# People Analytics must benefit the people. An ethical analysis of data-driven algorithmic systems in human resources management

Dr. Michele Loi
2 March 2020

Published by

ALGORITHM WATCH

Funded by

Hans Böckler Stiftung

# Contents

## Acknowledgements

I would like to express my deep gratitude to the following people for their contribution to this project, especially by making themselves available for the interviews that helped me frame the research questions:

Oliver Suchy, director of the department on digital workplace and workplace reporting of the German Trade Union Confederation (DGB-Bundesvorstand, Leiter der Abteilung Digitale Arbeitswelten und Arbeitsweltberichterstattung), Isabelle Schömann, Confederal Secretary, European Trade Union Confederation, Marco Bentivogli, General Secretary FIM CISL (Italian Federation of Steel and Mechanical Workers), Michele Carrus, general secretary of CGIL (Italian General Confederation of Labor) for the Sardinian Region, and Wolfgang Kowalsky, Senior Advisor, European Trade Union Confederation.

In no way should they be considered responsible for the shortcomings of the present work.

# 1. Goal and scope

The goal of this report is to analyze the implications of the emerging ethics of AI for the use of AI in the employment context. We focus on uses of AI which may affect the employee's well-being at work, employment-related opportunities, career and remuneration. The application of AI technology to human resources (HR) analytics is still in its infancy, even if one considers a generous definition of what kind of technologies AI refers to. HR analytics software products rarely involve automated decisions or even recommendations based on data-driven predictions. Rather, they often develop and visualize an array of HR metrics leaving evaluations and decisions entirely to human decision-makers. The function of these technologies is to enhance the analytical capacity of the decision-makers, by virtue of representing and packaging the information in a more usable and insightful format. These decision support systems are often labeled *descriptive analytics* (answering the question "what happens?") and *diagnostic analytics* (answering the question "why did something happen?"). While descriptive and diagnostic analytics is all but trivial, technically as well as conceptually, and ripe with ethical implications, it is not what is considered as AI for the sake of this report.

This report deals with the most socially controversial (and possibly for that reason, least developed) aspects of HR analytics, namely *predictive analytics* (answering the question "what will happen next?) and *prescriptive analytics* (answering the question "how should one act?"). The kinds of AI functionalities that interest us are primarily automatic systems of HR decisions (e.g. matchmaking) or at least recommendations for HR decisions. Typically, such activities include the profiling and scoring of employees, where profiling consists of assigning individual employees to abstract groups (e.g. productive and non-productive, reliable and unreliable) and scoring consists of assigning individuals to more fine-grained abstract groups of the same kind (e.g. all the employees that have a 0.7 reliability on a scale from 0 – not reliable – to 1 – fully reliable).

The impact of AI tools on HR management practices is important because they have the potential to profoundly influence, alter, and redirect the lives of people at work, particularly in education/training (1,2). One example is the insurance company AXA, which uses an educational/training system called the "online career assistant" to help employees seek new job opportunities within the organization, based on their abilities, aspirations, and personalities. Such matching is based on an automated analysis of the employee's CV, which generates a skill profile which is then matched with the skill profiles of positions available in the organization (3).

This analysis considers AI, which may use both business intelligence information such as "accounting and financial measures, productivity measures" (4), but also potentially information not primarily generated for tracking work-related processes, such as the data trail produced by wearable devices, e.g. if employees are invited to a fitness tracking program sponsored by the employer, or the video flow from wearable cameras, or from badges generating social sensing information (5). The hope of HR analytics is that the analysis of information may contribute to more efficient workforce planning and evidence-driven organization. Conceivably, predictive and prescriptive analytics may be applied to the same range of decisions that characterize the field of HR analytics today, for example:

- calculating the optimal number of staff members to deal with customers on the front desk;
- choosing the right kind of personality profile to deal with customers on the front desk;
- assessing the impact of health and wellness industry programs;
- measuring the ability to take initiative and using it to predict performance;
- using data when deciding personalized health and fitness programs for athletes, or for contract decisions. In the latter case, sports teams may try to predict risks inherent in hiring a promising athlete, who may become inactive due to accident or disease;
- analyzing the flow of information between team members to improve communication and problem solving by groups;
- analyzing employee satisfaction through surveys;
- collecting and analyzing key performance data, to assess personal achievements and alignment with the company's objectives;
- analyzing turnover and business opportunity to predict shortages or excesses of human capabilities before they happen;
- developing indicators that predict how likely it is that employees are to stay with the company, by identifying what employees value the most;
- optimizing a store's next day work schedule, based on predicted individual sales performance combined with other supply chain decisions (6).

AI-generated predictions and recommendations may be used to pursue all tasks currently considered in the domain of data-driven HR analytics, for example in order to personalize employment offers and contracts, manage employee's performance, optimize learning and talent development activities, manage employee engagement and communication, decide disciplinary, health and safety interventions, organize employees' holidays, absence, flexible working, maternity/paternity leave, and assign rewards (e.g. salary and benefits) (4).

# 2. Methodology

In the first step of this project, interviews were conducted with members of trade-unions[1] in order to single out significant areas of ethical concerns in the domain of human resources analytics. Insights from these interviews were combined with insights from the scientific literature on data ethics and algorithm ethics (7), not specific to HR applications. Subsequently, a philosophical framework for the analysis of the content of AI guidelines was developed, based on two sources: a) the values listed in a recently published inductive value analysis performed on a set of 84 AI guidelines (8); b) the work conducted by the author of this report in co-leading a working group to develop one such guideline (9). In this work, ethical recommendations for the development of any data-driven product were structured according to their relations to the data pipeline workflow, which starts from the data collection and ends with the deployment of a data-driven model on new data, affecting individuals. This data pipeline concept was taken into consideration as a framework to analyze the content of other guidelines. The main idea was to associate ethical requirements and recommendations with three distinct phases of the data pipeline, namely 1) data collection and generation; 2) knowledge accumulation/model building; 3) deployment of the model on actual individuals.

The last aspect of the conceptual framework was based on an in-depth analysis of the content of 20 guidelines – a subset of the 84 guidelines considered in the aforementioned global landscape review (8). The qualitative analysis of these 20 guidelines by the author which led to abstracting, by induction, different general concepts, compared to the review by Jobin and co-authors. While the analysis by Jobin, Ienca and Vayena identified the most general principles and value terms, our complementary analysis identified the most general *activity* types.

By combining trade-union concerns, different ethical values (from the analysis of Jobin and co-authors (8)), different stages of the data pipeline, and different types of activity prescribed, a new conceptual framework to analyze the content of ethical guidelines in AI emerges.

All guideline documents were retrieved using the webpage on the Algorithmwatch website. The 20 guidelines analyzed here were selected to represent different stakeholder types in a way that would be a sample of the diverse stakeholder types in the original 84 documents analyzed by Jobin, Ienca and Vayena. The selection was limited to sources available in English. Only European, international or supra-national guidelines were considered. Beside the EU High Level Expert Group on Artificial Intelligence with which we started our analysis, other 19 guidelines within the Jobin et al's 84 guideline set were selected[2] and analyzed. Table 1 contains all the guidelines

1 The following trade union representatives were interviewed: Oliver Suchy, director of the department on digital workplace and workplace reporting of the German Trade Union Confederation (DGB-Bundesvorstand, Leiter der Abteilung Digitale Arbeitswelten und Arbeitsweltberichterstattung), Isabelle Schömann, Confederal Secretary, European Trade Union Confederation, Marco Bentivogli, General Secretary FIM CISL (Italian Federation of Steel and Mechanical Workers), Michele Carrus, general secretary of CGIL (Italian General Confederation of Labor) for the Sardinian Region, and Wolfgang Kowalsky, Senior Advisor, European Trade Union Confederation.

2 We analyzed the first 19 guidelines (except those not accessible or not published in English) listed on the website *https://aiethics.herokuapp.com/*. This specific set turned out to be widely diversified in terms of: a) the stakeholder type issuing the guideline, b) the stakeholder addressed by it, c) the focus (e.g. general scope, discrimination, fairness, gender, privacy and data protection, freedom of speech, harmful AI, impact on work), and d) the type of document (e.g. short principle list, long guidelines, specialized white paper).

analyzed in the order in which they were retrieved and selected for inclusion.

Table 1

| Name of Document/ Website | Name of guide-lines/ principles | Issuer | Country of issuer | Type of issuer | Date of publishing | Target audience |
|---|---|---|---|---|---|---|
| Ethics Guidelines for Trustworthy AI (29) | Ethical Prin-ciples in the Context of AI Systems | High-Level Expert Group on Artificial Intelligence | EU | IGO/ supra-national | 8-Apr-2019 | multiple (all stakeholders) and interna-tional policy makers) |
| AI Guidelines (10) | AI Guidelines | Deutsche Telekom | Germany | Company | 11-May-2018 | Self |
| 10 Principles of responsible AI (11) | Summary of our proposed Recommenda-tions | Women leading in AI | n.a. | Think Tank | n.a. | public sector (national and international policy makers) |
| Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (12) | Principles for Accountable Algorithms | Fairness, Accountability, and Transpar-ency in Machine Learning (FATML) | n.a. | Community of researchers and practitioners | 24-Nov-2016 | Multiple (development and product managers) |
| Tenets (13) | Tenets | Partnership on AI | n.a. | Private sector alliance | 29-Sep-2016 | Self |
| Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, version 2 (14) | Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, version 2 | Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems | international | Prof. Association/ Society | 12-Dec- 2017 [3] | unspecified |
| Universal Guidelines for Artificial Intelligence (15) | Universal Guidelines for Artificial Intelligence | The Public Voice | international | Mixed (collation of NGOs, ICOs, etc.) | 23-Oct- 2018 | multiple (institutions, governments) |
| Declaration on ethics and data protection in Artificial Intelli-gence (16) | "... guiding principles ..." | ICDPPC | international | IGO/ supra-national | 23-Oct- 2018 | unspecified |

3 Oddly, the date of the press release of V2 is 12 Dec 2017 and the document itself has no date. V1 has a copyright notice dated 25 Mar 2019. Hence, V2 appears to precede V1 two years.

| Name of Document/ Website | Name of guide-lines/ principles | Issuer | Country of issuer | Type of issuer | Date of publishing | Target audience |
|---|---|---|---|---|---|---|
| Artificial intel-ligence: open questions about gender inclusion (17) | Proposals | W20 | international | IGO/ supra-national | 2-Jul-2018 | Public sector (states/ countries) |
| Charlevoix Common Vision for the Future of Artificial Intelli-gence (18) | Charlevoix Common Vision for the Future of Artificial Intelligence | Leaders of the G7 | international | IGO/ supra-national | 9-Jun-2018 | Self (gov) |
| The Toronto Declaration: Protecting the right to equality and non-dis-crimination in machine learning systems (19) | The Toronto Declaration: Protecting the right to equality and non-dis-crimination in machine learning systems | Access Now; Amnesty International | international | Miscellaneous (mixed NGO, NPO) | 16-May- 2018 | multiple (states, private sector actors) |
| Privacy and Freedom of Expression in the Age of Artifi-cial Intelligence (20) | Conclusions and Recom-mendations | Privacy Inter-national and Article 19 | international | NPO/Charity | 25-Apr-2018 | multiple (states, companies, civil society) |
| White Paper: How to Prevent Discriminatory Outcomes in Machine Learning (21) | Executive summary | World Economic Forum, Global Future Council on Human Rights 2016-2018 | international | NPO/Charity | 12-Mar-2018 | Private sector (companies) |
| The Malicious use of Artificial Intelligence: forecasting, prevention and mitigation (22) | Four High-Level Recommenda-tions | Future of Humanity Recommenda-tions Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Elec-tronic Frontier Foundation; OpenAI | international | Miscellaneous (mixed academic, NPO) | 20-Feb-2018 | unspecified |

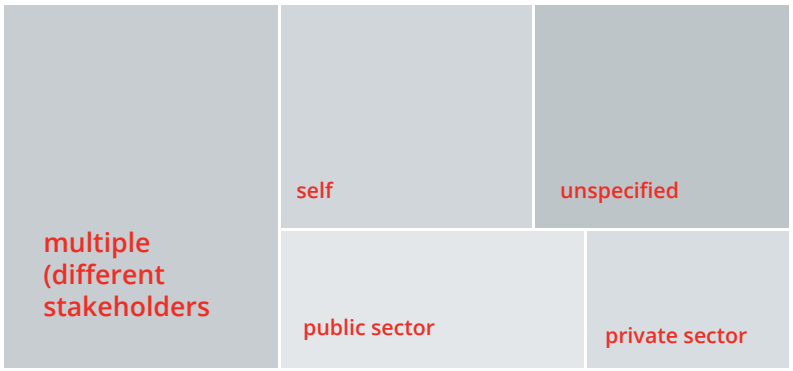
| Name of Document/ Website | Name of guide-lines/ principles | Issuer | Country of issuer | Type of issuer | Date of publishing | Target audience |
|---|---|---|---|---|---|---|
| For a meaningful Artificial Intelligence: Towards a French and European strategy (23) | "Part 5 — What are the Ethics of AI?; Part 6 — For Inclusive and Diverse Artificial Intelligence" | Mission Villani | France | Governmental agencies/ organizations | 29 Mar 2018 | Public sector (French government/ parliament) |
| Top 10 Principles for Ethical Artificial Intelligence | Top 10 Principles for Ethical Artificial Intelligence | UNI Global Union | international | Federation/ Union | 17-Dec-2017 | Multiple (union, workers) |
| ITI AI Policy Principles (25) | ITI AI Policy Principles | Information Technology Industry Council (ITI) | international | Private sector alliance | 24-Oct- 2017 | self (members) |
| Ethical Principles for Artificial Intelligence and Data Analytics (26) | Ethical Principles for Artificial Intelligence and Data Analytics | Software & Information Industry Association (SIIA), Public Policy Division | international | Private sector alliance | 15-Sep-2017 | Private sector (industry organizations) |
| Report of COMEST on Robotics Ethics (only section "Recommendations" taken into account) (27) | Relevant ethical principles and values | COMEST/ UNESCO | international | IGO/ supra-national | 14-Sep-2017 | unspecified |
| Artificial Intelligence and Machine Learning: Policy Paper (28) | Artificial Intelligence and Machine Learning: Policy Paper | Internet society | international | NPO/charity | 18-Apr-2017 | multiple (policy- makers, other stakeholders in the wider Internet ecosystem |

Table 1. 20 guidelines selected for analysis, which is a subset of Table S1 in Jobin, Ienca and Vayena, (30)a, with additional information for Issuer Type.

