapid proliferation of uses in a high-risk environment, discussions of what ethical principles should guide AI systems in the workplace remain both remarkably limited and imbalanced. The deployment of AI in this arena has rarely been paid sufficient attention, and when it has workers and their organisations have largely been left out.

The relationship between regulations and principles

AI systems are increasingly subject to specific policies aiming to codify ethical principles in regulatory structures. As of 2021 the OECD identified 700 AI policy initiatives from 60 countries, territories and the European Union.[6] Landmark legislation is being developed in both the EU and US, with the White House Office of Science and Technology consulting on an AI bill of rights[7] and the EU Commission proposing an Artificial Intelligence act, and draft directives on both platform work and AI liability.[8] These initiatives are themselves the subject of debate over principles and content[9] but their development marks a significant transition in the overall approach to AI system regulation.

Regulation is finally begin developed, but does this make discussions of ethics and politics redundant? No, as this would fail to account for how the debate around the ethics and politics of AI has been discursively framed. The introduction of legislation is a response to the difficulty of translating non-

[4] Brent Mittelstadt et al., "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* 3, no. 2 (December 2016): 205395171667967, https://doi.org/10.1177/2053951716679679; Andreas Tsamados et al., "The Ethics of Algorithms: Key Problems and Solutions," *AI & SOCIETY*, February 20, 2021, https://doi.org/10.1007/s00146-021-01154-8.
[5] Matthew Cole et al., "Politics by Automatic Means? A Critique of Artificial Intelligence Ethics at Work," *Frontiers in Artificial Intelligence* 5 (July 15, 2022): 869114, https://doi.org/10.3389/frai.2022.869114; Merve Hickok, "Lessons Learned from AI Ethics Principles for Future Actions," *AI and Ethics* 1, no. 1 (February 2021): 41–47, https://doi.org/10.1007/s43681-020-00008-1.
[6] OECD, "OECD.AI Database of National AI Policies," 2021, https://oecd.ai/en/.
[7] Eric Lander and Alondra Nelson, "Americans Need a Bill of Rights for an AI-Powered World," *Wired*, 2021, https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/.
[8] European Commission, "Proposal for a Directive of the European Parliment and of the Council on Improving Working Conditions in Platform Work" (Brussels: European Comission, September 12, 2021), https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2021)762&lang=en; European Commission, "Proposal for a Directive on Adapting Non Contractual Civil Liability Rules to Artificial Intelligence," September 28, 2022, https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2022)496&lang=en.
[9] Thomas Metzinger, "Ethics Washing Made in Europe," *Der Tagesspiegel*, April 8, 2019, sec. Politik, https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html; "An EU Artificial Intelligence Act for Fundamental Rights: A Civil Society Statement," European Digital Rights (EDRi), 2021, https://edri.org/our-work/civil-society-calls-on-the-eu-to-put-fundamental-rights-first-in-the-ai-act/.

binding principles into ethical practice, rather than a rejection of that process of principle formation in its entirety. Indeed, the debate over principles has often rehearsed the arguments that have later shaped legislation. Thus, ethical debates can indeed have a powerful indirect effect.[10] By shaping the deeper values that lie behind the latest legislative developments, statements of principle have played a formative role in generating forms of 'harder' regulation. If we are critical of how that process of principle formation happened, then this interpretation has an additional consequence: we need to correct the mistakes made during this formative process so that these corrections can eventually feed through into legislation. In addition, a principle-based approach to questions of AI system regulation is key for adaptability to both local contexts and new developments.[11] Whilst specificity is an advantage in some circumstances, it is a weakness in others – and in a field as fast moving as AI, it is important we also retain more general frameworks that can be applied to the latest developments in a range of contexts.

So, these principles have been developed with two goals in mind: first, to create a globally applicable set of guidelines that can assist in translating the OECD Recommendations into the specific subdomain of work; and second, to correct the lop-sidedness of previous principle-based discussions in a way that feeds through into policy and the concrete application of AI systems in the workplace. In the creation of these principles, we have sought to correct the underrepresentation of worker voice and advance a truly multi-stakeholder perspective that foregrounds immediate issues of ethical concern. In addition, we have sought to design the principles with impact in mind, by thinking through how we can adopt with proven non-statutory techniques for generating impact to bring these principles to bear.

Concept of fair work

The title of these principles refers to 'fair work'. But what exactly does that mean? 'Decent work' is a widely used concept with extensive history in the collective practices of the ILO and the UN. The concept is embedded in the sustainable development goals, and integrated into international policy on many levels. Decent work is a fundamental baseline that all workers, everywhere in the world, should be guaranteed. Our concept of 'fair work,' however, is distinct from this.

While we view also fair work as a fundamental baseline which should be universally guaranteed, the standards that define this concept are better attuned to the evolving challenges emanating from the application of AI to the world of work. The notion of fair work has been developed through extensive engagement with the reality of emerging forms of platform work across 39 countries through the Fairwork project that was founded at the University of Oxford. We view fairness as an evolving quality, which iterates over time as society develops. In the report that follows, we treat fairness as an orientation. The benchmarks included in the principles below do not lay out the final coordinates of fair work, but instead indicate the necessary direction of travel if fair work is to be eventually achieved. The benchmarks should be seen as minimum thresholds of fairness that will have to progress over time in response to the challenges and opportunities presented by technological development and its consequent social effects. The benchmarks we provide below are orientated towards a global context and a range of AI actors: both those who deploy AI systems and those who design them.

Connecting all the principles that follow is a fundamental optimism. A set of principles can be understood as a mechanism designed to avoid and mitigate risks, but it can also be used a way to identify opportunities. There is an emancipatory potential in technology that should be embraced with the goal of not only maintaining the wellbeing of actors involved in the workplace but expanding it and going beyond the status quo. Given the immense challenges of the 21st century, primary amongst which is the existential threat of anthropogenic climate change, making the most of these opportunities will be vitally important.

[10] Institute for the Future of Work, "Mind the Gap: The Final Report of the Equality Task Force," 2020, https://www.ifow.org/publications/mind-the-gap-the-final-report-of-the-equality-task-force.
[11] Alessandro Mantelero, "Artificial Intelligence and Data Protection: Challenges and Possible Remedies" (Strasbourg: Consultative Committee of The Convention for The Protection of Individuals with Regard to Automatic Processing of Personal Data, 2019), https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6.

The need for red-lines

There is an interesting pattern in some discussions of AI ethics: a problem is raised, a potential solution is identified, but the solution challenges the scope of what is technically possible given the resources available to the system developer or user. In this situation, it is all too common for the developer or user to suggest that this solution is non-viable, and therefore the identified problem is just an insoluble feature of technology that has to be endured.

This raises a very important question: When technologies that are being built clash with our ethical principles, how will we make decisions? Ultimately, when ethics are at odds with practice, one or the other will have to be prioritised. The view embedded in this report is that ethics should always come first. Our principles must determine the shape of our technology, and not vice versa.

In the place of a principled approach to fairness, some AI actors suggest we adopt an approach to ethics based on risk/reward evaluations. If an AI system delivers sufficient benefits, this logic goes, then certain problems can be overlooked or mitigated. These arguments tend to be made by those who see little of the risk and reap most of the reward – a conflict of interest, to say the least. By contrast, the principles we have developed exist to advance the process of identifying an absolute minimum position of fairness which should not be breached, regardless of the level of reward. Once these basic minimums have been met, then multiple stakeholders should be engaged in the process of collectively determining what decisions should be made on the balance of risk and reward above and beyond this basic level.

Consequently, if at any point in the AI system lifecycle it becomes clear that the below principles are not being or cannot be met, then that adherence to these principles dictates that this AI system is not fit for deployment. It is not sufficient to define AI as high risk and continue regardless with a mitigation-based approach. Ultimately, 'red lines' which cannot be crossed are necessary. This will inevitably mean that some applications of AI systems to the workplace are ruled out. Certain forms of invasive surveillance, for instance, seem impossible to reconcile with the principles stated below. But in these circumstances, we must accept that these functions are fundamentally contradictory to our ethics. They are not just high-risk use cases which need to be handled extremely cautiously – they should, in fact, be prohibited. This approach to deployment has been backed by members of European civil society.[12] When it is impossible to deploy an AI system without crossing clear red lines, then that AI system should not be deployed at all.

Finally, the authors wish to clarify that under no circumstances should these principles replace or supersede any relevant legal frameworks. Examples of voluntary compliance with these principles, nor other non-statutory frameworks, do not negate the wider need for enforceable AI system regulation that aims to further the goal of fair work. We have attempted to explain in detail the role we see these principles playing, and it is emphatically not that of a replacement for robust legislation.

Method

This report was produced by a research team based at the Fairwork project at the Oxford Internet Institute, working in collaboration with the Future of Work working group at the Global Partnership in Artificial Intelligence (GPAI). It has been wholly funded by the GPAI.