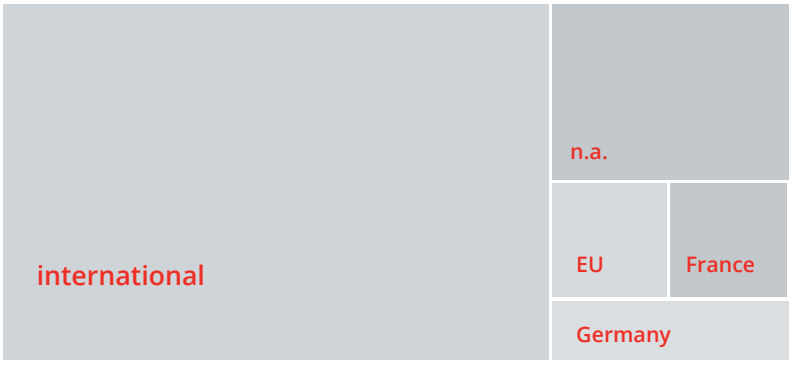
The different types of stakeholders listed in Table 1 are summarized in the charts below[4].

## Issuer


## target audience




## Issuer Country




4 NGO stands for non-governmental organization, ICO stands for information commissioner office, IGO stands for inter-governmental organization, NPO stands for non-profit organization

# 3. AI ethics guidelines

## 3.1. Towards a uniform conceptual scheme?

We now turn to the main section of this document, which is dedicated to the guidance offered by existing guidelines on ethical AI.

### 3.1.1. Values in AI ethics guidelines

Jobin, Ienca and Vayena have analyzed the global landscape of AI ethics guidelines, which includes 84 documents collected until April 23, 2019. According to their analysis, the following 11 distinct values are mentioned in the entire corpus, although no value is mentioned in *all* guidelines:

1. Transparency
2. Justice & fairness
3. Non-maleficence
4. Responsibility
5. Privacy
6. Beneficence
7. Freedom & autonomy
8. Trust
9. Sustainability
10. Dignity
11. Solidarity

But what is a value? Broadly speaking, a value is what characterizes an evaluative claim, a claim that does not simply describe the situation but assesses it for its desirability. In a narrow sense, a value is something that is good, a variety of goodness (31). Typically, value theory in the narrow sense is only concerned with intrinsic and/or final goodness (goodness as an end), not with anything that can serve as a means to promote any form of good. But many items in the above list are not "values", if by values one means *intrinsic* forms of goodness. Some of them, e.g. transparency, are more plausibly considered instrumental goods, that is, means to other objectives, which are really valuable, e.g. human well-being, or justice. Thus, I will use the term "value-laden concerns" instead of "values" to refer to the eleven items in this list, both goods that are valuable as ends and those that are valuable as means, as well as others (e.g. human rights), which appear in these guidelines.

### 3.1.2. Similar content, different scaffolding

It should be noted that there is no one-to-one correspondence between principles, values, or moral goals and the recommendations that are mentioned as practical ways of realizing those values. In other words, although we find a significant overlap of both high-level values and bottom-level recommendations across different guidelines, the mapping of practical recommendations into broad value-laden goals (or principles) is not coherent across the different documents. In other words, the guidelines mentioning the same value-laden words and similar recommendations do not share the same conceptual scheme.

To illustrate the diversity of conceptual schemes, consider the issue of algorithmic explainability. This is described as part of a right/duty to (for example):

- Transparency (15)
- Control ("We set the framework"(10), "Human in command approach" (24))
- Understanding (21)
- Explainability (12)
- Accountability (14)
- Interpretability (25)

What can we learn from the examination of this case study about explainability? Firstly, there is little theoretical coherence in the way value terms are used. These terms are never defined clearly or exhaustively and the logical inference from top down, value-laden goals or principles to recommendations is not guided by any coherent underlying theoretical construct. Rather these important value-laden words work as labels under which several practical ideas can be filed, in a way that is not entirely arbitrary, but significantly so.

As a consequence, it is legitimate to doubt that a structure relying solely or mainly on value-laden goals (or principles) will guide the users interested in the implementation of ethics guidelines to easily find the guidelines relevant to their tasks.

### 3.1.3. Different stakeholder groups

Moreover, guidelines differ in relation to the stakeholders that produced them and that they are addressed to. Jobin, Ienca and Vayena provide a breakdown of 84 guidelines in terms of stakeholders, showing that the majority are produced by private companies, followed by governmental agencies. In terms of the intended audiences, most guidelines examined by Jobin, Ienca and Vayena address multiple stakeholder groups, followed by guidelines that are self-directed (i.e. written by an organization to address itself or its members). Following Jobin and co-authors, the stakeholders in the target audience groups are the most varied: 1) the issuing organization managers and employees, including developers and designers), 2) developers and designers in general, 3) researchers, 4) the private sector in general, 5) the public sector in general, 6) generic "organizations", 7) everyone who can affect the development of AIs (8). The more limited sample on which the analysis of this document rests also includes private companies, governmental agencies, NGOs, researchers, private sector associations, professional organizations and entities, and address stakeholders as heterogeneous as the broadest set analyzed by Jobin, Ienca, and Vayena.

### 3.1.4. Different kinds of activities

As the guidelines are not coherently organized in terms of value-laden concerns (the so-called "values" in Jobin, Ienca and Vayena's paper), this report proposes a complementary conceptual scheme to analyze their contents, based on *types of activities*:

1. Knowledge and control of goals, processes, and outcomes
2. Transparency about goals, processes and outcomes
3. Accountability of goals, processes and outcomes
4. Outcome and process improvements

The three types of activities are not independent but related as follows:

a) Knowledge and control are both related to the activity of documenting i) what an organization tries to do (its goal), ii) how it does what it aims to do (its processes), and iii) what results from it (its outcomes). The act of documenting produces the human goods of knowledge and control. These goods are valuable *even before, and independently of, enabling transparency to the outside*

*and accountability*. Such knowledge and control are *presupposed* by transparency and accountability, but may be produced independently by internal processes that are neither transparent to outsiders nor associated with clear legal responsibility and moral blameworthiness.[5]

b) Transparency is achieved by combining knowledge with successful communication to the outside. Transparency *presupposes* knowledge: in order to be transparent about a goal, process, or outcome, you must first of all know it and document it. This explains why many other guidelines list activities of outcome and process evaluation and documentation under transparency.[6]

c) Accountability tasks result from the combination of the good of control with the ascription of i) moral or legal responsibility to organizations, and ii) organizational responsibility to individuals within organizations. Lack of accountability can be due to the lack of control (e.g. a failure of knowing and controlling what one does, and how one's outcomes have been achieved). But not necessarily: it can also be due to the lack of clear organizational responsibilities for the quality of the goals, processes or outcomes of an organization. Many recommendations concerning accountability describe social, administrative, political or legal tasks that enable or facilitate identifying who should be held responsible, morally or legally, for setting goals, supervising processes, monitoring outcomes, or improving them. Other recommendations concern the *technical* presupposition of responsibility, namely the *scientific knowledge* and the *techniques* enhancing a data-scientist's control of data pipeline, including the data, the training, and its outcome, namely the algorithm (or decision rule, as we shall later call it). In some cases it is hard to achieve meaningful knowledge of the algorithm, required for meaningful control and human responsibility (32). This challenge can be intrinsic to the technology – this is the issue of algorithmic opacity and black boxes, which will be considered later on.

d) Outcome improvement is the process whereby processes are modified to change outcomes in a desirable way. Improvement takes advantage of the knowledge generated and it requires some degree of control over goals and processes. For example, if a mathematical measure of bias or unfairness is provided, it may be used to constrain the utility function of an algorithm (goal control, in value-by-design approaches) and thus to improve "by design" some fairness-relevant properties of the algorithm. Improvement is a vague term and can be obtained in the direction of any of the substantive values referred above. For example, improving outcome can refer to building beneficial AI (beneficence), to building safe AI (AI non-maleficence), to building non-discriminatory AI (justice), and to preserve meaningful human control in interactions with AI, not just as a means to building safe and robust AI, but because human control is valuable *for its own sake* (autonomy)[7].

Summing up with an example: Measuring algorithmic fairness promotes *knowledge*, communicating the algorithmic fairness that has been measured promotes *transparency*, having identified modifiable causes of unfairness (e.g. data, definition of the mathematical function) achieves *control*, taking legal responsibility for those processes is a form of *accountability*, and mitigating unfairness in the outcomes is an *outcome improvement*.

---

5 Activities of knowledge (generation) and control tend to satisfy two of the requirements for trustworthy AI mentioned in the document of the independent expert group of the EU (29), namely "human agency and oversight" and "technical robustness and safety", with respect to the "robustness" part. While human agency and oversight (as requirements) and technical robustness are instrumentally valuable in relation to different ethical values and principles, promoting safety can be considered an aspect of the principle of non-maleficence, as suggested by Jobin and co-authors (30).

6 Transparency is also listed as one of the so-called "requirements" of trustworthy AI (29).

7 Activities of improvement correspond to the trustworthy AI (29) requirements of "societal and environmental well-being" (related to the substantive value of beneficence), of "diversity, non-discrimination and fairness" (related to the substantive value of justice), and of "privacy and data governance" (related to the substantive values of non-maleficence and autonomy).

These macroscopic action types can be further specified in relation to the tasks of data scientists and people responsible for implementing technology in the HR department of organizations. Consider the action type of knowledge production and control. Following a simplified, four-stage version of standard models of data science pipeline (9), we may distinguish:

- 1.1. Knowledge and control of the data (corresponding to the two stages of data acquisition and data management)
- 1.2. Knowledge and control of the algorithm (corresponding to the stage of algorithm design and testing phase)
- 1.3. Knowledge and control of the impact on humans (corresponding to the stage of deployment of the algorithm in concrete cases).

### 3.1.5. A matrix of values and action types

Some of the values identified in the analysis by Jobin et al are instrumental, procedural values, while others are substantive and intrinsic. Values such as transparency and accountability are procedural since they describe a certain way of doing things and they are instrumental since transparency and accountability are normally valued because they lead to morally better outcomes and better actions. Other values, at least justice and fairness, non-maleficence, privacy, beneficence, freedom and autonomy, dignity and solidarity are intrinsic in the sense that they are commonly regarded to characterize outcomes and actions that are intrinsically better in that they realize these values. (E.g. it is intrinsically better to be freer and have more autonomy compared to be less free and less autonomous; a society may be believed to be intrinsically superior to another if it is more just and if it involves stronger relations of fraternity and solidarity between its members.) Thus, we can summarize the two kinds of values in question and combine them with the distinction of the action types, leading to a tri-dimensional matrix including both procedural (task-related) and substantive (outcome-related) values, which relate to data, algorithms and affected humans.

Table 2. Matrix of recommendations in AI guidelines, based on procedural and substantive values

| Procedural-instrumental principles | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Outcome improvement along four dimensions (substantive ethical principles/values):[8] | | | Knowledge and Control (Document your chosen goals/procedures/achieved outcomes, concerning:) | | | Transparency (Communicate what you have documented concerning:) | | | Accountability (Define who is morally or legally responsible for documented goals, processes and outcomes concerning:) | | |
| V1. Beneficence (related to the values of well-being/sustainability/ trust) | | | Data | Algo-rithm | Humans affected | Data | Algo-rithm | Humans affected | Data | Algo-rithm | Humans affected |
| Data | Algorithm | Humans affected | | | | | | | | | |
| V2. Non-maleficence (related to the values of security/well-being/ privacy) | | | Data | Algo-rithm | Humans affected | Data | Algo-rithm | Humans affected | Data | Algo-rithm | Humans affected |
| Data | Algorithm | Humans affected | | | | | | | | | |
| V3. Justice/fairness (related to the values of solidarity/fundamental rights) | | | Data | Algo-rithm | Humans affected | Data | Algo-rithm | Humans affected | Data | Algo-rithm | Humans affected |
| Data | Algorithm | Humans affected | | | | | | | | | |
| V4. Autonomy (related to the values of freedom/dignity/funda-mental rights/privacy) | | | Data | Algo-rithm | Humans affected | Data | Algo-rithm | Humans affected | Data | Algo-rithm | Humans affected |
| Data | Algorithm | Humans affected | | | | | | | | | |

Most recommendations in these guideline can be described as a task or combination of tasks involving some activity of achieving control, and/or documenting and/or communicating and/or taking ownership for and/or improving processes, for the sake of bringing about improvements in one or more dimensions, typically more dimensions (e.g. producing good, minimizing harm, and mitigating injustice) simultaneously. The four main value dimensions corresponding to the four principles of biomedical ethics (33).[9] We have expanded this list to include the other intrinsic values in Jobin, Ienca and Vayena's list (8), which can be regarded as more closely related to them. Hence, beneficence, which involves doing good, is related to promoting well-being, and sustainability is also typically related to the possibility of promoting well-being in the future; non-maleficence is related to the values of security, harm avoidance, and some aspects of privacy; justice, fairness and solidarity are related to discrimination, and more generally to inequality in the distribution of the benefits and social inclusion vs. exclusion of the benefits produced by AI; freedom,

8 The values of autonomy, non-maleficence, and justice correspond to the ethical values mentioned in the EU ethics guidelines for trustworthy AI (29): (i) Respect for human autonomy, (ii) Prevention of harm and (iii) Fairness. In addition to these, the EU guidelines include the value explicability. For the reason why explicability is not included as a distinct value in our framework, see the footnote below.