We began the project by conducting a literature review on the question of AI ethics in the workplace which covered 90 items, ranging in type from AI ethics principles, to reports from civil society organisations, and research articles. Through this review, we identified nine recurring themes that seemed of significant relevance to the workplace context: explainability and accountability; bias and discrimination; decision making; the use and governance of data; impacts on job quality; AI production networks; occupational safety and health; technological unemployment; and collective worker voice. We also identified how all these themes interacted with the particular power asymmetries that result from

[12] European Digital Rights, "Civil Society Calls for AI Red Lines in the European Union's Artificial Intelligence Proposal," EDRi, European Digital Rights (EDRi), 2021, https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal/.

the specific modes of ownership and control that structure the global economy, and the way in which this led to worker perspectives being widely neglected and deprioritised in multi-stakeholder initiatives. We then continued to conduct a further sub-review of the relevant literatures for each of these themes, drawing on a range of cross-disciplinary materials from sociology to law, philosophy and computer science. When the review was completed, it covered 311 items. On the basis of this initial work we developed a first set of draft principles, which were then further developed through consultation with the GPAI Future of Work working group. This then resulted in a second version, which was the basis on which we began external consultation.

We completed two rounds of engagement with global stakeholders to understand their perceptions of the draft principles and engage with them on questions we felt we had yet to adequately address. Over the course of the first round, we completed 21 in-depth interviews with stakeholders. We split these interviewees into 4 categories: trade unions and worker representatives, governments and quasi-governmental organisations, academics and experts, and private sector representatives. We intentionally designed our sample to correct for the previous failure to include worker perspectives in discussion of AI ethics in the workplace. The final sample was made up of 8 trade union and worker representatives, 4 academics and expert representatives, 5 private sector representatives, and 3 government and quasi-governmental organisation representatives. The stakeholders interviewed represented a range of organisations including major technology firms and AI developers; global labour platforms; international trade union confederations; and information regulators. Each interviewee was sent a copy of the second version of the principles before taking part in a one-hour, semi-structured interview which reviewed the principles and discussed issues of specific salience to the participant.

Following on from this initial interview process, we also held one focus group with 4 participants from a range of stakeholder groups to attempt to understand the convergences and divergences in their perspectives. During this focus group we discussed two case studies of AI implementation in the workplace and one case study of a work-orientated data broker to understand their perceptions of risk and how the principles could engage with these circumstances to produce positive change. We then integrated insights from both the interviews and the focus group into a third version of the principles.

The second round of qualitative engagement with stakeholders consisted of a survey that presented participants with sections of this third version of the principles. We asked for their specific comments and responses regarding the rationale of each principles and the measurable benchmarks associated with them. We also included a ranking exercise at the start of the survey to evaluate the relative priority stakeholders attributed to each principle. By making this exercise zero-sum (ranking one principle more highly necessarily implied ranking another less so) we hoped to minimise acquiescence bias.

This survey was distributed broadly via email to key stakeholders and via both the GPAI and Fairwork networks. Between the 18th of August and the 20th of September, we received 117 responses. We conducted a thematic analysis on the resulting data.[13] This analysis resulted in 71 codes divided into 12 themes. We then designed a series of edits to the principle text and the wider report that responded to these themes before finalising the report.

Impact

As discussed above, there have been widespread concerns about the ways in which statements of ethical principle get translated into practice. As a result, the AI for Fair Work project has been planned with an extended impact phase in mind, during which the report authors will conduct further work to ensure that the principles go on to have a tangible impact on AI system deployment in the workplace. This phase will be implemented by Fairwork, a project partly based at the Oxford Internet Institute and led by members of the research team. The phase will continue into 2023, and consist of two key work

[13] Victoria Clarke and Virginia Braun, “Thematic Analysis,” *The Journal of Positive Psychology* 12, no. 3 (May 4, 2017): 297–98, https://doi.org/10.1080/17439760.2016.1262613; Gareth Terry et al., “Thematic Analysis,” in *The SAGE Handbook of Qualitative Research in Psychology*, ed. Carla Willig and Wendy Stainton Rogers (1 Oliver’s Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd, 2017), 17–36, https://doi.org/10.4135/9781526405555.n2.

packages: first, workplace case studies of AI system deployment; and second, using the Fairwork methodology to score a variety of organisations deploying AI in their workplaces against a modified version of the benchmarks laid out in the principles below.

Fairwork is a global research project currently focused on fair working conditions in the platform economy. Since being founded in 2018 the network has expanded significantly and is now made up of research teams spread across 39 countries. Our methodology consists of three parts: first, conducting multinational monitoring of platforms with a consistent methodology; second, evaluating these platforms against a set of standards for fairness; and third, implementing a scoring approach that leverages platforms to proactively make positive, pro-worker changes and thereby demonstrating the feasibility of fair work. The approach has proven highly successful, with the project so far incentivising 33 platforms to make over one hundred pro-worker changes to policies and practices. Given the rapid expansion of Fairwork into new countries, this number should only increase as teams conduct more rounds of scoring. This experience has allowed us to experiment with different ways of producing impact. Now, Fairwork is integrating a workstream on AI, which will aim to reproduce some of these impacts in relation to this new field. The Fairwork impact model can be segmented by intended audience and cross applied to the impact phase of the AI for Fair Work project to illustrate our plans more clearly for the next year.

Policy Makers

The implementation case studies will offer policy makers an in-depth understanding of the issues raised by AI deployment in a real-world context, and what tensions exist between the Fairwork principles and existing practices. They will also enable the project to identify possible routes for regulatory action at all stages in the development of regulatory systems. The scoring process will allow policy makers to monitor how some key players match up to the standards of fairness detailed below, and where points of convergence and divergence with these principles exist.

Workers and Unions

These principles were drafted in part to address the significant underrepresentation of worker voices and issues of concern to workers in the AI ethics debate. As such, the impact phase offers workers and unions significant opportunities to share their perspectives. The Fairwork methodology engages both workers and trade unions in the process of empirical research on the labour process ahead of scoring, and facilitates the provision of evidence about risks and benefits that can otherwise go ignored. With the benchmarks written into the principles and investigated by the case studies, we aim for these to be useful reference points for workers and unions. We hope they will be used as tools that facilitate an understanding of how workplace risks can be minimised and benefits maximised. We anticipate that trade union research departments and similar bodies may also find the information in scoring useful in terms of monitoring ongoing trends in AI system deployment.

Private Sector

Given the complexity of AI systems, a common barrier to the implementation of ethical principles arises around the question of technical feasibility. The goal of our case study approach is to develop research which addresses this question directly, and demonstrates strategies adopted by stakeholders to put the principles into practice in real-world environments. For organisations being scored, the impact phase will apply public scrutiny to their use of AI and potentially generate an additional incentive for internal change. The Fairwork methodology also entails extensive engagement with the subjects of scoring, during which time collaborative discussions are possible regarding how an organisation can take productive steps to meet benchmarks and achieve the standards of fair work.

Our aspiration is for this impact phase to mark the start of Fairwork's long-term engagement with the issue of AI and fairness, and for the project to continue to follow-up this impact phase with further to support the implementation of the principles in the future. Potential collaborators interested in pursuing this goal are invited to contact the project directly.

The Principles

1. Guarantee fair work
2. Build and maintain fair production networks:
3. Promote explainability
4. Strive for equity
5. Make fair decisions
6. Use data fairly
7. Enhance safety
8. Create future-proof jobs
9. Avoid inappropriate deployment
10. Advance collective worker voice

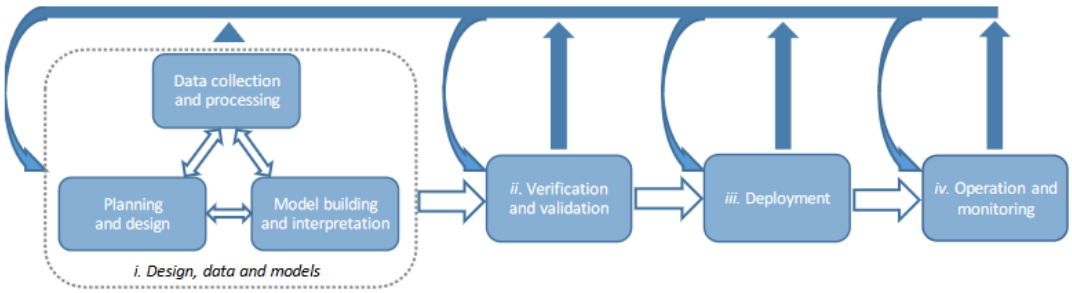
Definitions

Artificial Intelligence (AI) is a set of technologies that "seeks to make computers do the sorts of things that minds can do."[14] Different kinds of AI can perform an ever-increasing range of tasks, from image recognition to language processing, using a variety of computing methods.

An **AI System** is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.[15]

An **AI System Lifecycle** is a series of phases: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.[16]

Figure 1.0 The AI System Lifecycle[17]



Fair Work is work that is optimised to produce individual and collective benefits for both workers and society as a whole. These principles are orientated towards that objective. However, the benchmarks associated with each principle only establish the necessary direction of travel, rather than laying out the final coordinates of the destination.

[14] Margaret A. Boden, *AI: Its Nature and Future*, First edition (Oxford, United Kingdom: Oxford University Press, 2016), 6.
[15] OECD, "Recommendation of the Council on Artificial Intelligence," 2019.
[16] OECD.
[17] OECD, *Artificial Intelligence in Society* (OECD, 2019), https://doi.org/10.1787/eedfee77-en.

Organisations are actors who directly or indirectly pay workers to engage in a labour process that the organisation primarily directs and controls.

Workers are actors who are directly or indirectly paid by an organisation to engage in a labour process that is primarily directed and controlled by that organisation.

Trade unions are independent bodies created by workers for the purpose of representing their collective interests. In some rare circumstances, the role of a trade union can be fulfilled by another form of collective worker organisation without the legal status of a trade union, so long as this organisation is fully independent of the organisation.

These definitions are independent of, and do not imply legal statuses. The nature of employment, self-employment, subcontracting and other relationships show significant global differences that they require the use of broad definitions in these principles.

1. Guarantee fair work

The concept of fair work as developed by our project applies to more than just the deployment of AI systems in the workplace is just one of its many components. It would be nonsensical to only consider the fairness of this one component in isolation from others, such as wages and conditions. It is important that the following principles are understood on this basis: the fair use of AI in the workplace cannot be achieved without also making progress on a broader fairness agenda beyond the bounds of AI systems. As a result, this principle holds that a wider set of more general standards on fair work must be met alongside the AI-specific benchmarks that follow before progress can be made towards fairness.

This runs counter to some of the existing trends of AI deployment, particularly in the case of 'geographically-tethered platform work.'[18] Digital labour platforms have often used the introduction of new forms of technology as an opportunity to attempt to fundamentally alter the relationship between workers and organisations. These changes, such as the frequently illegitimate avoidance of the employment relationship, have frequently resulted in workers being deprived of established rights. This flies in the face of one simple reality: in the absence of deeper changes in social structure, the application of AI to the workplace does not *fundamentally* alter the relationship between the two parties, despite the important circumstantial changes associated with its application.[19] The technological 'disruption' caused by AI in no way justifies walking back long-established rights, any more than the introduction of the telephone or the computer did. This trend, which risks using technological change to regressively redefine workers' rights, is directly contradictory to developing AI for Fair Work.

Organisations must demonstrate a broad commitment to fair work beyond the specific issues of AI to meet this principle. The benchmarks below establish a baseline against which organisations can be measured on issues like freedom of association, wages and conditions, and health and safety. But these are not in themselves exhaustive. The benchmarks also include several standards developed by the ILO with explicit reference to other qualifiers of job quality, such as 'decent,' 'good,' and 'fulfilling' work. We treat these concepts as distinct but non-contradictory with the central concept of fair work deployed by these principles. (For more detail on this relationship see the dedicated section in the introduction above.).

Benchmarks

- Organisations should commit to meeting the principles laid out in the ILO *Declaration of Fundamental Principles and Rights at Work*, namely: freedom of association and the effective recognition of the right to collective bargaining; the elimination of all forms of forced or compulsory labour; the effective abolition of child labour; the elimination of discrimination in respect of employment and occupation; and a safe and healthy working environment.

[18] Jamie Woodcock and Mark Graham, *The Gig Economy: A Critical Introduction* (Cambridge; Medford, MA: Polity, 2020).
[19] Cole et al., "Politics by Automatic Means?"

- Organisations should additionally commit to supporting the goals of the ILO's decent work agenda, in particular the four areas of the agenda highlighted in the ILO's *Centenary Declaration on the Future of Work*: respect for workers' fundamental rights; an adequate minimum wage; maximum limits on working time; and safety and health at work.
- Organisations should additionally show concrete evidence of their commitment to the UN's Sustainable Development goals – particularly decent work as embedded in SDG 8. Specifically, organisations should follow the example of participants in the UN Global Compact and commit to paying all workers a living wage.[20]
- The organisation should proactively comply with the spirit and letter of all applicable national and international statutory labour standards.

2. Build fair production networks:

AI system development is thoroughly dependent on human labour, material infrastructure and computer hardware. Like any other technology, global supply chains and production networks feed into design, testing, dataset labelling, and more.[21] Currently many discussions of the negative impacts of AI on workers focus only on the operation of AI in the workplace in which it is finally deployed. This is understandable, because these workers are usually the only audience which directly engaged with AI as a workforce. But the lifecycle of an AI system has impacts on a much wider workforce, who interact with that AI system through the process of production. At each of the four stages of the AI lifecycle ('design, data and models'; 'verification and validation'; 'deployment'; and 'operation and monitoring'), workers beyond the deployment site may undertake work tasks which allow for the AI system to function. Indeed, there is a strong argument that we should consider the workers who enable AI system lifecycles themselves as worker's impacted by that AI system. From mineral extraction to data labelling and Graphics Processing Unit Manufacture, the working conditions in these supply chains are shaped by the way in which AI systems are produced.

Imagine a fast-food worker, who produces and serves food to order based on voice orders given to a customer-facing Natural Language Processing AI. One of the core requirements for this kind of system would be that it is able to identify the names of specific products from a range of voices, so significant attention would be paid to training it with a wide range of labelled voice data. The labelling of this data would likely be done by a cloud worker via an online labour platform. Whilst the fast-food worker is the only one of these two workers who engages with the AI system directly, both have their working conditions shaped by it. These principles proceed on the basis that the work experience of both workers should be considered when determining if any one AI system is being implemented fairly.

The need for action on fairness in AI system production networks is clear. There is significant evidence that workers in parts of these networks already face extreme risks. 2021 research by Fairwork has shown that none of the seventeen global cloudwork platforms studied could provide evidence that they met five basic principles of fairness.[22] Other research on cloudwork in the global south has found evidence that such work has net-negative impacts on the communities cloudworkers are situated within. .[23] An additional issue of concern that comes to the fore when considering the production network is sustainability. Given the existential threat posed by anthropogenic climate change, the varied uses of electricity, land, water and resources in the production of AI systems should be subject to additional scrutiny. AI production networks must pursue just and fair outcomes not only with regards to the workers directly enmeshed within them, but also with regards to the societies that enable them.

There are positive examples of action which address concerns like this. In the textile industry, the ACT agreement between IndustriALL and 32 global retailers commits all parties to achieve living wages in

[20] UN Global Compact, "Achieving the Living Wage Ambition: Reference Sheet and Implementation Guidance" (New York: UN Global Compact, 2021), https://www.unglobalcompact.org/library/5887.
[21] Kate Crawford, *Atlas of AI: Power, Politics and the Planetary Costs of Artificial Intelligence* (Yale and London: Yale University Press, 2021).
[22] Fairwork, "Work in the Planetary Labour Market: Fairwork Cloudwork Ratings 2021" (Oxford, United Kingdom, 2021).
[23] Julian Posada, "Embedded Reproduction in Platform Data Work," *Information, Communication & Society* 25, no. 6 (April 26, 2022): 816–34, https://doi.org/10.1080/1369118X.2022.2049849.

the garment industry through collective bargaining at an industry level.[24] Other organisations like Unilever have leveraged their purchasing power to ensure that every worker in their supply chain will be paid a living wage by 2030.[25] Increasing fairness in the supply chain can have ripple effects that produce ethical outcomes on a larger scale and can act as a force multiplier across the global division of labour. On the specific questions of sustainability, the UN Global Compact have developed guidance for practical action by private sector actors.[26]

It can be complex for organisations to understand the details of that a diffuse production network. However, this complexity is not a viable excuse to avoid accountability. Organisations have a duty to fully understand two things when procuring good and services: first, the functionalities of what they are buying (particularly when AI systems are being purchased); and second, the labour conditions involved in the production of that good or service.

- Organisations must actively pursue the implementation of fair work standards (as defined in principle one) throughout the production network, including but not limited to: a living wage for all workers; freedom of association and collective bargaining; the elimination of child and forced labour; equality of opportunity and treatment; fair working time; and health and safety. In the first instance, this pursuit should consist of a production network labour standards audit, followed by the development of an action plan which uses the organisation's procurement leverage to push for ethical change. This plan of action should be developed and implemented through a mechanism which includes collective worker voice (see principle 10 below).
- The above action plan must also contain concrete steps towards sustainability that reflect the existential threat of anthropogenic climate change. These should pursue goals such as: UN SDG 7 (affordable and clean energy); SDG 11 (sustainable cities and communities); SDG 12 (responsible consumption and production); 13 (climate action); 14 (life below water); and 15 (life on land.)
- In addition, organisations should ensure that any cloudwork involved in AI development is conducted in fair conditions, in line with the Fairwork cloudwork principles.[27] Organisations should consider the guidance provided by the Partnership on AI specifically relating to cloudwork as useful supplementary tool to use in the development of internal best practice guidance on cloudwork.[28]

3. Promote explainability

Workers have a need to understand the processes and technologies governing their own work. A wide range of characteristics associated with high job quality are linked to this form of transparency: task discretion and autonomy; task clarity; training and learning opportunities; opportunities for self-realisation; intrinsic rewards, and more.[29] Transparency is therefore an important component of any high-quality job. However, transparency is often impeded by the concentration of knowledge and power about the organisation and the production process in the hands of managers. The deployment of AI systems can either reinforce or challenge this concentration. If the opportunities of AI implementation