9 Floridi and Cowls's unified framework for AI ethics includes the four ethical values of our framework, plus the value of explicability. We think that the value of explicability is both instrumental and at a different level of abstraction, being one of the main presuppositions of the values of both accountability and transparency and being close to the instrumental values of knowledge and control.

autonomy and dignity are related to some aspects of privacy (e.g. control over one's data and information) but also the idea that human beings should remain in control of their lives is something valuable for its own sake, as opposed to being manipulated or controlled from the outside.

### 3.1.6. Different implementation infrastructure

Finally, the tasks recommended in response to value-laden concerns can be distinguished in terms of the kind of social roles and organizational structure they presuppose, in order to be executed. It is fruitful to distinguish three main kinds of solutions, which differ in terms of the social roles and tasks they presuppose:

- I1: **Technical solutions.** These presuppose mainly technical tools to be executed. For example, the FAT ML guideline on accuracy recommends "Perform a validity check by randomly sampling a portion of your data (e.g., input and/or training data) and manually checking its correctness. This check should be performed early in your development process before derived information is used. Report the overall data error rate on this random sample publicly". This is a technical task that only requires the usual social role of the data-scientist, and the technical infrastructure typically available to the machine learning specialist working at an organization.[10]

- I2: **Organizational solutions.** These presuppose an infrastructure of rules (and rule-governed behaviors) which can be set up within a single organization. For example, the second edition of the IEEE guidelines (14) requires that "Governance frameworks, including standards and regulatory bodies, should be established to oversee processes assuring that the use of A/IS does not infringe upon human rights, freedoms, dignity, and privacy, and of traceability to contribute to the building of public trust in A/IS".[11]

- I3: **Institutional solutions.** These presuppose an infrastructure of rules (and rule-governed behaviors) that cannot be set up within a single organization. This includes, for example, recommendations concerning new laws, policy objectives, and the promotion of new civil society bodies or stakeholders. For example, the French report on AI "For a meaningful Artificial Intelligence: Towards a French and European strategy" (23) includes the following recommendation concerning gender equality: "Educational efforts on equality and digital technology are obviously vital, but greater diversity could also be achieved with an incentive policy aimed at achieving 40% of female students in digital subject areas in universities, business schools and their preparatory classes out to 2020."[12]

10 The ethics guidelines for Trustworthy AI (29) contain a similar distinction between the "methods" for achieving trustworthy AI. The solutions mentioned in this paragraph correspond to the "methods" called "Architectures for Trustworthy AI", "Ethics and rule of law by design (X-by-design)", "Explanation methods", "Testing and validating",.

11 In relation to the Trustworthy AI guidelines (29), organizational solutions include "Quality of Service Indicators", "Codes of Conduct", "Standardization", "Certification", "Accountability via governance frameworks", "Education and awareness to foster an ethical mindset" (within the organization), "Stakeholder participation and social dialogue" and "Diversity and inclusive design teams".

12 In the trustworthy AI guidelines (29), these correspond primarily to "methods of trustworthy AI" called "regulation". It also includes the methods of "Accountability via governance frameworks", "Education and awareness to foster an ethical mind-set", and "Stakeholder participation and social dialogue".

# 4. Implications of AI ethics guidelines for HR analytics

The recommendations included in twenty documents have been analyzed here in order to assess their relevance to the ethics of using AI in HR. The structure of the document reflects the main distinction between three different topics, namely the activities of data collection, of building an HR tool (algorithm) and of using this HR tool to aid decisions in the field of HR. For each topic, the four *procedural values* of i) knowledge/control, ii) communication/transparency, iii) accountability/"owning" the process or outcome are considered sequentially.

## 4.1. Knowing, communicating, owning and improving data collection, access and storage

The sub-topic "knowledge and control of the data" includes all the recommendations related to *knowing, documenting* and *monitoring* the processes of data acquisition and generation, data storage and data access, in particular when data from identifiable natural persons (i.e. personal data) are involved.[13] The generation of knowledge about the data and the process of data collection is presupposed by recommendations about transparency and accountability. Data can be related to AI in different ways: they can be the data used for training a statistical model, or the data based on which AI makes a recommendation or decision about a concrete individual case. Clearly, the processes of data collection and access must be described and documented by any organization that wants to be transparent and accountable about it. For example, the transparency principle of IEEE (principle 4 in the 2nd edition) requires securing and recording data from sensors (which are used by the AI), such as in a flight data recorder. But the detailed knowledge and control of all processes of data collection, access and storage is valuable for protecting privacy (e.g. from unauthorized access), and for enhancing the robustness and reliability of AI, even when these processes are not communicated to outsiders and even independently of legal responsibilities. Thus, it is not surprising that many recommendations require the documentation of the data that is collected, and used by algorithms, also in connection to other values, for example privacy.

Several recommendations include prescriptions or checklist items requiring that data collection, access, and storage, is to be used with AIs, and always done in a way that is controlled, transparent, and accountable:

> *"Is the data collected in an authorized manner? If from third parties, can they attest to the authorized collection of the data? Has consumer been informed of data collection, and where appropriate, provided permission about its use?"*(26)

> *AI systems must be data responsible. They should use only what they need and delete it when it is no longer needed ("data minimization"). They should encrypt data in transit and at rest, and restrict access to authorized persons ("access control"). AI systems should only collect, use, share and store data in accordance with privacy and personal data laws and best practices.*(28)

> *We enrich and simplify our customers' lives. If an AI system or the usage of customer-related data helps us to benefit our customers, we embrace this*

13 This sub-topic correspond to the requirement "privacy and data governance" of trustworthy AI (29).

> *opportunity to meet their demands and expectations.* (10)

In the texts analyzed here, these activities are associated with transparency (14), respect for privacy and personal data (16), and accountability. Transparency, however, should not be considered achieved when the processes of data collection, access and storage have been documented.[14] Transparency, in the sense we understand the value here, essentially involves communicating these facts in a simple and effective way to the interested parties. Commitments to data transparency are included in Deutsche Telekom's "AI guidelines" (Principle 4)(10), in the "Declaration on ethics and data protection in Artificial Intelligence" (16)(Principle 5) explicitly mentioning the right to information and the right to access.

The principle of accountability requires measures whereby an organization, or a person, or role within an organization, takes responsibility or "ownership" for a process involving data. One important concept is that of *source traceability*. When an AI takes a decision for an organization, the data used to make that decision should be known and usable by some person – that is, an auditor internal or external to the organization – to explain the decision of the AI. The idea of source traceability is expressed in different ways in different guidelines. Some guidelines include concrete suggestions, for example, the recommendation for secure storage of sensor data (14), also described as the "ethical black box" (24):

> *Applied to robots, the ethical black box would record all decisions, its basis for decision-making, movements, and sensory data for its robot host [...]* (24)

Like data protection law (e.g. the General Data Protection Regulation – GDPR – in the EU), ethical guidelines stress the importance of guaranteeing appropriate levels of *cybersecurity*:

> *We ensure that our security measures are up to date while having a full overview of how customer related data is used and who has access to which kind of data. We never process privacy-relevant data without legal permission. [...] Additionally, we limit the usage to appropriate use cases and thoroughly secure our systems to obstruct external access and ensure data privacy.* (10)

### 4.1.1. Relevance to HR analytics

It is not surprising that the topic of data control, transparency and accountability is not one of the most widely discussed in guidelines on AI ethics. After all, this ethical territory overlaps strongly with privacy and data protection law, so it has the lowest novelty value and it is the one least needing to be codified by new ethics guidelines. Issues related to data protection and privacy protection were perceived as essential, by the trade unionists who were interviewed: a sort of necessary condition for any other ethical guideline to apply. The importance of these recommendations for HR analytics can be explained as follows:

1. Recording data from sensors involved in employee monitoring would be instrumental to assessing causes of counter-intuitive HR decisions, by models trained with the data. (Relevant for V1 – trust and V3 – justice). However, if such data are not anonymized (which may not be feasible in many circumstances) this generates further threats to employee privacy and increases the risks associated with their surveillance, hence the importance of guidelines on data privacy. (Relevant for V2 – non-maleficence.)

2. Requiring the employees' *informed consent* to the use of their data may be appropriate, at least in some context. (V4 – freedom)

14 The first edition of the IEEE's guidelines also included the principle of "data agency" that interprets digital sovereignty and control over identity as control over data, including through digital agents (14). Interestingly, these ideas were entirely removed from the second edition. Such an idea was not found in the other 19 guidelines reviewed here.

3. In addition to *informed consent* and also in cases in which a different legal ground for collecting employee's data is invoked, the principle of data minimization could be invoked in attempts to limit the range of data not pertinent to employment. This will place some limits on the employers' invasion of an employee's private life. (V4 – freedom/privacy)

4. Effective communication about how data is used within a company may also be used to confer some protections to employees against abuses, e.g. if documented data processes are made accessible to workers' representatives, or external auditors. From the point of view of individual employees, transparency about data use is a precondition of informed consent to their use. (V1 – trust, V2 – non-maleficence, V4 – freedom)

5. Transparency about the building of profiles contributes to protecting employees against the collection of data that may be unreliable or incorrect, or used for illegitimate forms of discrimination. (V2 non-maleficence, V4 freedom)

6. Accountability for data processes provides incentives against unethical uses of data. (Relevant to V1, V2, V3, V4.)

### 4.1.2. Open challenges

On the other hand, the focus on data knowledge, transparency, control, accountability, and outcome improvement is insufficient to protect employees from threats to their privacy and from morally objectionable forms of discrimination. It is true that employers are prohibited (by some privacy laws) from requesting access to social media data. But, depending on the content of the laws in the country of operation of the employer, the principle of purpose limitation may be compatible with employers collecting data from different sources. Consider data from social media. E.g. an employee's public tweets may be collected by one employer to be further analyzed. Since machine learning algorithms are used to discover new and surprising correlations, an employer may claim, for example, that the use of social media may provide a legitimate basis to evaluate and predict the reputational risk, for example, associated with the social media exposure of its employees. (E.g. an organization dealing with migrants may be concerned that its employees do not publicly express xenophobic views on social media.) What data protection law does not provide is the definition of this boundary, i.e. what kind of data may be legitimately collected by employers to be analyzed for the sake of HR decisions? For example, should an NGO working with migrants adopt a tool that predicts the likelihood of its employees publicly uttering xenophobic or discriminatory views, based on past social media interactions?

Even when limits are set regarding the type of employee data that an employer may legitimately collect and analyze, this is not yet sufficient to protect employees from an invasion of privacy. The challenge to an approach that is focused on the *type of data* that an employer should consider, or not consider, is that it does not protect the privacy of employees from (somewhat inaccurate) inferences that can be made from apparently innocent data (34). The possibility of using statistical methods (e.g. machine learning) brings the possibility that data apparently not about gender, sexual inclination, social class or family status be inferred, with a given degree of uncertainty, from other data an employer may more easily collect (34). Machine learning models may enable an employer to assess, with a given degree of uncertainty, the lifestyle choices of his employee (e.g. whether an employee is planning to have a child) based on data originating in the context of employment, e.g. an employee's schedule, special requests, etc. Privacy is not guaranteed by controlling *what kind of data* one shares. Privacy – if it means freedom from the influence of the employer within a sphere of personal decisions – is threatened by AI even if no sensitive data are shared, if AI provides a technology that makes informed guesses about sensitive characteristics from data that are legitimately collected in the workplace.