[24] The ACT Initiative, "The ACT Initiative: A Global Commitment on Living Wages" (Berlin, 2020), https://actonlivingwages.com/app/uploads/2021/04/ACT-on-Living-Wages-1.pdf.
[25] Unilever, "How We'll Help Build a More Equitable and Inclusive Society," Unilever, 2021, https://www.unilever.com/news/news-search/2021/how-we-will-help-build-a-more-equitable-and-inclusive-society/.
[26] UN Global Compact, "Supply Chain Sustainability: A Practical Guide for Continuous Improvement (2nd Ed)," 2015, https://respect.international/supply-chain-sustainability-a-practical-guide-for-continuous-improvement-second-edition/.
[27] Fairwork, "Work in the Planetary Labour Market: Fairwork Cloudwork Ratings 2021."
[28] Partnership on AI, "Responsible Sourcing of Data Enrichment Services," 2021, https://partnershiponai.org/paper/responsible-sourcing-considerations/.
[29] Sandrine Cazes, Alexander Hijzen, and Anne Saint-Martin, "Measuring and Assessing Job Quality: The OECD Job Quality Framework," OECD Social, Employment and Migration Working Papers, vol. 174, OECD Social, Employment and Migration Working Papers, December 18, 2015, https://doi.org/10.1787/5jrp02kjw1mr-en For more on the OECD Job Quality Framework, see principle 8 below. .

are to be seized then we must guarantee this second outcome and make AI systems transparent to the people who work alongside them by providing comprehensible explanations of AIs' design, function, limitations, and behaviours.

Explainability can also foster trust in AI systems.[30] This gain in trust can lead to more effective working patterns and increased cooperation with novel AI systems. However, it is important to remain attentive to the negative side of this trust. Research has begun to identify the insufficiency of human supervision to correct against the failures and inaccuracies of AI systems. The widespread assumption that such 'human in the loop' policies mitigate the risks of AI system decision making does not accord with the emerging empirical evidence.[31] In fact, some studies indicate that the growth in trust gained by providing human supervisors with explanations may make them more prone to blindly trusting the AI system, thereby reducing decision quality.[32]

With these necessities in mind, organisations must ensure that they make AI *explainable*. That is to say, they should translate technical processes and decision outputs into intelligible, comprehensible formats which are suitable for evaluation by workers.[33] This concept has already been the focus for significant debate and discussion.[34] Machine Learning centric non-symbolic approaches to AI are difficult to explain. The weights assigned to nodes in a neural network do not carry any symbolic meaning, and simply function to make the network produce the intended output. It is now widely understood that the designers of such networks may have almost as limited understanding of how the network makes a decision in conceptual terms as the worker who directly engages with the AI system. Despite this challenge, there are a huge range of methods both in development and already available which can be used to attain the required level of explainability.[35]

This principle stresses the need for AI systems to be explainable in two different senses. First, workers need to be able to understand the 'model behaviour' of the AI systems that impact their labour process,

---

[30] Ehsan Toreini et al., "The Relationship between Trust in AI and Trustworthy Machine Learning Technologies," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona Spain: ACM, 2020), 272–83, https://doi.org/10.1145/3351095.3372834.

[31] Ben Green, "The Flaws of Policies Requiring Human Oversight of Government Algorithms," *Computer Law & Security Review*, 2021, https://doi.org/10.2139/ssrn.3921216.

[32] Gagan Bansal et al., "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21: CHI Conference on Human Factors in Computing Systems, Yokohama Japan: ACM, 2021), 1–16, https://doi.org/10.1145/3411764.3445717; Maia Jacobs et al., "How Machine-Learning Recommendations Influence Clinician Treatment Selections: The Example of Antidepressant Selection," *Translational Psychiatry* 11, no. 1 (June 2021): 108, https://doi.org/10.1038/s41398-021-01224-x; Vivian Lai and Chenhao Tan, "On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19: Conference on Fairness, Accountability, and Transparency, Atlanta GA USA: ACM, 2019), 29–38, https://doi.org/10.1145/3287560.3287590.

[33] Jessica Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI" (Berkman Klein Center for Internet& Society, January 15, 2020), 42, https://dash.harvard.edu/handle/1/42160420.

[34] Amina Adadi and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* 6 (2018): 52138–60, https://doi.org/10.1109/ACCESS.2018.2870052.

[35] Adadi and Berrada; Wojciech Samek and Klaus-Robert Müller, "Towards Explainable Artificial Intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed. Wojciech Samek et al., vol. 11700, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2019), 5–22, https://doi.org/10.1007/978-3-030-28954-6_1; Carlos Zednik, "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence," *Philosophy & Technology* 34, no. 2 (June 2021): 265–88, https://doi.org/10.1007/s13347-019-00382-7; Plamen P. Angelov et al., "Explainable Artificial Intelligence: An Analytical Review," *WIREs Data Mining and Knowledge Discovery* 11, no. 5 (September 2021), https://doi.org/10.1002/widm.1424.

either directly or indirectly.[36] Second, workers need to be able to understand specific decisions taken about them in order to ensure fairness (see principle 5 below). This explainability should be combined with resources which seek to make visible the ways in which AI systems can prove fallible in order to enable workers to engage critically with the technology deployed in their workplace. The explainability of an AI system must be further supported over time through training processes and social dialogue with both individual and collective worker voice. This ongoing process of information sharing will ensure that the right to understand is meaningful in practice.

Benchmarks

- Organisations must provide workers with suitable and comprehensible explanations of the model behaviour of AI systems which directly or indirectly impact them in the workplace. These can be produced via any suitable of methodology but must be shown to address any fundamental questions raised by workers and include all essential information, such as what forms of data are inputs to the system.
- Additionally, organisations must have the capacity to provide specific explanations for specific AI system decisions (see principle 5.)
- Organisations should facilitate ongoing processes of training that provide workers with a critical understanding of AI systems and their potential fallibility.
- Workers and trade unions must have access to a process for making further enquiries about the data involved in and behaviour of an AI system.

4. Strive for equity

Some of the most pronounced risks of AI systems arise from the way in which they interact with patterns of social discrimination. Not only can they reproduce them, but they can also scale them up and embed them within apparently objective processes of prediction, decision making, and classification. To take just one example, language models can encode the hegemonic world views expressed in their training data, reproducing and amplifying the biases contained within it – and then those language models can subsequently be used in a huge range of use cases.[37] As this example demonstrates, training data of all kinds, from text scraped from the internet to statistics collected by government bodies, all have cultural, moral, social, economic and ethical norms encoded within them.[38] This reproduction of power dynamics can occur in surprising ways as a result of the social complexity of power itself. A correlation between a seemingly uncontroversial variable and a protected characteristic can reproduce discrimination by proxy in ways that are very difficult to predict in advance.[39] All of which is to say, the datafication of a phenomenon is not a neutral process, even if it appears like one. The patterns of power that shape data also shape the outcomes of AI systems trained on that data – and sometimes they do so in surprising ways.

In this sense, AI systems can provide a mirror in which we can recognise existing regimes of power and their discriminatory outcomes anew. Consider the famous example of the Pro Publica investigation into Northpointe's risk score software, which informed decisions on bail across the US justice system. Rather than a pattern of discrimination informed by social power being reproduced by the decisions of thousands of discrete individuals, all of whom appear to be acting in an uncoordinated manner, the investigation found a centralised pattern of differential decision making based on race.[40] At the same time, AI systems have the potential to initiate a contradictory scenario in which a centralised decision maker reproduces patterns of discrimination at huge scale whilst appearing objective to outside

[36] Samek and Müller, "Towards Explainable Artificial Intelligence."
[37] Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada: ACM, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.
[38] Crawford, *Atlas of AI: Power, Politics and the Planetary Costs of Artificial Intelligence*.
[39] Michal Kosinski, David Stillwell, and Thore Graepel, "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior," *Proceedings of the National Academy of Sciences* 110, no. 15 (April 9, 2013): 5802–5, https://doi.org/10.1073/pnas.1218772110.
[40] Julia Angwin et al., "Machine Bias," *Pro Publica*, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

observers, whilst also centralising the agency of a diffuse structure of social power into a single technological artefact in such a way that that diffuse structure may prove more contestable to those on the wrong end of it. The challenge posed by this principle to all stakeholders is that they should look to prevent the first phenomenon, and act on the opportunities presented by the second.

This is no small challenge, given the way in which AI system production and deployment is led by some of the most powerful actors in government and the private sector. Compliance would demand that these actors often make decisions which contradict the structure of power they directly benefit from. As such, a more likely source of positive change is the agency of the disempowered, whose interests directly align with the spirit of this principle. One of the mechanisms through which this subaltern agency can be brought to bear is through collective bargaining, discussed further in principle 10 below.

If this challenge is to be met, then it is vital that the AI lifecycle involved repeated opportunities to detect and eliminate sources of discrimination. This requires designing AI systems in such a way that they are auditable over questions of discrimination, and so that there are multiple internal assessments conducted at different points in time over the entire course of the system's operation. As discussed in the policy recommendations below, governments should create systems to ensure that these audits can be subject to external scrutiny by relevant experts in order to guarantee high standards.

Benchmarks:

- Organisations should create an internal audit process that operates across an AI system's lifecycle and enable independent observation and replication of this process.[41] This audit should target both bias as broadly understood and patterned inequality.[42] The six-stage process for an Equality impact assessment that builds on such an audit to generate equitable outcomes (through collective bargaining with a trade union) proposed by the Institute for the Future of Work can act as a template to assist organisations in this process.[43]
- Public scrutiny should be utilised as a tool that can incentive equitable outcomes. Organisations should harness this by publishing a statement of principle outlining the values that guide the organisation's development and implementation of workplace AI. This statement should be subject to regular review through a collective bargaining process or anonymous consultation in workplaces where workers are not represented by trade unions (see principle 10 below). An anonymous feedback mechanism should then be created through which concerns about the reality of workplace AI implementation for both individual and group-level decisions can be fed into the process discussed above.

5. Make fair and accountable decisions

Every workplace is a site of complex decision making. Some of these decisions are fair, but many are not. As discussed above, the structures of social power that exist in a workplace inevitably play a role in determining which stakeholder interests are served by most decisions, most of the time. As AI systems increasingly supplement and/or replace elements of existing decision-making processes, they will become entrapped in the complex questions of fairness inherent to them. Once again, AI systems bring with them a dual potential to both reproduce unfairness at scale and centralise unfairness in a way which makes it open to challenge.

One of the dominant responses to the risk posed by AI systems has been to propose that they are not allowed to make decisions without human supervision. This argument has led to a common specification that decision loops must always involve a human. The assumption of legislation like Article 22(1) of the GDPR (which prohibits decisions made solely on the basis of automated data processing) is that such

[41] For a discussion of existing audit tools, their limitations, and the variation in possible definitions of bias, see Logan Graham et al., "Artificial Intelligence in Hiring: Assessing Impacts on Equality" (Institute for the Future of Work, 2020).
[42] Benjamin Eidelson, "Patterned Inequality, Compounding Injustice, and Algorithmic Prediction," *American Journal of Law and Equality* 1 (September 1, 2021): 252–76, https://doi.org/10.1162/ajle_a_00017.
[43] Graham et al., "Artificial Intelligence in Hiring: Assessing Impacts on Equality."

inclusion will make decisions fairer. However, as discussed above, the evidence suggests this is not the case. Human supervision of algorithmic decision making can often provide only superficial protection from unfair outcomes because of skills, knowledge and time shortfalls which impair the human actor's supervisory ability.[44] To that widely recognised list of potential problems, we should also add the contradictory incentives provided by divergent stakeholder interests. A human-in-the-loop is not a panacea.

Placing humans within the decision-making loop does not significantly mitigate risk - but it is our belief that empowering the human who is the subject of that loop may well do so. There is significant precedent for this approach: in judicial decision-making systems, for example, the subject of a decision is the one who uses appeal processes to challenge the fairness of that decision. Rather than solely relying on an external 'court observer' to check judgements for fairness and initiate appeals, we empower the agent who is most directly impacted by the decision and who has a strong material interest in fair decision making. This leads us to the conclusion that the best currently available defence against the interest structure of various workplaces and the unfair decisions they may produce is a strong set of rights vested in all workers directly or indirectly impacted by AI systems. These rights are fourfold: to receive a specific explanation of a decision made by an AI system; to challenge that decision via a multi-stakeholder process if they feel the decision was unjust; to receive direct remedy to undo the results of an unfair decision; and to see that the decision-making process is updated in a way which prevents such unfairness reoccurring.

In addition, the necessity of moving beyond excessive faith in the human-in-the-loop allows for us to adopt approaches which stress the accountability of key actors, whose influence on the earlier stages of the AI system lifecycle is so important. An institutional oversight approach, which has been theorised in a public sector context, provides a useful analogy.[45] Rather than making humans-in-the-loop the key point of accountability for AI decisions, this proposal shifts the burden up the chain to institutional leaders. Where negative consequences result from AI system deployment, those with the decision-making power to shape AI systems must be held accountable.

Benchmarks:

- Respect workers' rights to receive explanations of decisions affecting them, to make appeals when they consider those decisions to be unfair, to receive remedial action if their appeal is upheld, and to ensure that improvements are made to prevent the recurrence of any issue.
- Create easily accessible processes through which workers can ask for an explanation for a specific decision taken or informed by AI systems which impacts upon their work.
- Create an appeals system, with significant worker voice representation, through which workers can appeal any decision taken or informed by AI. This system should have the capacity to overturn decisions and apply remedial steps in the wake of its findings.
- This system should feed directly into a process for revising the implementation of AI in wake of successful appeals, and reporting back to the appellant on the changes made.
- Recognise the accountability of senior stakeholders for AI system decision making, and act accordingly.

6. Use data fairly

Datasets are the raw material of contemporary non-symbolic artificial intelligence. Without it, machine learning is impossible. As a result, they are a central component of AI system development and have become assets of significant importance for the contemporary economy. This valuation incentivises organisations to collect and process vast amounts of data, and the trend extends to the workplace. But the collection of data in the workplace has the potential to concentrate power in existing structures in a way that reduces social cohesion and creates and exacerbates risks.[46] The datafication of the workplace

[44] Green, "The Flaws of Policies Requiring Human Oversight of Government Algorithms."
[45] Green.
[46] 'Advancing Data Justice Research and Practice: An Interim Report for the 2021 GPAI Paris Summit' (Alan Turing Institute, 2021), https://gpai.ai/projects/data-governance/data-justice/advancing-data-justice-research-and-practice-interim-report.pdf.

needs to be restrained if fair outcomes are to be achieved.

In the first instance, this means taking care to limit the kinds of data collected. Data minimization is now a widely established principle that has been through multiple historic iterations,[47] and documents like the European Commission's proposal for a directive on platform work (2021) recognise the need to extend this principle to the work context. It is a central component of any fair approach to data in the workplace, because minimisation pre-empts a wide range of possible misuses of data by not collecting it in the first place. As with the GDPR, the principle of data minimisation can be simply expressed as collecting the minimum amount of personal data required to fulfil a legitimate purpose. The blanket collection of potentially useful personal data is not legitimated just by the fact that it occurs in a workplace context. The EU Commission's proposal is a useful guide as to what kinds of data should not be collected.

Legitimate data collection and processing must respect the interests of workers, not just the interests of companies, governments and other organisations. Data collection for the purposes of increasing the productivity and safety of the labour process is not justified in and of itself, if such collection also has potential negative impacts for the workers concerned. As such, the decision to collect a particular form of data must weigh these potentially divergent interests against one another. Where decisions are being made about what appears to be a zero-sum trade-off between safety and privacy, workers must be involved at every stage. The most effective way to weigh these factors is always likely to be through multistakeholder engagement that includes collective worker voice.

However, even when the principle of minimization is rigorously applied, it is also essential to consider additional complexities such as the use of consent as a justification for data processing. Given the imbalance of power between workplace stakeholders, 'consent' may be a problematic basis for data processing in the workplace. 'Take it or leave it' data policies leave workers unable to genuinely exercise judgement about data collection and do not provide evidence of informed consent. If refusing to consent means losing access to earnings or employment, then any agreement will have been reached by coercive means. As such, opt-out clauses or other sources of justification should be sought in most instances.

When data have been collected on a justified basis, it is important to recognise the complexities of workplace data ownership. Data regulation increasingly recognises that data is not solely owned by the party that collects it, but also the subjects who produced it. In order to accord with this principle, organisations should demonstrate an ongoing commitment to exploring potential ways of sharing data ownership, such as the creation of democratic data trusts and the granting of collective data rights to trade unions.

In addition, workers have a range of rights to data protection and access that must be consistently applied. Workers should be able to access three forms of personal data: input, observed and inferred.[48] Importantly, this means that workers do not just have a right to access whatever data they have provided directly during their work, but also data that has been generated about them and their work through observation and processing.

Benchmarks

- Collect only personal data that is strictly necessary for the labour process. For example, do not collect data on private conversations, the emotions or health of workers, or excessive detail on their moment-to-moment practical activity. The general principle should be that data should be gathered to inform or improve the labour process, not in order to discipline, surveil, or intensify control over workers. This includes applications that track worker behaviour to enhance worker productivity or mental/physical well-being. These applications should not be run in the background without the direct and informed consent of workers, and workers should be invited to opt-in, rather than opt-out.

[47] Peter Hustinx, "Privacy by Design: Delivering the Promises," *Identity in the Information Society* 3, no. 2 (August 2010): 253–55, https://doi.org/10.1007/s12394-010-0061-z.
[48] Worker Info Exchange, "Managed by Bots: Data-Driven Exploitation in the Gig Economy," 2021, https://www.workerinfoexchange.org/wie-report-managed-by-bots.

Workers who opt-out from such applications should not be disadvantaged as a result of their decision.

- Recognise workers' right to request all relevant personal data (input, observed and inferred). These requests should use the GDPR as a basic minimum set of rights, even in jurisdictions where the GDPR does not directly apply. The results should be provided in a portable format, and the principle of explainability must be applied to ensure that the processing and use of any data collected should be transparent to the workers who produced that data.
- Demonstrate an ongoing commitment to exploring potential ways of sharing data ownership with workers, such as the creation of democratic data trusts and the granting of collective data rights to trade unions.