## 4.2. Knowing, communicating, owning and improving the algorithm

Most guidelines analyzed here concern the nature of the algorithms themselves rather than the process of data collection. In the context of HR analytics, by "algorithm", or "AI", the guidelines can mean two distinct sets of rules implemented electronically: the learning algorithm, which can be, for example, a process used to infer a general rule based on historical data; and the (algorithmic) *decision rule* that is a rule which is applied to concrete individuals and, after processing data about those concrete individuals through some kind of model, generates a prediction, recommendation or decision about them. The *decision rule* could in many cases also be called the "learned algorithm", as nowadays decision rules are rarely hand-coded on the basis of domain knowledge. Rather, machine learning algorithms are used to infer (learn) general rules which then generate outputs (e.g. predictions, classifications, recommendations, decisions) about particular cases. In the case of *predictive analytics* the *decision rule* only outputs a prediction about an individual and lets the human decide what to do with it. In the case of *prescriptive* analytics, the *decision rule* could be, for example, a recommendation (to give a salary raise to an employee, or to direct more work on certain tasks) which a human decides, or an automated decision (e.g. to efficiently assign holidays to workers on specific days). Some confusion arises from using "AI" to refer to both the learning algorithm and the decision rule. In a neural net algorithm, for example, the learning algorithm is some kind of mathematical rule determining how to set the weights of the nodes of the net in response to the training data. Computer scientists are entirely aware and fully understand these mathematical rules (e.g. the back-propagation algorithm, which uses calculus to minimize errors). The application of these algorithmic rules to a (for example statistical) learning process produces a *decision rule*, e.g. a neural net which recognizes pictures of cats. The intrinsic logic of the decision rule of a neural net can be an extremely complex criterion – it can only be described as the computation resulting from the interaction of individual neurons combined and weighted in a specific way, which is not determined a-priori but established by a learning algorithm. When one says that a neural net decides whether an image is a cat, here "neural net" means the learned *decision rule* (not the learning algorithm!). The *decision rule* has a given performance, such as accuracy, e.g. it classifies cats correctly 90% of the time. The generation of knowledge about the algorithm and of documenting it may refer to both the learning algorithm and the resulting decision rule, whose performance can be assessed in the lab, for example, by running tests with known data (data about individual cases, whose labels, e.g. being a cat or a dog, are known in advance to the data scientist).

Some recommendations require organizations that produce or deploy AIs to document the learning process, including the learning algorithm (e.g. the type of algorithm and its human-defined parameters) and the data used to train it and test it. Other guidelines are best understood as requirements to produce knowledge about the decision rule, for example those requiring producers or deployers of AI systems to assess the accuracy (12,29), reliability and reproducibility (29) of AIs' decisions or predictions. For example, the IEEE guidelines (14) require manufacturers, operators, and owners of AIs to register: "Intended use, Training data/training environment (if applicable), Sensors/real world data sources, Algorithms, Process graphs, Model features (at various levels), User interfaces, Actuators/outputs, Optimization goal/loss function/reward function". In the Trustworthy AI guidelines this is called "traceability" and is listed as an element of *transparency* (29), but clearly traceability is also instrumental to robustness[15] and safety. Many guidelines insist on the importance of ensuring the quality of the training and test data. For example, the World Economic Forum prescribes:

15 For example, if you can verify that a model to differentiate wolves from dogs has been trained with pictures of wolves, predominantly including snow in the surrounding, and dogs, predominantly including no snow in the surrounding, you can more easily produce valid hypotheses about the failure of the model to generalize in real-world situations.

> *Determine whether certain data sets fit internally agreed upon standards of "adequate" and "representative" data (looking to both quantitative and qualitative metrics); identify opportunities to expand data collection efforts where contextually appropriate, viable, and possible to do so without violating privacy* (21)

This idea – requiring "adequate" and "representative" data and "opportunities to expand data collection efforts" – is found in different guidelines.[16] One guideline even includes the recommendation to "Exclude data that is not relevant to predicting the outcome"(26) where it is unclear if such "relevance" should be assessed *in advance* of the algorithmic learning process, or *a posteriori*.[17]

Besides the data used in training, another element of the data pipeline that, according to the guidelines, should be documented is the training process, i.e. the algorithm used and the parameters used to set it up. The IEEE guidelines require manufacturers, operators and owners of AIs to register the "algorithms" used to generate a model, and more specifically their "[m]odel features (at various levels)" and the "optimization goal/loss function/reward function" (14). Similarly, the guidelines of the Software & Information Industry Association ask software developers to produce "[a]n inventory and documentation of all models used for automated decisions" (26) and to declare "what the model is intended to predict" (26). Other guidelines require model building to be sensitive to "the norms and values of specific populations affected by the output of AI systems" (21) and to generate and publicize information about algorithmic performance. The IEEE document (14) recommends organizations to assess the *transparency* of the AIs in question through rigorous metrics. The Pilot assessment list of the Trustworthy AI guidelines requires designers to declare the accuracy goals, to assess that adequacy is accurate, to improve accuracy (under the assessment of "accuracy"), to communicate "the reasons and criteria behind the AI system's outcomes", "the purpose of the AI system and who or what may benefit from the product/service" and to communicate its "characteristics, limitations and potential shortcomings" (29).

As already mentioned, some recommendations require manufacturers, operators and owners of AIs to define and record the goals behind the algorithms (e.g. what it should predict) and test their performance. These are standard steps of the workflow of data science, but arguably they become ethically salient in combination with three additional requirements:

1. The requirement that information about these steps is used to enhance the *transparency* of the system, ideally in the sense that it can be made available to independent auditors, when the AI's ethical credentials should be questioned. (Relevant for V1 Beneficence/Trust, and V2 non-maleficence as robust and trustworthy AI is likely to produce more benefits and to be more secure). Indeed, in several guidelines documenting the algorithm is considered an element of transparency and adequate communication.[18]

16 For example, the "Pilot assessment list" of the guidelines for trustworthy AI (29) includes the following questions: "Did you put in place measures to ensure that the data used is comprehensive and up to date? Did you put in place measures to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?"

17 One problem with this proposal is that many algorithms of machine learning can be considered statistical methods to determine the relevance of data for a given prediction task. So, the relevance of information cannot be assessed in advance of including the data in the data pipeline of machine learning. The guideline appears redundant, if it is meant to imply that a decision rule (identified via a machine learning algorithm) should not make predictions based on data that is discovered a posteriori (that is, through machine learning) to not contribute to the accuracy of the model. Understood in this way, it will be the machine learning algorithm itself that determines if any source information is worthy of collection and analysis by a decision rule. But such evaluation is a posteriori which means that R&D departments are justified in collecting any kind of information to determine if a model can be learned from it. Alternatively, the claim may be interpreted as an attempt to use the domain expert's prior knowledge of the factors that can increase the accuracy of predictions as a criterion to decide which data to exclude from the machine learning pipeline, a priori. If so, it is a very conservative requirement, since it is one of the goals of machine learning and big data to discover new, and even non-intuitive correlations, which do not belong to the consolidated knowledge of domain experts.

18 E.g. the communication of the algorithm's goal, criteria, and limitations, is listed under "explainability" in the assessment list of the guidelines for trustworthy AI (29).

2. The requirement that the goals of the algorithm be defined in a way that is sensitive (or at least not insensitive) to the values and norms of the stakeholder. Ideally this is meant to avoid the disconnect between actual practice and automation associated with products that are developed by engineers with little knowledge of the real-world, lived existence of the people impacted by their products. (Relevant for V1 and V2.)

3. The requirements that performance assessment be made in a way that is both scientifically and ethically defensible. Some guidelines explicitly mention the accuracy (29,35) of AI or machine learning models. But the apparent accuracy of a model may be a smokescreen for the fact that the model is highly inaccurate when applied to a different population, from the one the data come from (an instance of "overfitting"). Or the way the accuracy of AI is measured may have little "ecological validity", i.e. be a poor indicator of how the AI behaves in real-world settings, including in the altered circumstances in which people know their behaviors are being measured and evaluated. This is why some guidelines mention reliability and reproducibility (29),[19] in addition to accuracy. Alternatively, one may speak about *robustness* and *veracity* (adherence to facts and reality). These values may be compromised due to a poor choice of training data (hence the emphasis on data quality in many guidelines) or from the inability of the designers to foresee and assess additional environmental variables influencing the model's performance, associated with the real-world deployment of AI. (Relevant for V1, V2 and V3 – fairness.)

Such recommendations appear in other guidelines as implementation methods for the principles of respect for human rights (14), transparency (14,19), the goal of value-sensitive design [(27–29)], the principle of "accountability" (14), "active inclusion"(21), control ("We set the framework"(10)), "accuracy"(12), transparency and communication (29).

### 4.2.1. Improving fairness

Another widely addressed topic concerns discrimination by an algorithm, or more or less equivalently, its fairness. More precisely, the algorithm whose fairness or discriminatory properties are in question is what we have called the *decision rule*.[20] Two distinct aspects of this problem are highlighted in most codes, with a minority of codes that emphasize both. One aspect is that the idea could be labelled as fairness, equal robustness or veracity. The idea here is that one needs "robust and representative data" to be both fair and (equally) accurate when making predictions and decisions about all groups (25). Proper datasets are often mentioned as a precondition of fair and non-discriminatory algorithms, in particular the idea that data ought to be representative of the different groups found in society, including minorities. Other documents only talk about avoiding bias or biased data in general terms, without using the language of discrimination (which typically refers to social groups) explicitly[21]. The boundary between the question of accuracy, reliability, and that of fairness is not very well-defined. In some cases, an unfair or discriminatory algorithm is one that "fails in certain kinds of setting"[22], namely in a setting in which minority groups (e.g. black women) are involved. Other documents are far more specific in their referring fairness and equal treatment to groups, e.g. groups

19 For example, the assessment list of the Trustworthy AI guidelines (29) includes the following questions, under the heading of "reliability and reproducibility": "Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility? Did you put in place verification methods to measure and ensure different aspects of the system's reliability and reproducibility? Did you put in place processes to describe when an AI system fails in certain types of settings? Did you clearly document and operationalize these processes for the testing and verification of the reliability of AI systems? Did you establish mechanisms of communication to assure (end-)users of the system's reliability?"

20 This corresponds to the requirement of "diversity, non-discrimination, and fairness" in the guidelines for Trustworthy AI (29), in particular the "avoidance of unfair bias".

21 For example, the recommendation that "AI systems should not be trained with data that is biased, inaccurate, incomplete or misleading." (28).

22 This is a question of the assessment list of "reliability and reproducibility" in the Trustworthy AI guidelines (29).

defined by gender or race (36–38), as it is typical of most anti-discrimination laws. For example, the document of Women 20 (the group responsible for digital inclusion of the World Wide Web Foundation) recommends governments to produce "open gender disaggregated datasets, so the machine learning systems can improve their performance", with an explicit reference to gender (17). With respect to the training data, the guidelines for Trustworthy AI include the following checklist items as methods of "unfair bias avoidance" (29):

- *Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets?*
- *Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?*
- *Did you research and use available technical tools to improve your understanding of the data, model and performance?* (29).[23]

Some guidelines mention the value of fairness explicitly (e.g. "*unfair* bias avoidance" (29) as opposed to "bias avoidance") and stress that defining what fairness means for a decision rule may be difficult and highly context-dependent. For example, the guidelines on "how to prevent discriminatory outcomes" by the World Economic Forum explicitly recommend that

> *People involved in conceptualizing, developing, and implementing machine learning systems should consider which definition of fairness best applies to their context and application, and prioritize it in the architecture of the machine learning system and its evaluation metrics* (21)

Besides acknowledging the plurality of fairness definitions, this and other prescriptions highlight another important set of ideas concerning fairness-by-design: fairness should be measurable by technical methods as an aspect of algorithmic performance (e.g. along with accuracy and other measures) and imposed as a goal in the data pipeline process.[24] The appropriate balance of fairness, privacy and accuracy is an aspect of algorithmic design, that is implemented technically. Expressions referring to the technical implementation of fairness are, for example, "fairness-aware data mining algorithms"(12), or "values which may need to be embedded in the machine" (11).[25] The FAT-ML principles include further implementation suggestions, such as calculating error rates as well as types (i.e. it is necessary to distinguish the prevalence of false positives from false negatives) for different sub-populations.[26] It also talks explicitly about "disparate impact" which is the US-law expression for indirect discrimination.

The vast majority of references to fairness are related to the avoidance of discrimination which is typically defined as unequal or unfair treatment in relation to *social groups*. The FAT-ML principles require developing awareness of inequalities in access to goods for *groups* defined by "race, sex, gender identity, ability status, socio-economic status, education level, religion, country of origin" (12). (This is different from requiring the removal of all inequalities, which would be problematic, as we shall later explain.) Some documents include the more abstract concept of "groups that may be advantaged or disadvantaged *by* the algorithm"(21). Some guidelines identify the relevant groups by reference to past and existing social

23 Analogously, the ICDPPC document recommends taking reasonable steps to ensure the personal data and information used in automated decision-making is accurate, up-to-date and as complete as possible" in relation to "unlawful biases or discriminations(16).

24 Analogously, the Trustworthy AI guidelines' assessment list includes the following item: "Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems? Is your definition commonly used? Did you consider other definitions before choosing this one? Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?".

25 Analogously, the Trustworthy AI guidelines' assessment list includes the following items: "Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system? [...] Did you establish mechanisms to ensure fairness in your AI systems? [...]" (29).

26 The same requirement appears in other documents, e.g. (21).

injustices and, more broadly, inequalities. A reference to historical inequality and injustice is the prescription to avoid "self-fulfilling markers of success and reinforce patterns of inequality"(19), and the warning that "[e]xisting patterns of structural discrimination may be reproduced and aggravated" (19).

Summing up, guidelines that devote significant attention to the issue of fairness include the following recommendations:

Document bias, discrimination, unfairness, and the selected fairness goals

1. Acquire knowledge of norms related to the context of the deployment of an algorithm, and related to the cultural context in general, including social and legal norms (14) and of the different possible definitions of fairness which may be adopted in relation to these specificities (21,29). (Relevant for V1 trust and V3 justice.)

2. Consult domain experts and take into consideration interdisciplinary insights to understand potential biases and unfairness (21) (Relevant for V1, V2, V3, and V4.)