7. Enhance safety

In 2022, workers' right to occupational safety and health was added to the ILO's fundamental principle and rights at work, meaning that all member states are now committed to conventions No.155 and No.187. This demonstrates both the fundamental significance of occupational safety and health at work, and the degree of consensus around this principle.

Positive developments in occupational safety and health are one of the major opportunities afforded by AI systems – but the reality is not so straightforward. The US warehousing sector is a case in point. Examples of workplace AI application already demonstrate that a resulting increase in labour productivity leads to unsafe outcomes.[49] In fact, some research has found that the greater the degree of technological complexity in a warehouse work environment, the higher the injury rate.[50] Concerns also abound in areas like platform food and parcel delivery, where initial work suggests significant reductions in safety and health as a result of AI-led transformations of the labour process.[51] The introduction of AI is a major component of contemporary effort-biased technological change, a factor which has historically contributed to significant increases in work intensity.[52] Given the ongoing trend toward work intensification, any developments which might contribute to a further acceleration of this potentially dangerous trend must be closely monitored.[53] There is growing evidence of a cross-sectoral trend whereby the introduction of AI systems that manage work task distribution and/or the coordination and supervision of the labour process can come into conflict with workers' rights to occupational safety and health. Organisations should apply specific caution when implementing any AI system which might contribute to this trend. In particular, organisations must recognise that technology which intensifies work towards a maximum notionally safe level may, rather than mitigating risks, introduce new ones.

Different stakeholder groups will have different understandings of various trade-offs impacting on occupational safety and health particularly on issues, hence the importance of negotiating a common position through a process of collective bargaining and potentially co-design wherever possible (see principle 10.) Alongside this commitment to social dialogue, organisations have a responsibility to devote resources to ensure the labour process is safe. Article 7 of the EU Commission's proposal on platform work (2021), which requires organisations to employ an independent internal reviewer to assess the

[49] Gutelius Beth and Theodore Nik, "The Future of Warehouse Work: Technological Change in the US Logistics Industry." (UC Berkley Labour Centre, 2019), http://tankona.free.fr/warehousework1019.pdf.
[50] Strategic Organising Centre, "Primed for Pain," May 2021, https://thesoc.org/amazon-primed-for-pain/.
[51] Nicola Christie and Heather Ward, "The Emerging Issues for Management of Occupational Road Risk in a Changing Economy: A Survey of Gig Economy Drivers, Riders and Their Managers" (London: UCL Institute for Transport Studies, August 20, 2018), http://www.ucl.ac.uk/news/news-articles/0818/200818-gig-economy-drivers-traffic-collisions; Callum Cant, *Riding for Deliveroo: Resistance in the New Economy.* (Cambridge: Polity Press, 2019); Hua Qin et al., "An Observational Study on the Risk Behaviors of Electric Bicycle Riders Performing Meal Delivery at Urban Intersections in China," *Transportation Research Part F: Traffic Psychology and Behaviour* 79 (May 2021): 107–17, https://doi.org/10.1016/j.trf.2021.04.010.
[52] Francis Green et al., "Working Still Harder," *ILR Review*, January 27, 2021, 001979392097785, https://doi.org/10.1177/0019793920977850.
[53] Matea Paškvan and Bettina Kubicek, "The Intensification of Work," in *Job Demands in a Changing World of Work*, ed. Christian Korunka and Bettina Kubicek (Cham: Springer International Publishing, 2017), 25–43, https://doi.org/10.1007/978-3-319-54678-0_3.

occupational health implications of algorithmic management practices, should be seen as a baseline in workplaces where AI is deployed. At all four stages of the AI system lifecycle, safety must be a consistent consideration.

Benchmarks

- Organisations must guarantee a safe and healthy environment for workers, considering all dimensions of both physical and mental health. This guarantee should be actively maintained through monitoring of the labour process, considering all points of interaction with AI systems.
- Where an AI system sets work task allocation, it should not be capable of demanding workers engage in an unsafe pace of work. The determination of a safe pace should be made through multistakeholder processes which make significant room for collective worker voice and must be adapted to individual circumstances as and when necessary.
- Organisations should use the implementation of AI in the workplace as an opportunity to actively improve health and safety, where the opportunity arises. To do so, they should embed processes of reflection into safety monitoring which look at potential improvements that can be fed back into the design of the labour process.
- Organisations should respect the distinction between working and non-working time and implement a rigorous 'Right to Disconnect' in circumstances where the combination of AI systems and remote communications risk undermining that distinction.

8. Create future-proof jobs

AI systems are already transforming the way workers experience work. From the task-by-task organisation of the labour process to the terms and conditions of employment, they are shaping the daily lives of millions of people. This impact will only continue to grow over time. At the heart of this principle is the idea that such processes of transformation and creation must be harnessed for collective social benefit. However, the current course of development indicates that there are significant risks that this will not be the case. These risks can be schematised into two camps: first, job destruction; and second, reductions in job quality.

AI systems tend to have a net neutral impact on the level of overall employment when considered at the scale of a whole economy.[54] But on a smaller scale, the automation of specific tasks may lead to certain job roles within an organisation becoming redundant. Job destruction is linked to significant negative effects ranging from a labour force drop out and earnings loss[55] to declining mental and physical health.[56] These impacts are profoundly social in nature: they are best understood as collective phenomena with widespread impact, rather than as an individual malaise. As a result, governments have a widely recognised responsibility to lead other economic actors in a course of development which minimises job destruction and mitigates any negative effects of AI system introduction on a social scale.

The OECD has an established Job Quality Framework with three components: earnings quality, labour market security, and quality of working environment.[57] The introduction of AI systems to a workplace or industry can interact with all three of these components in a positive or negative way. Take for example how last mile delivery services have been transformed through the advent of the platform economy in a way that has often negatively impacted all three. It is self-evident that any reduction in earning quality of labour market security as a result of the introduction of AI systems contradicts this principle, but it is worth discussing the third factor, quality of the working environment, at some more length. The OECD

[54] Spencer, D., Cole, M., Joyce, S., Whittaker, X. and Stuart, M. (2021). *Digital automation and the future of work*. Brussels: European Parliament. https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)656311 (accessed 5 February 2021).
[55] Kristiina Huttunen, Jarle Møen, and Kjell G. Salvanes, "How Destructive Is Creative Destruction? Effects of Job Loss on Job Mobility, Withdrawal and Income," *Journal of the European Economic Association* 9, no. 5 (October 2011): 840–70, https://doi.org/10.1111/j.1542-4774.2011.01027.x.
[56] Robert J. Hironimus-Wendt, "The Human Costs of Worker Displacement," *Humanity & Society* 32, no. 1 (February 2008): 71–93, https://doi.org/10.1177/016059760803200105.
[57] Sandrine Cazes, Alexander Hijzen, and Anne Saint-Martin, "Measuring and Assessing Job Quality."

have provided guidelines for understanding the quality of the working environment through several objective job characteristics.[58] Characteristics that pose challenges to quality are termed as 'job demands', whilst those which tend to increase it are termed as 'job resources':

Table 1. Dimensions and characteristics of quality in the job environment

Dimensions	Job characteristics	
	Job demands	Job resources
A. Physical and social environment	A.1. Physical risk factors A.2. Physical demands A.3. Intimidation and discrimination at the workplace	A.4. Social support at work
B. Job tasks	B.1. Work intensity B.2. Emotional demands	B.3. Task discretion and autonomy
C. Organisational characteristics		C.1. Organisation participation and workplace voice C.2. Good managerial practices C.3. Task clarity and performance feedback
D. Worktime arrangements	D.1. Unsocial work schedule	D.2. Flexibility of working hours
E. Job prospects	E.1. Perceptions of job insecurity	E.2. Training and learning opportunities E.3. Opportunity for career advancement
F. Intrinsic aspects		F.1. Opportunities for self-realisation F.2. Intrinsic rewards

To accord with this principle, AI systems should contribute to the diminishment of the demands of a job and expand the resources that a worker has available to them. From this perspective we can see more clearly how some of the risks of AI systems can manifest themselves. The use of technology to partially automate a job and break up what remains into smaller tasks and thereby reduce the skill required to do the job has been termed 'deskilling.' [59] Such a process is likely to diminish a wide range of resources (B.3.; C.3.; E.2.; E.3.; F.1.; F.2.) and increase a wide range of demands (A.2.; B.1.; E.1.). Such deskilling would, clearly, never be in line with this principle. On the contrary, the introduction of AI systems must lead to improvements in job quality along the lines of any of the three components laid out in the OECD framework. The specifics of what improvements should be made and how is a question that can only be settled through a multistakeholder approach.

That is why AI systems should provoke all parties to engage in a process of participatory redesign, through which the benefits of technological development can be passed on to workers in a form which they have an active role in determining. The possibilities are huge: productivity gains realised through the automation of mindless work could create the space for a four-day working week; for new forms of positive social impact; for greater creative freedom; for higher wages; and many other potential benefits. It is not enough to avoid harmful uses of AI systems – all parties have a positive responsibility to create beneficial outcomes.

In the limited circumstances where the destruction of tasks leads to a situation where jobs are at risk, then an absolute emphasis must be placed on job redesign, retraining, and the creation of new opportunities, before any compulsory redundancies are considered. Through any such process an emphasis must be placed on allowing workers the absolute maximum degree of choice and control over their future. This will require extensive engagement with collective worker voice throughout.

Take, for example, warehouse automation. A warehouse in Oslo underwent a broad redesign to increase the level of automation. Against the suggestion of one of the technology firms which was advising the company which operated the warehouse, they did not implement compulsory redundancies for the old warehouse workers – instead, these workers were upskilled into roles as technicians to run the new facility. This significantly increased the resources associated with the job: workers went from physically picking goods to monitoring automated systems and problem solving to maintain overall

[58] OECD, *OECD Guidelines on Measuring the Quality of the Working Environment* (Paris: OECD Publishing, 2017), https://doi.org/10.1787/9789264278240-en.
[59] Harry Braverman, *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century* (New York: Monthly Review Press, 1975).

efficiency. This was by no means a perfect example – productivity gains remained very unevenly distributed between stakeholders – but it demonstrates that AI systems can be introduced into a workplace without job losses.[60]

Benchmarks:

- Use the introduction of AI as an opportunity to increase job quality. This exploration should be part of the process of engaging with worker voice prior to any major change (see principle 10 below.) Opportunities for job quality increases may involve increases in task discretion and autonomy; intrinsic rewards; and opportunity for self-realisation, amongst others.[61]
- Give significant weight to the negative impacts of job destruction and deskilling in the development process and intentionally direct design to minimise those negative impacts wherever possible. Where job destruction appears unavoidable, consultation with workers should begin as early as possible to begin to look for alternatives such as reemployment in new roles with full retaining, and upskilling. If no alternative is found after substantial consultation, then in the final instance organisations must offer significant redundancy packages and investigate avenues for the long-term support and retaining of those made redundant.

9. Avoid inappropriate deployment

When dealing with the risks posed by AI systems, our goal should be to prevent harms in advance rather than in retrospect. It is not acceptable to deploy under-tested systems in a high-risk environment and the iterate to deal with problems as they arise. Proactive action is essential to limit harm, maintain social dialogue and trust between all stakeholders, and maximise the possible benefits of AI systems. This means that the first two steps in the AI system lifecycle (data, design and models, and verification and validation) need to be the subject of significant attention if optimal outcomes are to be achieved.

It is not enough for a sole stakeholder to dedicate resources to this process without wider engagement. Only with a diversity of perspectives will organisations be able to identify all the necessary steps required to mitigate harms and maximise opportunities. Participatory design is showing promise as an approach which can allow for the integration of worker preferences into the creation of AI systems, particularly when those AI systems fall within the bounds of 'algorithmic management.'[62] Real-world examples of this approach applied on a small scale have demonstrated that it is possible to build highly effective models that successfully represent the preferences and beliefs of human users and produce positive outcomes. The unique characteristics of AI systems which make use of Machine Learning may pose challenges to this approach, particularly as systems evolve over the deployment phase in response to novel data.[63] However, if regular chances for vertical feedback and the renegotiation of system specifics

[60] Interview with Victor Figueroa of the International Transport Federation.
[61] Sandrine Cazes, Alexander Hijzen, and Anne Saint-Martin, "Measuring and Assessing Job Quality."
[62] Min Kyung Lee et al., "WeBuildAI: Participatory Framework for Algorithmic Governance," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 1–35, https://doi.org/10.1145/3359283; Min Kyung Lee et al., "Participatory Algorithmic Management: Elicitation Methods for Worker Well-Being Models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event USA: ACM, 2021), 715–26, https://doi.org/10.1145/3461702.3462628; Angie Zhang et al., "Algorithmic Management Reimagined For Workers and By Workers: Centering Worker Well-Being in Gig Work," in *CHI Conference on Human Factors in Computing Systems* (CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans LA USA: ACM, 2022), 1–20, https://doi.org/10.1145/3491102.3501866.
[63] Tone Bratteteig and Guri Verne, "Does AI Make PD Obsolete?: Exploring Challenges from Artificial Intelligence to Participatory Design," in *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2* (PDC '18: Participatory Design Conference 2018, Hasselt and Genk Belgium: ACM, 2018), 1–5, https://doi.org/10.1145/3210604.3210646.

are integrated into stage 4 of the system lifecycle (operation and monitoring) these may prove less significant than feared.

As well as stakeholder participation, strong testing processes are also characterised by going beyond basic validation. A model may well be shown not to over or underfit and generalise to produce useful outcomes, but it also necessary to test how the AI system operates in the context of the labour process. Pilot schemes which introduce AI systems into a low-risk, highly monitored environment for a fixed period and then conclude with multistakeholder assessment are essential for most deployment use cases. The use of synthetic data may play a role in these kinds of tests, particularly in the early stages.[64] Such pilot schemes should also include dedicated testing in abnormal scenarios to develop a stronger understanding of model behaviour in the kind of varied conditions that exist in the workplace. This phase also offers organisations a chance to further develop impact assessments that specifically address how the proposed system interacts with some of the issued raised in the principles above. The degree of testing required should be varied depending on the significance of the risks associated with the system, with riskier systems subjected to more prolonged testing with higher bars to clear before deployment is considered.

In circumstances where organisations are procuring AI systems from an external provider, this principle also reemphasises the need for a full and rounded understanding of the system in question. A full assessment of how an AI system was trained, what data it utilises, what functions it is capable of performing within the labour process, what implications it has for all stakeholder groups, and how any of those may potentially change over time is a fundamental precondition for any procurement decision. Procured systems must be subjected to adequate testing by the purchaser prior to deployment.

Benchmarks:

- Organisations must be able to evidence that they have sufficiently tested all AI systems which are deployed to the workplace. This testing should reflect the scale of risk posed by the system itself: those systems with higher potential harms should be subjected to a more extended regime of testing, with a higher bar to clear before deployment is considered.
- All stakeholders should have a role in designing and testing AI systems that are deployed into their workplaces. Organisations must consult and/or negotiate with workers and/or trade unions in an appropriate manner throughout phases one and two of the AI system lifecycle (see principle 10 below.) Participatory design approaches should be considered best practice in most circumstances.

10. Advance collective worker voice

Previous statements of ethical AI principles informed by multiple stakeholders have struggled to be translated into practice because of a lack of suitable mechanisms for resolving different interests when discussing concrete details regarding implementation. What has been missing has been a mechanism which allows for ongoing negotiation at the workplace level. Collective bargaining can play exactly this role.[65]

By ensuring workers can participate in decision making, measures to facilitate worker voice will allow for the codetermination of the details of the labour process by all stakeholders. Organisations and trade unions should engage in a positive process to negotiate the details of the ethical implementation of workplace AI. This process should go further than shallow consultation and embrace collective bargaining as an ongoing iterative process which is always responding to changes in the labour process. Attempts to classify issues relating to the technology used in the workplace as 'managerial prerogative' and therefore exempt from bargaining fundamentally fail to respect the multi-stakeholder nature of the labour process. Ethical AI implementation requires practical processes that respect the interests of all stakeholders.

[64] Sergey I. Nikolenko, *Synthetic Data for Deep Learning*, Springer Optimization and Its Applications, volume 174 (Cham: Springer, 2021), https://doi.org/10.1007/978-3-030-75178-4.
[65] Valerio De Stefano and Simon Taes, "Algorithmic Management and Collective Bargaining" (Brussels: European Trade Union Institute, 2021), https://www.etui.org/publications/algorithmic-management-and-collective-bargaining.ht.

This approach has proven successful in contexts such as H&M stores in Germany, where the trade union *ver.di* is involved in negotiating the introduction of RFID technology to avoid negative impacts on workers such as deskilling, work intensification, unwarranted increases in managerial control, workforce segmentation, and increases in precarity.[66] In these circumstances, workers who attempted individual negotiation with management over the introduction of technology would be unlikely to achieve this outcome. As a result, collective bargaining creates a relationship through which workers can be treated as agents in the design and implementation of AI systems, not just as subjects of those systems.

But given the limited coverage of collective bargaining agreements in the global economy, additional measures are required to allow for multistakeholder participation in decision making regarding AI systems. Organisations without existing agreements have a duty not to obstruct workers from organising by respecting their fundamental rights in line with the ILO *Declaration of Fundamental Principles and Rights at Work* (see principle 1), but also to go further and actively facilitate collective worker voice. In line with Article 15 of the EU Commission proposal on platform work (2021), steps must be taken to enable workers to communicate and organise. It is also important that whilst the ideal situation of formal collective bargaining is being pursued other approaches for considering collective worker voice are implemented, such as meaningful surveys and consultations.

Benchmarks:

In workplaces where workers are represented by a trade union:

- Use collective bargaining as a tool to guide the implementation of these principles in practice. In particular, the bargaining process should involve a regular review of the overall use of AI in the workplace; and specific negotiations whenever a new AI system is: a) being introduced for the first time or in a new function or b) an existing AI system is being substantially modified.
- In line with principle 3 above, organisations should support trade unions to access the relevant information and expertise to conduct informed bargaining.