3. Detect unequal error rates by population, considering different kinds of errors (12,21). (Relevant for V3 fairness.)

4. Detect indirectly discriminatory effects/disparate impact (when groups benefit in different degrees from algorithmic decisions (12,21,26)), e.g. by documenting disparate impact for individuals of a class that deserves special protection. It is acknowledged that disparate impact cannot be eliminated entirely and so its avoidance should be "proportional and necessary considering the costs involved"(26). The SIIA document even recommends collecting sensitive information (race, gender, ethnicity, and religion) and using it in data analytic systems, "where permitted by law" for the purpose of assessing disparate impact (26).

5. Some guidelines stress the importance of measuring unfairness not only in relation to the test data (which are historical data, typically from the same source as the data used for the training) but also by monitoring the performance of the application of the algorithmic *decision rule* as applied to the business case at hand. E.g. The Toronto Declaration mentions the importance of "accurate pre-release trials, and the set-up *of an ongoing evaluation system* throughout the life cycle of the product" (19), and the Trustworthy AI guidelines require organizations employing AI to assess themselves by asking "[d]id you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system?" (29).

Improve the fairness of outcomes

6. Collect more, or more representative data, in order to reduce discrimination, or generate disaggregated datasets, including opportunities to expand data collections (17,19,21,29). (Relevant for V1 trust, V3 fairness.)

7. Do not use an algorithmic system if the same goal can be achieved with an algorithm that has a lesser disparate impact (26). This recommendation seems to be derived from the US disparate impact law, which, however, applies to a limited sphere of decisions (those affecting employment and housing). The SIIA's document even posits an ethical obligation to re-design a system with less disparate impact on vulnerable groups, even if this sacrifices "organizational effectiveness" when the cost-benefits balance of such step is defensible (26). (Relevant for V3 fairness.)

8. Other documents invoking discrimination prevention or mitigation are very generic concerning how define unfair discrimination,

the kind of discrimination that ought to be removed.[27] Some guidelines characterize unfairness and discrimination in terms of errors and biases where "such errors or biases may disproportionately affect certain groups of people"(20). But notice that removing disparate errors is a conception of unfairness as *disparate mistreatment* (39): mitigating disparate impact (as in point 7) is another goal entirely and may even lead to *augment* the disproportion of errors across groups.

9. Gender inequality and gender bias are given a special emphasis. The document by the Global [workers'] Union features a "Genderless, unbiased AI" principle (24). One guideline can be read as suggesting affirmative/reverse discriminatory action (engineered through algorithmic design) in areas where women are disadvantaged in society, because it recommends "algorithmic equitable actions to correct real life biases and barriers that prevent women from achieving full participation and equal enjoyment of rights" (17). (Relevant for V3 justice.)

10. Introduce redress mechanisms, such as an "emergency procedure to correct unforeseen cases of unfairness" (21). (Relevant to V1 trust and V3 fairness.)

Transparency about discrimination removal/fairness promotion

1. Some guidelines require documenting the forms of unfairness and discrimination discovered and efforts made to identify, prevent and mitigate against discrimination in machine learning systems (19,21), e.g. "Introduce a new 'Certificate of Fairness for AI systems'" (11) (Relevant to V1 trust and V3 fairness.) Interestingly, one document even requires explaining "when a model is built with discrimination as a desired outcome and hold the relevant parties accountable"(21).

2. Allowing third parties to monitor, signal, and assess forms of algorithmic bias and discrimination. E.g. the FAT-ML document includes a principle of auditability(12), and the guidelines on Trustworthy AI guideline's assessment list includes a request to "ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system"(29).

Inclusive R&D teams

1. Some guidelines stress the importance of diversity in R&D development teams for AI. The guidelines for Trustworthy AI (29) even describe "[d]iversity and inclusive design teams" as a *non-technical method* for achieving trustworthy AI. The goal of inclusion may be specified in different ways. For example, the White Paper by the World Economic Forum (21) requires that organizations responsible for developing AI systems "bring different perspectives together", and "afford insights into whether certain populations are adequately included and represented in training data". Understood in this sense, the proposal goes well beyond gender inclusion or the inclusion of minority members in R&D teams as it may justify including researchers or experts with a different (i.e. non computer science-related) disciplinary profile. (Relevant to V1 trust and V3 fairness.)

2. An issue that is sometimes related to questions of fairness and discrimination is the one of countering adverse stereotypes, e.g. gender stereotypes in toy robots and sex companions (27). (Relevant to V3 fairness.)

27 E.g. The Toronto Declaration which requires that organizations "take effective action to prevent and mitigate discrimination"; the Universal Guidelines by "The Public Voice" recommend that "Institutions must ensure that AI systems do not reflect bias or make impermissible discriminatory decisions" – where the qualifier "impermissible" suggests that some discriminatory decisions may be permissible; the ICDPPC document writes that "Unlawful biases or discriminations [...] should be reduced and mitigated". Notice that discrimination "as defined by international law" (mentioned in the Toronto document) may have a narrower scope than fairness.

3. One guideline claims that stereotyping (such as, for example, mentioned in point 2 above) is caused by "the low involvement and marginal inclusion of women in the coding and design of AI and machine learning technologies" (17). Consequently, it recommends the following actions to governments:

— "[...] take proactive steps towards the inclusion of more women in the workforce that design AI systems [...]

— [...] require companies to proactively disclose the gender balance of their design teams

— [...] require recipients of research grants to disclose the gender balance of the applying research teams

— [...] ensure that decision-making spaces are adequately gender balanced [...]

— [...] fund women-owned technology firms working in AI [...]

— [...] incentivize other firms to have more diverse staff at all levels." (17).

Similar recommendations aiming for better gender balance in technical teams responsible for AI are found in other guidelines for example:

— "an incentive policy aimed at achieving 40% of female students in digital subject areas in universities, business schools and their preparatory classes out to 2020" (23) (Relevant to V3 fairness.)

### 4.2.2. Improving intelligibility

Transparency does not consist simply in documenting goals, processes and outcomes. It also consists in communicating these in a way that can be understood. One of the most widely and intensely debated issues is the degree to which algorithms can and should be understood by different stakeholders, including the wider public. Again here the subject matter of intelligibility is the *decision rule,* that is, the *learned* algorithm, not the machine learning algorithm. The *decision rule* makes predictions, recommendations or decisions about individual cases in concrete applications. One especially important aspect of being subjected to, or impacted by, decisions based on rules determined by algorithms is the possibility of understanding the ground, or reasons behind, such decision. Thus, the aspiration that algorithms (meaning here: decision rules) be comprehensible or explainable relates to an alleged moral or legal right to explanations:

> *All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome.* (15)[28]

A frequently mentioned idea is that intelligibility and explainability are relative, not absolute, properties. A good explanation for a data scientist may not be intelligible to a lay person, and an explanation that may satisfy the curiosity of a lay person may be considered obscure by a data scientist. Hence some guidelines recommend taking this audience-relativity into account, e.g.

> *Explainability Guiding Questions: Who are your end-users and stakeholders?* (12)[29]

There are fundamentally two distinct strands on explainability:

1. One strand conceives explainability in a holistic way: understanding *why* an AI does something is a matter of documenting, together, or in a combined way:

28 Analogously "Workers must also have the 'right of explanation' when AI systems are used in human resource procedures, such as recruitment, promotion or dismissal" (24).

29 Analogously, the guidelines for Trustworthy AI (29) mention different stakeholders in the assessment list, under "Communication", which is a sub-heading of "Transparency": end-users, other users (those embedding a technology in another service), third parties and the broader public.

a) The source code (24) (although this is widely considered both insufficient, and unnecessary)

b) The data sources, i.e. when and where data is collected, for both training data and test data (11,12,29)

c) The process of data cleaning or data transformation (11,12)

d) The features used to train an algorithm/make decisions (11)

e) The weightings of these features (if known) (11)

f) The algorithm type, the extent of its opacity (11,12)

g) The different performance metrics (11,12)

h) The procedure of validation (11,12)

i) The algorithm's generic goal and purpose (29), in the sense of "intended use" (14)

j) The algorithm's mathematical goals, in the sense of its "optimization goal/loss function/reward function" (14)

In more general and abstract terms, this approach to achieving intelligibility consists in communicating "the factors, the logic, and techniques" (15), "the reasons and criteria behind the AI system's outcomes" (29), and, in some cases, the "cooking book", i.e. how such logic was implemented technically. Transparency about the design goals, reasons and criteria can be achieved even when the *decision rule* itself, i.e. "the internal workflow of the model" (29) is not transparent, because it is too complex to be grasped in its entirety by the human mind (40). In other words, this kind of intelligibility consists of explaining goals and desiderata of the *decision rule*, and does not focus on individual decisions (41–44). Notice that the tasks recommended for intelligibility, understood in this way, include the tasks of documenting the model design, described above. But intelligibility is best understood as an aspect of algorithmic transparency, which involves the element of purposive *communication* (to the outside), beside (internal) documentation. If intelligibility is an aspect of *transparency*, it may not be enough for the data scientists to document the logic and implemented methodology in ways other data scientists or technically proficient auditors can understand. In fact, a requirement of *transparency* may have different target audiences, than colleagues and auditors. Understood as an aspect of *transparency, intelligibility* is also the requirement to explain the *rationale* and generalizations behind the rule in a way that is accessible to specific different target audiences[30] (see point 3, below).

2. The second strand of discussion about intelligibility concerns the explainability of the *outcome* of *applying a decision rule* on particular individual cases. It is exemplified by recommendations such as:

a) The proposal of building a "why did you do that" button in AIs interacting with humans (14)

b) The idea that "[t]he data provided by the black box could also assist robots in explaining their actions in language human users can understand" (24)

c) The idea that "[i]n some cases it may be appropriate to develop an automated explanation for each decision" (12)

These recommendations could be associated with the idea of what have been called in the literature *post-hoc* explanations of individual decisions by the AI, which is further examined below (4.2.4). By analogy with human explanations, the relevant explanation

30 The element of communication (besides documentation) is evident in the Trustworthy AI guidelines' assessment list, which includes the following items "Depending on the use case, did you consider communication and transparency towards other audiences, third parties or the general public? [...] Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system? In case of the system's development: to whoever is deploying it into a product or service? In case of the system's deployment: to the (end-)user or consumer?".

will not be holistic but rather identify the most important factor, or a limited set of important factors, and characterize those factors as the reason(s) behind a particular recommendation or decision (44–46).

Some intelligibility guidelines do not indicate or entail a preference for the first (holistic) or the second (*post-hoc*) approach to explanation. They simply emphasize the importance of providing explanations that are at the right level of simplicity for a target audience, while remaining agnostic on what that is. The background idea is acknowledging uninterpretable black boxes (47)[31] and yet documenting, explaining and communicating the logic behind the adoption of these rules, in ways that make the assessment of this logic assessable, for all the relevant pragmatic purposes, e.g. legal ones (45,46,48). Such guidelines may be characterized as "agnostic" with respect to the form of intelligibility at stake. The majority of explainability recommendations in the twenty guidelines analyzed here are of this type. For example:

d) The requirement that an organization assesses:

   i) "the extent to which the algorithm is opaque (a black box)" (17)

   ii) "how much of your system/algorithm can you explain to your users and stakeholders?" (12)

e) "[h]ave a plan for how decisions will be explained to users and subjects of those decisions" (12)

f) "consider whether a directly interpretable or explainable model can be used." (12)[32]

g) "the systems must be able to provide an explanation of their decision-making that is understandable to end-users and reviewable by a competent human authority. Where this is impossible and rights are at stake, leaders in the design, deployment and regulation of ML technology must question whether or not it should be used" (21).

So far, we have considered guidelines that stress the importance of *technical* solutions. However, some guidelines stress *organizational* solutions for promoting intelligibility, such as, for example, providing a human contact point for people affected by the decisions who want to know about the logic involved. For example:

> *Artificial intelligence systems' transparency and intelligibility should be improved, with the objective of effective implementation, in particular by: [...] c. making organizations' practices more transparent, notably by promoting algorithmic transparency and the auditability of systems, while ensuring meaningfulness of the information provided [...]* (16)[33]

> *Introduce a new 'Certificate of Fairness for AI systems' alongside a 'kite mark' type scheme to display it. Criteria to be defined at industry level, similarly to food labelling regulations.* (11)

Other guidelines (typically those addressing governments, public sector organizations, or multiple stakeholders) also propose *institutional* solutions for promoting the adoption of more easily intelligible, and therefore transparent, AIs.[34] E.g.

> *Establish an AI regulatory function working alongside the Information Commissioner's Office and Centre for Data Ethics – to audit algorithms, investigate complaints by individuals, issue notices and fines [...] and ensure algorithms must be fully explained to users and open to public scrutiny.* (11)

---

31 Sometimes the reason why these are black boxes is intrinsic to the type of technology, sometimes it is a choice not to reveal their inner workings.

32 Analogously, the guidelines for Trustworthy AI ask organizations "Did you research and try to use the simplest and most interpretable model possible for the application in question?" (29).

33 These are "non-technical methods" in the sense of the guidelines for trustworthy AI (29), respectively "accountability via governance frameworks" and "certification".