In workplaces where workers are not represented by a trade union:

- Conduct annual anonymous consultation processes to collect feedback on workplace AI, with additional consultation processes whenever a new AI is being introduced or an existing AI is being substantially modified. The results of these consultations must be available to all workers and carry significant weight in organisational decision making.
- Facilitate collective worker voice by ensuring the availability of non-supervised communication channels for workers, e.g., forums for remote workers, in line with article 15 of the EU Commission proposal on platform work.[67]

In all workplaces:

- Facilitate collective worker voice by ensuring the availability of non-supervised communication channels for workers, e.g., forums for remote workers, in line with the EU directive on platform work.[68]

[66] Tatiana López et al., "Digital Value Chain Restructuring and Labour Process Transformations in the Fast-fashion Sector: Evidence from the Value Chains of Zara & H&M," *Global Networks*, December 16, 2021, glob.12353, https://doi.org/10.1111/glob.12353.
[67] European Commission, "Proposal for a Directive of the European Parliment and of the Council on Improving Working Conditions in Platform Work."
[68] European Commission.

Policy recommendations

The above principles establish an expansive vision of the fairer deployment of AI systems in workplaces. However, meeting the benchmarks above may require steps which individual stakeholders could either prove resistant to or struggle to take on their own. In this section, we discuss what role policymakers can play in creating an environment in which fair work can flourish and bring with it a range of widely recognised social benefits. Overall, we make six policy recommendations: observe and monitor AI system deployment; update regulation; empower labour inspectorates; empower information regulators; foster social dialogue; and distribute the gains of AI.

Observe and monitor AI system deployment

AI systems are already transforming labour markets and labour processes throughout contemporary societies. The implications of this transformation are profound, but unless qualitative and quantitative data on the specifics of how this change is occurring are gathered and analysed, governments will unable to shape this process effectively. Gathering this information is a precondition to the majority of effective policy action. As a result, governments should dedicate specific resources to the ongoing monitoring and observation of AI-related trends. The kind of data gathered and analysed will no doubt need to vary based on the specifics of different national contexts, but fundamental questions about the nature, scope and impacts of AI system deployment on the economy, workers, and citizens are likely to be relatively universal. Observation and monitoring activities should be conducted with the potential for novel technology to produce novel outcomes in mind, but this awareness must be balanced with the need to prioritise actually existing current forms of risk over potentially existing future forms of risk. It would be a mistake to concentrate resources on, for instance, monitoring the potential emergence of AGI (Artificial General Intelligence) instead of on a range of already-widespread use cases.[69] The process of data gathering should be designed with the interests of all stakeholders in mind, particularly those who are likely to have their voices marginalised as a result of inequalities in social power.

Update regulation

AI systems can challenge the adequacy of existing labour regulation. While existing standards like minimum wages should remain invariant regardless of the specific technology used in a workplace, some other kinds of labour regulation can either be rendered inoperable or actively circumnavigated by organisations using AI systems. Given these possibilities, governments must take an active approach to updating regulation in order to maintain an adequate legislative baseline of guaranteed standards. The OECD has said that AI systems should be fair, equal, and socially just.[70] Given the variety of AI actors and our collective experience of AI system deployment so far, it is clear that meeting this goal will require governments to take coordinated legislative and non-legislative action to update and enforce labour rights and regulations. The principles and benchmarks above can offer some guidance to regulators as to which areas may require particular attention – but specific information relevant to national contexts will need to be gathered to guide further action.

Empower labour inspectorates

Updated regulation will prove useless if not enforced. Labour inspection systems have long been vital to the implementation of labour regulations in practice across a huge range of national contexts.[71] When supported with sufficient resources and located within a suitable institutional framework, these systems can help turn fairness from a set of agreed principles into a concrete social reality. In the context of widespread and accelerating AI system deployment, these inspection systems remain vitally important. Actors within the labour inspection system should be capable of performing a series of vital functions if AI system deployment is going to reach high standards. For example: ensuring updated labour

[69] Cole et al., “Politics by Automatic Means?”
[70] OECD, “Recommendation of the Council on Artificial Intelligence.”
[71] Gianni Arrigo, Giuseppe Casale, and Mario Fasani, *A Guide to Selected Labour Inspection Systems: (With Special Reference to OSH)* (Geneva: ILO, 2011); International Labour Conference, ed., *Labour Administration and Labour Inspection: International Labour Conference, 100th Session, 2011; Fifth Item on the Agenda*, 1. ed, Report / International Labour Conference, 100,5 (Geneva: International Labour Office, 2011).

regulation is being implemented via inspections and sanctions; externally scrutinising audits and impact assessments; and maintaining high levels of OSH by investigating workplaces with excessive levels of issues associated with work intensification, such as musculoskeletal injuries and stress.

Empower information regulators

The centrality of data to AI systems means that information regulators have a huge role to play in ensuring fairness. The use of data to train and operate systems has to be conducted in a way which is compliant with the relevant regulation, and information regulators need adequate resources and powers to make sure that this is the case.

Given the expanded role that datasets are playing in all kinds of economic activity, there is also an expanding role for information regulators in actively supporting and facilitating the development of new forms of data ownership and governance. Governments should facilitate this by providing resourcing and access to policymakers so that ideas like democratic data trusts can move from theory to practice across the public sphere, thereby providing examples and structures that can be drawn upon across society.

Foster social dialogue

We have proposed collective worker voice as one of our ten principles above because the translation from ethical ideas to ethical practices requires multistakeholder participation. The experience of the Ethical AI field so far has made clear that only through the negotiation of multiple potentially opposed sets of interests can concrete progress be made in increasing the fairness of AI system deployment. Governments have a vital role to play in enabling this process by creating an environment conducive to collective bargaining and social dialogue. A variety of measures could be used to achieve this outcome: chief amongst these being those which directly support enterprise and sector-level collective bargaining. In addition, governments should prevent any restrictions on the right to organise, including activities associated with the expansion of workplace surveillance and AI systems. Other measures, such as mandating workers membership on company boards and creating a widespread 'right to access the workplace' for trade unions would all contribute to the same ends. Finally, governments can directly foster tripartite engagement by bringing a range of social partners into discussion over policy questions.

Distribute the gains of AI

The fair deployment of AI systems can produce substantial gains in efficiency and productivity. However, if these gains are monopolised by powerful stakeholders, these benefits will not be experienced throughout society. Governments should investigate the plethora of possible policy measures which could ensure the equitable distribution of AI-related gains, which range from a four-day working week with no loss of pay to universal basic services models.[72] Investment in the development of AI by public actors for socially beneficial ends may also result in significant benefits. Governments should consider the creation of public compute resources which can be accessed by a range of non-profit actors to support wider processes of AI development and oppose existing tendencies towards AI monopolisation.[73] All such measures should, of course, be designed to complement wider goals such as the UN SDGs and an ongoing just transition towards a zero-carbon future.

[72] Autonomy, "The Shorter Working Week: A Radical and Pragmatic Proposal," 2019, https://autonomy.work/portfolio/the-shorter-working-week-a-report-from-autonomy-in-collaboration-with-members-of-the-4-day-week-campaign/; Jonathon Portes, Howard Reed, and Andrew Percy, "Social Prosperity for the Future: A Proposal for Universal Basic Services" (London: UCL Institute for Global Prosperity, 2017), https://www.ucl.ac.uk/bartlett/igp/sites/bartlett/files/universal_basic_services_-_the_institute_for_global_prosperity_.pdf.

[73] Nick Srnicek, "Data, Compute, Labour," in *Digital Work in the Planetary Market*, ed. Mark Graham and Fabian Ferrari (The MIT Press, 2022), https://doi.org/10.7551/mitpress/13835.001.0001.

A Final Note: Structural barriers to fairness

In our process of consultation, we found that some stakeholders believe that serious structural barriers exist for the realisation of fairness in the workplace. From the incentive structures created by the dominant financial system, to the concentration of political and economic power, and the non-existence of democratic decision-making structures across many terrains of contemporary society, they foresaw significant difficulties in the implementation of what they considered fair practices around AI in the workplace. We agree with this analysis.

There is significant latitude for positive action at a variety of levels which can improve fairness: employers can change their practices for the better; governments can introduce new, beneficial regulation; regulators can design and implement effective policy; and unions and workers can negotiate better deals. All stakeholders have a responsibility to pursue these attainable improvements. But it seems indisputable that sooner or later, making progress towards realising the principles outlined above will mean challenging the structures of power that are encoded into a wide range of social structures. The precise nature of the barrier presented by these structures of power will change depending on the specific context where fairness is being advanced, but the generic scenario seems obvious: where increased fairness promotes the interest of a less powerful stakeholder against the interests of a more powerful stakeholder, and no pre-existing mechanism for the resolution of this difference exists, it is likely that progress will be – at the very least – stymied.

It is beyond the scope of this report to recommend how stakeholders engage with these challenging scenarios. However, we believe it is important to close the report by emphasising that strategies and methods for overcoming these barriers must be identified and implemented if the development of AI systems is to achieve its emancipatory potential.