34 These are also instances of a "non-technical method for trustworthy AI" (29), namely "regulation".

*Introduce a 'reduced liability' incentive for companies that have obtained a Certificate of Fairness to foster innovation and competitiveness.* (11)

*Introduce a mandatory requirement for public sector organizations using AI for particular purposes to inform citizens that decisions are made by machines, explain how the decision is reached and what would need to change for individuals to get a different outcome.* (11)

### 4.2.3. Relevance to HR analytics

Let us now explain the relevance of these ethical recommendations for AI used in HR analytics. First of all, let us consider the prescription to know, assess, and document how an AI (learned algorithm) works (i.e. its performance metrics and limitations) and why (its sources, its goal, parameters, and how it was made). The knowledge about the algorithm is ethically valuable in so far as it contributes to building trustworthy technology. Following Ferrario, Loi and Viganò (49), the trustworthiness of AI is understood here as AI having those properties that make it rational for users to *simply trust it,* i.e. it is rational for users to rely on the technology even with little knowledge and control over how it works. That is, when the technology is used for the purpose for which it is designed and recommended, even in the absence of detailed knowledge and control by the end-user, it produces a positive payoff or utility for the average user. Trustworthy technology is convenient and safe to use. It can only exist when those designing the technology understand *both* the technology itself and the context of its use (and misuse) well, so they can anticipate possible problems and avoid them. Thus, trustworthy AI used in HR is designed to benefit HR managers who rely on it, even when HR managers do not have the capacity to technically evaluate its trustworthiness, as long as the technology is used in accordance to its specified purpose and well-defined limits. Trustworthy AI in HR produces benefits for HR managers who cannot fully understand *why* the technology is trustworthy (50). Trustworthy AI in HR is a goal, and it is not yet clear if it is achievable. Improving the process of knowledge and control of the algorithm should make AI more trustworthy, since it involves knowing and clearly documenting the limits of the technology (e.g. which populations it will classify accurately and which inaccurately), and how it may be misused. This should lead to better design, and through transparent communication, mitigate the possibility of misuse.

However, in the case of AI this simple trust (relying on the producers to do all the technical and evaluative steps correctly) is arguably not sufficient to deliver trustworthy technology. A different kind of trust, *reflective* trust (49), should be in place to prevent such technology from harming a user. The idea here is that market competition is not sufficient to guarantee that successful companies produce trustworthy technology.[35] Arguably, AI in HR needs *watchdogs* (e.g. auditors, certifying entities, NGOs, investigative journalists, etc.) to monitor and criticize AI used in HR solutions, as a complement to market incentives, for businesses to deliver trustworthy technology. The assessment of competent and independent watchdogs can help to build *reflective* trust – trust based on the belief that a technology is worth relying on (for the purpose for which it is marketed), which is ultimately based on a trustworthy assessment of its quality (49). Thus, reflective trust implies *transparency* about the design of AI for HR solutions, with watchdogs as the intended audience of communication. The designer's knowledge about the algorithm should be communicated appropriately to these entities.

The reason why transparency is particularly important for predictive and prescriptive AI in HR and less for existing products lies in the difference between AI and current products. First, low quality AI may not be immediately identified by consumers of AI as low quality. The inaccuracy of AI-driven decisions in HR, its biases, and the unfair decisions that it may inspire may not be detected for a long time. Meaningless, harmful decisions can be made and harm employees,

35 E.g. one may be skeptical that Facebook, clearly a successful company in terms of marketability, produces trustworthy technology – technology that benefits users who simply trust it (without understanding it). E.g. users should not have trusted Facebooks algorithms (at least in the past) to display news content worth paying attention to.

before flaws in the AIs are recognized. Management decisions are considered fair when it relies on clear and consistently applied rules (51). With predictive and especially prescriptive HR powered by AI, the rules are embedded in the HR analytics solution, and may be embedded opaquely, e.g. if they are too mathematically complex. The opacity of such rules may undermine trust in management. Meta-analytical work in organizational behavior has demonstrated that the different sub-dimensions of organizational justice uniquely and positively contribute to performance, trust, job satisfaction, and organizational commitment, and negatively relate to turnover intentions and absenteeism (52,53). Mechanisms to implement transparency and accountability about the algorithm can be thus regarded as a means to achieve reflective trust in AI in HR. In conformity with this idea, one guideline says:

> *Workers must have the right to demand transparency in the decisions and outcomes of AI systems as well as the underlying algorithms [...].* (24)

It is clear however, that not all workers can obtain transparency about the algorithm in the sense of the algorithm being explained directly to them. Rather the algorithm may realistically be made transparent to auditors and other watchdogs. Auditing and certification may be supported by new legal instruments, creating incentives for companies to be transparent and allow auditors to do their job. It is yet an open question whether reputational concerns are sufficient to generate the necessary incentives to improve algorithms, or whether legislative pressures are also required. For the moment, it may be hoped that reputational concerns create a market for auditors and certifiers of trustworthy AI, and that reputational concerns drive producers of AI in HR analytics to seek such forms of transparency. Transparency without accountability does not lead to improvements. If no one is responsible when AI systems do not work as they are reasonably expected to, transparency about the problems in AI will not translate into process improvements, unless the roles responsible for the problem and correcting it are identifiable.

On the other hand, workers may be provided directly with explanations of algorithmic *decisions*. This can be different from explaining the *algorithm*, i.e. its goal and general logic. In fact, a significant strand of research on algorithmic transparency and explainability concerns methods that make individual algorithmic *decisions* more easily *interpretable* (45). The relevance of this type of explainability for HR analytics is twofold. Firstly, average end-users of AI may need explanations of AI decisions that are simpler than those provided by holistic models of transparency. Without such explanations, they may not develop (reflective) trust in the model (45). For example, an HR manager implementing AI recommendations may regard it as irresponsible to choose an employee based on a recommendation whose rationale they do not fully understand. This is highlighted in the UNI guidelines, which say:

> *"For users, transparency is important because it builds trust in, and understanding of, the system, by providing a simple way for the user to understand what the system is doing and why."* (24)

Secondly, employees affected by the decisions of AI used by end-users in HR may demand explanations to HR decision-makers. Just like end-users, employees affected need explanations they can reasonably be expected to understand. Understanding the rationale of decisions, especially decisions involving inequality, is known to augment the perception of a decision as just fair (54). A further idea is that if people adversely affected by a disadvantageous algorithmic decision are aware of the reason behind the decision they may take a reasonable course of action which would result in a different outcome, more favorable to them (46,55).

One approach to achieve this is to avoid algorithms that are considered "black boxes" and instead use models that are easily interpretable. Another

approach consists in using more complex models, typically regarded as non-interpretable (so-called black boxes) and make them interpretable.[36] Some models can be designed to be so simple that most users can intuitively understand what they do. These models are not as accurate as many algorithms that tend to be black boxes. Hence, a fashionable way to achieve intelligibility consists in building a simple algorithm that approximates the behavior of a black box, at least for a limited range of inputs of interest to the stakeholder (40). Such algorithms can only provide an *approximate* explanation of what drives AI's predictions or decisions, which does not reveal the authentic internal logic of a very complex decision rule (45). Hence, such attempts to make AI explainable cannot explain the full range of behavior of a complex inscrutable algorithm in all potential scenarios of operation.

One approach of this type focuses on quantifying the influence of different inputs to a decision (55). E.g. if an employee is refused a promotion, she can be told what are the factors affecting the decision how they affect it (positively or negatively) and what are their respective weights. Alternatively counterfactual models provide a list of the most important features that the employee would need to possess in order to obtain the desired outcome (46). For example, a counterfactual explanation may be "you would have obtained the position if you had had a better level of English and at least 3 additional years of experience in your present role". The psychological effects of these and other kinds of explanations have been tested empirically, including in a fictional scenario concerning the use of algorithms to decide a promotion decision (55). The study showed that providing people with explanations generally improves the perception of a decision as just, as predicted by the psychological literature. But the study did not prove that "input influence" explanations or counterfactual explanations[37] are more successful than other kinds of explanations, in enhancing fairness perceptions.

Plausibly one would rationally make individual decisions intelligible and transparent in the service of several substantive values:

1) *Post-hoc* explanations of HR decisions by AI to AI *operators* may enhance the trustworthiness of the technology in real-world applications. This is achieved if post-hoc explanations of individual decisions enable HR operators with a limited grasp of the technology to detect flaws in AIs, or to avoid clear cases of AI misuse. This information from HR operators may enable designers to better understand how the AI behaves, and thus, how (and why, and in which scenarios) it may misbehave. When this is the case, *ad hoc* explanations are required by the principles of beneficence, maleficence, and justice (e.g. when the misuse involves illegitimate discrimination).

2) *Post-hoc* explanations may enhance the freedom of individuals subjected to algorithmic decisions. In the HR context, this is the case if *post-hoc* explanations of HR recommendations by the AI to the workers enables workers to change their behavior in ways leading to better outcomes *for them*. This is supported by the principle of autonomy.

All potential ethical advantages of explainability, transparency and accountability and their relation to substantive ethical values are highlighted in the graphics below.

36 The distinction here is not between neural networks being black boxes and everything else being intelligible. The behavior of linear models with a huge number of interacting factors may also defy intuitive human comprehension (40).

37 Of all psychological purposes in this study, the concept of "sensitivity explanations" sufficiently resembles the concept of counterfactual explanations.

| | | | |
|---|---|---|---|
| Knowledge and control of the algorithm | Robustness and trustworthiness | Simple trust benefits those who trust the technology | Beneficence and non-malefic |
| Knowledge and control of the algorithm | Design transparency and acccountability | Reflective trust by watchdogs; simple trust by average users | Beneficence, non-maleficence and justice |
| Interpretable algorithms or X AI for black boxes | Post-hoc explanation of individual decisions | Robustness in deployment | Beneficence, non-maleficence and justice |
| Interpretable algorithms or X AI for black boxes | Post-hoc explanation of individual decisions | Agency enhancement in decisions subjects | Autonomy |

### 4.2.4. Open challenges for algorithmic transparency and accountability:

There is a mystique about neural networks being entities that intrinsically escape control by their designers. This mystique should be rejected. Designers can always control the algorithm in the sense of taking responsibility for its goal, for the way the goals are translated into a mathematical function, the quality of the data fed to it, how its performance is evaluated, for monitoring the performance in the live context, and for recommending its use in that context (41–44). There are however several challenges; we shall examine three of them.

#### First challenge to transparency: manipulation

Many calls for transparency about the algorithm ignore that in certain cases transparent algorithms can be morally problematic. In some cases, transparency about the algorithm or its decisions may reveal sensitive private information about individuals. In others, it may lead to gaming the algorithm. For example, if the algorithm used to identify fiscal evasion is well known, people who want to avoid paying taxes will optimize strategies for avoiding being caught. Even in the context of HR analytics this could be a problem, i.e. algorithmic transparency may invite attempts to game the algorithm: knowing the proxies that are used for performance diverts the objective of the activity, i.e. from working properly to maximizing whatever measure is used as a proxy of performance.[38] Few of the guidelines examined here highlight this risk, while most of them propose algorithmic transparency as an unqualified good:

> *Can you provide for public auditing (i.e. probing, understanding, reviewing of system behavior) or is*

38 This is analogous to the problem of students learning in order to pass a test or scholars focusing on maximizing citations. Test scores and citations are at best an imperfect proxy of, respectively, learning and the value of scientific contributions. Aiming at the maximization of proxies diverts the energies of the people involved to pursue goals that are strategically valuable but sub-optimal from the point of view of the good that should be promoted, e.g. learning or scientific knowledge. Beside gaming the system (when the indicator becomes the target in itself (56,57)), other effects in academia are risk avoidance (highly innovative or interdisciplinary topics are avoided because they do not score well (58)) and task reduction (teaching and public engagement are avoided in order to focus on published research (59)). Take the case of the Italian academia, where the aggregate impact factor has, counter-intuitively, been growing in spite of significant cuts in funding in the last decade, due to practices designed to boost the individual impact factors of researchers. The explanation is that, while research funding was cut, measures such as citation counts were made a legal condition for certain career progressions to be authorized (60).

*there sensitive information that would necessitate auditing by a designated 3rd party?*

*How will you facilitate public or third-party auditing without opening the system to unwarranted manipulation?* (12)

**Second challenge to transparency (and accountability): reliability and reproducibility**

The second challenge is that the proper testing of ethically salient qualities of an algorithm may be difficult or impossible if limited to *lab conditions* for a combination of technical and social reasons. While it is true that learning-based models are tested and validated with historical data, there is no strict guarantee that the model will achieve its intended goal when applied to new cases, in a new context, with new data. This is why a system ought to be monitored when it is used in the real world, as recognized for instance by the Trustworthy AI guidelines:

- *Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility? [...]*
- *Did you put in place processes to describe when an AI system fails in certain types of settings? [...]*
- *Did you test for specific populations or problematic use cases?* (29)

The methodological challenge is that only a limited set of scenarios can be predicted and tested in the lab. There can be unpredicted, critical scenarios that will only occur in reality but are not anticipated. That means that situations may emerge in which the AI responds in ways that are not expected. One cannot be transparent and ensure that the model will behave fairly in such contexts. This is an engineering problem of identifying reasonable safe, robust, yet feasible testing procedures. For deep neural networks, practices that are sufficiently robust with ordinary software, e.g. extending the testing phase to a few iterations in which the software's behavior is assessed in practice, may not be sufficient. The problem is amplified if the neural net trains itself based on the results of the decisions it takes during operation. Such problems are not ignored by some of the recommendations we have examined. For example, one recommendation of Women leading in AI says:

*"Introduce a regulatory approach governing the deployment of AI which mirrors that used for the pharmaceutical sector".*(11)[39]

The thought underlying this is plausibly that if one's goal is to know and document how AI behaves, the testing with historical data performed in the lab is not sufficient. In terms of the data pipeline, that means monitoring the real outcomes of algorithms (decision rules) and using this information to improve the phase of AI design. Most widely used methodologies for the data pipeline already include iterations (61). However, it may be necessary to demand even more iterations and continuous monitoring of real-world outcomes when the stakes of a data-driven model are high and the models are black boxes.

**Third challenge for transparency: black box models**

The third challenge concerns the explanation of black box models. The technologies mentioned above produce their ethically beneficial consequences only if some speculative psychological hypotheses prove true, namely that providing explanations of AI decisions enhances the awareness of AI end-users (who are not data scientists) of the possible flaws and misuses of the AI, and that providing explanations to people affected empowers these people to find reasonable alternative courses of actions in their interests. Moreover, the idea of explaining a black box with a simpler model or by providing a counterfactual explanation may appear more resolutive than it is in fact. If a black-box model in fact relies on a large number of factors to decide, the reduction of its dimensionality may in fact require some arbitrary choices. E.g. there may be ten factors, all of which

39 Analogously, one checklist item of the guidelines for Trustworthy AI asks: "Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications? (29).

with roughly the same weight, and the explanation model will present only four, ascribing to these four an importance they do not really have. A counterfactual model may suggest a course of action that is in fact not feasible, such as changing a person's past salary or educational choices. And if it is designed to only suggest feasible courses of action, it may suggest none that is desirable for the person affected by the decisions (44).

### 4.2.5. Algorithmic fairness and HR analytics

The problem of discrimination and unfairness of predictive and prescriptive analytics in AI is clearly one of the most important challenges when using AI to guide HR decisions. Nowhere is the potential for a clash of visions stronger, that is, between the statistical justification of decision-making, and fairness understood in terms of the values that regulate labor relations. The importance of this topic for ethical AI is also reflected in the fact that risk of unfairness and/or discrimination is also mentioned by the majority of the ethical codes examined here. When this concern is not mentioned explicitly as fairness, or equal treatment (and its opposite, discrimination), it sometimes appears as a discussion of biases and errors that happen to disproportionately affect certain groups.

Apparently, recommendations concerning fairness, discrimination, and bias are the most straightforward. For example, the principle to "Ensure a Genderless, Unbiased AI" says:

> *In the design and maintenance of AI, it is vital that the system is controlled for negative or harmful human-bias, and that any bias—be it gender, race, sexual orientation, age, etc.—is identified and is not propagated by the system.* (24)

This is obviously relevant in the context of HR analytics, since a problem with AI tools is that they can introduce an implicit bias – an inclination to favor individuals who belong to certain groups due to statistically significant similarities between them – into the decision-making process. This happens for example with AI based on (statistical) machine learning, which learns similarities between individuals from historically gathered information. The machine learning process does not need to be told about race and gender explicitly, for example, in order to perceive and take into account a similarity that groups together (disproportionately) members of the same race or gender group (37). Hence, populations that have suffered past human and structural biases — also referred to as protected groups — are susceptible to damage from inaccurate projections or resource allocations, which reinforces historical inequalities. For example, AI helps by offering job openings to the "right" kind of people on personal job boards, such as ZipRecruiter, that learn about the preferences of recruiters for certain job candidates over time, which may be biased in favor of people of a given sex, race, or social class(62,63). The different native languages of workers may introduce unjustifiable disparities in selecting individuals for a given educational/training course or job(64).

The recommendations concerning fairness include the four fundamental action types described in the introduction: 1) documenting unfairness, 2) making unfairness transparent, 3) assigning moral or legal responsibility for unfairness, 4) mitigating unfairness. All of these are relevant to HR.

### 4.2.6. Open questions for fairness in machine learning

Unfortunately, there is a gap between the everyday concept of fairness and the attempted statistical definitions of it. Furthermore, there is a conceptual gap between fair prediction/classification and the legal concept of non-discrimination. This makes it difficult to document unfairness, make it transparent, mitigate it, and attribute blame or moral responsibility for outcomes that may reasonably appear unfair to some stakeholder groups. All these activities *presuppose* an objective, or at least inter-subjective, measure of unfair bias or unfair indirect discrimination. The difficulty, moreover, is not purely technical. There are deep, conceptual questions that are still open – the academic literature has barely begun

to interpret from the moral and legal point of view the normative importance of these statistical issues. One of the few agreed ideas is that what should be considered (wrongly or illegitimately) discriminatory or unfair is context-dependent. This is reflected in the more prudent language found in some guidelines, previously mentioned. The WEF guidelines mentioned above (21) explicitly acknowledge that fairness is contextual, that some forms of "discrimination" are intrinsic to the task of the AI, that there are different forms of fairness. They also acknowledge that different fairness definitions may be appropriate depending on the context, and that in order to determine what is fair in a specific context it is necessary to involve domain experts and interdisciplinary insights.

A relatively simple element is to detect the unequal error rates by population, considering different kinds of errors, in particular distinguishing false positives and false negatives. The challenge is, however, how to interpret an inequality in such error rates, since, as it will be shortly explained, there is no single measure of unfairness of inequalities in error rates and ascribing bias to an algorithm is a less obvious claim than it may appear at first sight.

We will now illustrate some hard, moral problems of fairness in machine learning through two hypothetical examples, both related to the HR context[40]. Suppose that you are training an HR analytics tool to advertise programming positions in your company – a vast multinational corporation with branches in many countries. Your target is to display the job opening to employees who are likely to be programmers and avoid showing it to employees who are not, so that they are not burdened by the information and can focus on positions more relevant to them. You want a model that predicts who will be interested in the ad by analyzing the browsing history of your employees, who have agreed to share such personal data with you exclusively for the purpose of such processing. Your model claims that the probability of clicking on such job ad is higher than average for an employee who visited stackexchange.com and lower than average for an employee who visited pinterest.com. The reason for this is that, in your training data, a very large proportion of people who visited stackexchange.com were in fact computer programmers interested and able to take such jobs. This is not surprising given the nature of this website, which is visited mostly by people who know how to code or are learning to code. On the other hand, the negative weight attribute to pinterest.com is due to the fact that most people accessing this website in the training set were women and the women landing on this website were also very unlikely to be computer programmers. When you use the algorithm, it turns out that 95% of employees who were shown the ad for a job opening as a computer programmer were males. In both law and philosophy, this is known as "indirect discrimination" ("disparate impact" in US legal language). What indirect discrimination means is that a facially neutral criterion produces unintentionally different results when applied to different populations.

There is not widespread agreement in the machine learning community on how to treat such cases. If fairness consists in demographic parity (66), it is unfair to implement a rule that is more likely to display the ad to a woman – men and women should be on average equally likely to be shown the ad, irrespective of their actual skill set and interests in such ads. According to a different standard of fairness, called equalized odds (67) or equal mistreatment (39), it is unfair to implement a rule that is more likely to display the ad *when the employee is actually interested in the job* if the employee is a woman. A possible form of auditing, based on equalized odds, would first have to find out whether the employees are in fact interested and qualified for computer programmer jobs and then determine if those who are, are equally likely to be shown the ad, irrespective of their gender.

One can imagine arguments for either fairness standard. In favor of equalized odds, one could say that it is not unfair if employees, who are not able and interested in such a job, are not shown the ad. If these employees are disproportionately of one gender, a departure from demographic parity

40 A similar example is found in Gilbert (65). The case here is modelled after the "short hair/long hair" example found in (48).

showing ads predominantly to members of the other gender is justified. Against equalized odds, and in favor of statistical parity, one may argue that this is a way to reproduce the *status quo*. It is important to show to women (even women who are not computer programmers and not interested in applying for such jobs) how many good opportunities there are in this field. This may provide an incentive for women to learn coding in their free time, or to ensure that their daughters are offered a fair opportunity to learn to code.

Another critique of equalized odds (the equalizing of the false positive and false negative rate[41]) follows from arguments that favor a criterion of *predictive* parity, the equalizing of *specificity*. Like equalized odds, predictive parity is compatible with the probability of being offered a job (or being shown an ad for a job) being different across groups, e.g. in so far as one group contains more members qualified or interested in that kind of job (66,68,69). But predictive parity gives even more importance to the statistical tendencies associated with group membership than equalized odds do. In fact, it is possible for a decision rule to satisfy predictive parity and yet for two people who are equally suitable for a job to be given different chances to be shown a job ad, if one belongs to a group that is statistically more likely to have the required characteristics and the other belongs to a group that is statistically less likely to have them (68,69). Some statisticians have claimed that the appropriate fairness criterion is predictive parity, not equalized odds. Predictive parity requires that, when an organization makes an HR decision based on a prediction (e.g. the decision to display an open position in the organization to an employee, based on the prediction that the employee will be interested in that opening and click the ad), the same proportion of positive predictions should turn out to be correct, irrespective of the group to which the employee belongs. E.g. the algorithm to display programmer openings in the company is fair even if men are more likely than women to see the ad, as long as the proportion of women (who are shown the ad) who click on the ad is (roughly) equal to the proportion of men (who are shown the ad) who click on the ad. While seemingly analogous to the equalized odds fairness criterion, this criterion is in fact mathematically incompatible with it in all but rare circumstances (70).[42]

One argument for predictive value parity (and thus, against equalized odds in most circumstances) is that it is what an employer would do, if she values the contribution to the company of all prospective employees equally, irrespective of their groups. To see why this may lead to a violation of equalized odds (to unequal false-positive and false-negative rates), consider the following hypothetical scenario. Suppose that you are training a machine learning algorithm to purge a huge pile of CVs of potential candidates for a position as school bus driver. You do not want to hire anyone for that position who cannot be trusted, hence you want to exclude those drivers who are more likely to drive while drunk. You have access to data about drivers who have been fined or had their driving license removed due to being caught drunk driving. You also have access to the internet browsing history of these users. Your training data is representative of different human populations and has as many data from populations with a predominantly Muslim religion and from populations with a predominantly Christian religion. It turns out that the highest accuracy is achieved by an algorithm that considers whether a person has visited an alcohol-related website (e.g. of a merchant specialized in alcoholic drinks, or wine ratings). We furthermore assume that there are no biases in the data due to the practice of stopping drivers for alcohol checks, namely, there was no preferential stopping of drivers based on their

41 This implies that the sensitivity of a classifier is the same for both groups.

42 The difference amounts to the following: while predictive parity is achieved only if women and men who are in fact shown the ad are equally likely to click on the ad, equalized odds requires that the women and men who (when asked) display interest in such ads (e.g. by clicking them) be equally likely to be shown the ad (68). In the example in question, equalized odds and predictive parity can be achieved simultaneously only if at least one of the following (unlikely) conditions are satisfied: (a) the proportion of people actually interested in the ad is exactly the same in the two groups (i.e. exactly the same proportion of female and male employees would click on the ad if shown) or (b) it is possible to predict with perfect accuracy (100% correct predictions) whether an employee will click on the ad.

perceived religion or other traits statistically correlated to religion.

We can explain why predictive parity equalizes the prospective gain for the employer independent of group membership, by considering how a rule based on predictions treats individuals whose actual label of driving while drunk is known. Predictive parity between Christian and Muslims is achieved if, by looking at the profiles of those candidates that the *decision rule* has excluded from the job, the proportion of drivers caught drunk is the same for Christian and Muslims, and if by looking at the profiles of those candidates that the *decision rule* has not excluded, the proportion of drivers actually caught driving while drunk is the same for Christians and Muslims. Suppose that Christians and Muslims who visit alcohol-related websites are on average equally likely to end up driving while drunk; and that Christians and Muslims who do not visit alcohol-related websites are on average equally likely to stay sober while driving. The decision rule produces the same benefit for the company (the benefit of selecting a driver who is not unsuitable) irrespective of the religion of the applicant: whether the company hires a Christian or a Muslim, after the recommendation of the tool, the company is equally likely to have hired someone who will drive while drunk, irrespective of the religion to which they belong.

But while this is achieved, *equalized odds* cannot be achieved, that is, there will be unequal false positive and false negative rates for Christians and Muslims if, (as it seems plausible) there are more Christians, compared to Muslims, among visitors to alcohol-related websites. Christians who do not drink and drive are more likely to be excluded than Muslims who do not drink and drive, because they are more likely to have visited alcohol related websites.[43] Thus, one cannot in situations such as this satisfy both predictive parity and equal odds. The choice between the statistical criteria of predictive value parity and equalized odds is a *moral* not a *technical* one. Since it is a question of *value*, there is no obvious answer to this question from the experts in fairness in machine learning (meaning, based on the authority given by being experts in this field, or statistics). Possibly, there is not a single right choice but different statistical conditions will be appropriate in different scenarios. Philosophers and machine learning theorists are still discussing how to make sense of such choices (72,73). The ultimate ground of the discussion is not merely technical (although the discussion concerns the technical notion of conditional probabilities) but ethical: it concerns the most appropriate interpretation to give to the idea of "treating people equally, if they are equal in the relevant respect" (e.g. should people who achieve the same outcomes be given the same chances, even when the probability that they achieve the relevant outcome differs? Should our decision about individuals depend on the information we are able to collect about them, even when the information in question does not concern them as individuals, but is only informative of their similarity to other people?). This is why, in spite of a somewhat simplistic ultimate declaration that an algorithm has been found to be "discriminatory" by the press, most claims about *bias* and *discrimination* are naturally controversial – they rely on moral assumptions about what is (and what is not) fair in a decision guided by probability, that deserves to be further discussed. The scientific debate on how to reason from the features of a situation (that are morally relevant) to the choice of an appropriate statistical constraint is only just beginning to emerge (72).

43 This inequality is possible for a rule that is equally accurate when judging Muslims and Christians. E.g. if there are only 3 Muslims who visit such site, and 3'000'000 Christians, a rule with 2/3 accuracy (for the positives) will falsely predict drunk driving for 1 Muslims and 1'000'000 Christians. If there are only 3 Christians who do not visit alcohol related websites, and 3'000'000 Muslims, a rule with 2/3 accuracy (for the negatives) will falsely predict not-drunk-driving for 1'000'000 Muslims and 1 Christian. This rule may be considered biased in favor of Muslims, since it falsely classifies as (for simplicity) "drunk drivers" 1'000'000 Christians and 1 Muslim, while it wrongly classifies as "not drunk drivers" 1'000'000 Muslims and only 1 Christian. These inequalities in the rate of misclassification (for both positives and negatives) may plausibly be considered advantageous for Muslims and against Christians. A conceptually similar case of "unequal mistreatment" (in this case, of White vs. Blacks) was found by ProPublica in relation to the COMPAS recidivism tool (71). The company's defense of the tool was to show that, in spite of the very clear unequal odds for Whites and Blacks, it almost achieved perfect predictive parity (69).

## 4.3. Kndowing, communicating, owning, and improving the human impact

### 4.3.1. Goals and ambitions of generating knowledge about the human impact of algorithms

The above discussion (in 4.2.4) shows that not all the morally relevant features of algorithms (i.e. decision rules) can be established by assessing and testing algorithms in the lab. It may be the case that an algorithm has been trained with valid and unbiased data, equally representative of all human populations; it may be the case that its accuracy has been measured and is entirely satisfactory given the task for which it is used; it may be the case that all the metrics relevant for fairness have been measured and the algorithm has been modified to achieve a distribution of errors between the different populations, that most people would consider fair, for the context in question. And yet, all those assessments and measures are based on data from the past. If the data for training and testing have been chosen appropriately, the future behavior of the decision rule should resemble quite closely the behavior displayed in the lab. But one cannot simply be sure that this will be the case. The assessment based on data from the past may not be sufficient to evaluate the performance of a decision rule (produced by a learning algorithm) when it is applied to the new employees in the real world. This problem is typically indicated by saying that the AI is not *robust* – that it fails to *generalize* appropriately.

This uncertainty about the future is not only due to the fact the behavior of "black box" AI, e.g. neural networks, is essentially unknown for all possible inputs. And it is also not limited to those AIs that evolve as they are used, even if these tend to be the least predictable, as they modify their rules based on how reality responds to them. The unpredictability of all AIs derives from a much broader phenomenon, namely its embedding in a human context that is always, because of its very nature, somewhat unpredictable, in ways that are not reflected in the training and test data. This is, of course, not a problem concerning AIs in particular, but rather a potential problem for all technologies. But in the case of AI, in particular, AI in the workplace, the risks associated with a negative human impact that is not foreseeable in the lab appear particularly serious.

When one considers the impact of AI on society, one can distinguish at least two different levels: the impact on individuals (e.g. how incorrect was the prediction about an individual employee's performance?) and on groups (e.g. is the *decision rule* making it harder for black people to be hired?). Moreover, one can distinguish two kinds of individuals and groups: those involved in contractual and economic relations with the organization deploying the AI in HR (in particular, but not exclusively, job candidates and employees) and everyone else. In this section, we are interested in the outcomes *on individuals who have a contractual and economic relationship with the organization* using AI for HR analytics.

There are several guidelines that recommend generating knowledge about the actual impact of algorithms on humans. The philosophical and legal normative concept used to indicate such guideline is not always the same in all guidelines. For example, some guidelines discuss human impact in relation to *fundamental human rights* (16), others use the concepts of *risk prevention and mitigation* (14), *ethics by design* (16), *public safety obligation*(15), and *responsible deployment* (28). In a sense, this is the least concrete aspect of the guidelines examined so far. Many guidelines stress the importance of monitoring the execution of AIs when they are actually implemented in an organization, yet they offer very few concrete organizational solutions to implement this requirement. Statements of principle without concrete guidance dominate the landscape. Here are some examples:

> *We propose that companies work on concrete ways to enhance company governance, establishing or augmenting existing mechanisms and models for ethical compliance."* (21)

> *"As part of an overall "ethics by design" approach, artificial intelligence systems should be designed*

*and developed responsibly[...] in particular by: [..]b. assessing and documenting the expected impacts on individuals and society [...] for relevant developments during its entire life cycle [...]"* (16)

*"Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices."* (15)

*The capacity of an AI agent to act autonomously, and to adapt its behavior over time without human direction, calls for [...] ongoing monitoring.* (28)

*Adopt and maintain policies and procedures reasonably designed to collect information sufficient to conduct assessments that would detect any significant disparate impacts, including, if necessary, collecting sensitive information such as race, gender, ethnicity, and religion or constructing accurate proxies for such sensitive information.* (26)

*A/IS* [Autonomous/Intelligent Systems] *should prioritize human well-being as an outcome in all system designs, using the best available, and widely accepted, well-being metrics as their reference point.* (14)

### 4.3.2. Recommendations

Apart from general and vague declarations of principles (and equally general assessment requests associated with the same principles), a few concrete recommendations surfacing through the twenty guidelines here are worth considering:

Firstly, *competence alignment.* There should be skill and competence alignment between those functions in an organization that employ AIs and those that are responsible for designing and testing them. The first edition of the IEEE guidelines on Ethically Aligned Design (74) contained an interesting recommendation for both creators and users of AI products. Namely: creators of AI products must specify the level of background knowledge and skill necessary for AI operators to operate AI products safely; and organizations should ensure that operators have the required skills and competences. In the context of HR analytics, this implies that AI-based tools should only be used by HR professionals with a required level of understanding of the logic of the decision rule and its potential flaws. One could imagine that AI products in HR analytics are always provided with an "ethical instruction manual" that explains the limits of the models, including the circumstances that may lead the model to malfunction, how the decisions of the model should be explained, especially if contested, what has been done to make the model fair, what potential problems should be expected, how to monitor and assess them, what steps must be taken to adapt software to the social circumstances (and avoid unfair bias) and how to communicate the problems. An important aspect of this competence alignment is, as highlighted by the FAT-ML recommendation, that the data-scientists "[d]etermine[s] how to communicate the uncertainty / margin of error for each [AI-driven] decision" (12).

Secondly, a concrete implementation guideline is the procedure of, *providing a feedback mechanism*, often amounting to a right to *challenge or correct algorithmic decisions*, especially those that are entirely automated. This idea is to be found in more than one guideline:

*"5. [...] guarantee[...], where applicable, individuals' right not to be subject to a decision based solely on automated processing if it significantly affects them and, where not applicable, guarantee[...] individuals' right to challenge such decision, [...]* (16)

*"Develop a process by which people can correct errors in input data, training data, or in output decisions"* (12)[44]

44 Similar items appear in the assessment list of Trustworthy AI (29), i.e. "Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system? Did you establish clear steps and ways of communicating on how and to whom such issues can be raised? Did you consider others, potentially indirectly affected by the AI system, in addition to the (end)-users?"

Thirdly, some guidelines recommend that organizations implementing AI driven solution have *redress procedures* in place to deal with cases in which, over time, significant harm or unfairness emerges as a result of the application of the AI:

> *Access to Redress: Leaders, designers and developers of ML systems are responsible for identifying the potential negative human rights impacts of their systems. They must make visible avenues for redress for those affected by disparate impacts, and establish processes for the timely redress of any discriminatory outputs."* (21)

Many guidelines also require organizations that are developing AIs or those implementing AI-driven solutions (or those that do both) to be aware of the implications of AI and AI-driven solutions on society as a whole, not only their clients. With respect to such "whole society" questions, concrete guidance seems to be entirely lacking within the documents reviewed here:

> *"Make AI Serve People and Planet"* (24)
>
> *"Share the Benefits of AI Systems"* (24)
>
> *"Secure a Just Transition and Ensuring Support for Fundamental Freedoms and Rights"* (24)
>
> *"Ban AI Arms Race"* (24)
>
> *Responsible Design and Deployment: We recognize our responsibility to integrate principles into the design of AI technologies, beyond compliance with existing laws. [...] As an industry, it is our responsibility to recognize potentials for use and misuse, the implications of such actions, and the responsibility and opportunity to take steps to avoid the reasonably predictable misuse of this technology by committing to ethics by design.* (25)
>
> *"We will seek to ensure that AI technologies benefit and empower as many people as possible"* (13)
>
> *Societal and Organizational Impact: the AIA needs to highlight the impact on the workforce as well as society / community as a whole. For example, it needs to demonstrate how the system augments human capabilities and how the algorithm does not become policy, thus removing human autonomy in wider decision-making* (11)

Possibly, such recommendations are conceived while having specific use cases of AI in mind, in particular AIs governing the spread of (mis)information and affecting fundamental political rights, in particular, democratic rights. (The ICDPPC explicitly mentions "technologies that influence personal development or opinions" and "respecting related rights including freedom of expression and information" (16). Outside these peculiar domains, the broad societal implications of AIs are either unknown or (given our current prediction capabilities) not foreseeable. Another domain in which broad societal risks have been mapped relates to the use of AI in cybersecurity and autonomous weapons (Future of Humanity Institute et al, UNI). Very little discussion of global risk exists outside this domain (with the exception of the debate about humanity's extinction (or domination) due to malevolent super-intelligence, and the like). This may be due to the fact that applications in other domains have no clear broad societal implications, or that awareness of the risk they pose is not yet mature, except in a few clear cases, such as the spread of fake news and online hate (75).

The guidelines on Trustworthy AI (29), even in the pilot assessment list, do not seem to provide operational questions that help much in terms of concrete guidance. Several claims about the need to assess human impact tell organizations what they *should do* but say nothing about *how* they should do it:

*Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights?*[45]

*Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?*[46]

*Does the AI system enhance or augment human capabilities?*

*Which detection and response mechanisms did you establish to assess whether something could go wrong?*

*Did you verify how your system behaves in unexpected situations and environments?*

*Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity?*

*Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?*

*Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?*

*Did you assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders?* (29)

### 4.3.3. The importance of stakeholder engagement

There seems to be widespread awareness of the problem of the lack of expertise for wide-ranging issues connected to AI in general. This applies also to HR applications in particular. As a response to this, most guidelines include appeals to stakeholder engagement as a practical procedure to fill out the knowledge gaps. "Stakeholder participation" appears as one of the requirements of "diversity, non-discrimination, and fairness" in the Trustworthy AI guidelines (29). It is by far the strongest recommendation in the *TENETS* guidelines by Partnership for AI (13), which mentions some form of it in guidelines 2, 3, 4, 5 and 8. The reason for such wide appeal of this notion is that stakeholder engagement is the variable for whatever is needed to fill an ethical, legitimation, or knowledge gap. Stakeholder engagement is expected to play a role in relation to:

a) Providing feedback on the focus of ethical inquiry, *i.e.* are companies and/or auditors identifying all the relevant risks and vulnerabilities? (10,13)

b) The need for open, interdisciplinary research (13), in particular to identify and bring together different *skills and competences* (11)

c) Collecting the interdisciplinary competences required to identify *potential biases and forms of discrimination* (19,20,25). According to some guidelines, the input from the stakeholders

45 the human rights recognized by the international community are many, as reflected by the several different international covenants, treatises and declarations that many states have signed (76). However, international law includes broader and narrower lists and not all human rights are recognized by all countries. The guidelines provide no guidance on which selected human rights should be considered (if all, the impact assessment may not be feasible for any organization). Nor do they offer any guidance on what practical steps are necessary to assess the impact on human rights on such a list. This is problematic since human rights assessments are not customary activities for most organizations.

46 Since the concept of human autonomy has many meanings, this appears so vague as to be quite useless in practice. E.g. does targeted advertising affect human autonomy by interfering with the (end) user's decision process (i.e. to vote for a particular candidate) in an unintended way? Arguably not, because the effect of the interference is wholly intended. It is not clear if the fact that the answer to the checklist is "no" makes that type of interference with human autonomy ethically acceptable for trustworthy AI.

should be sought to "identify the entire range of data types necessary to adequately train an [sic] ML in a given context" and "understand how to appropriately source the data needed" (21)

d) Knowing the *norms and values* of the data subjects, or populations affected by AI-driven decisions (11,21)

e) Identifying the different *stakeholders impacted* by AI research (13)

f) Identifying *domain-specific concerns* (13,21,25)

g) Avoiding *fears and confusions* regarding AI (14)

h) Promoting new forms of governance that include "various stakeholders" such as "civil society, government, private sector or academia and the technical community" (13,27)

Besides mentioning the different *goals and concerns* (a-g) that stakeholder engagement is supposed to address, these recommendations also indicate different practical forms it may take. These are summarized in table 4 below.

Table 4

| Type of stakeholder | Engagement procedure | Guideline |
|---|---|---|
| Business community | "partnerships with other companies, offer our know-how [...] to jointly tackle the challenges ahead"(10) | (10,13,25) |
| Citizens, broader community | "citizens/stakeholder panels at all stages of the development process and the inclusion of an ethics policy"(11)<br>""It is recommended that public discussions be organized about the implications of new robotic technologies for the various dimensions of society and everyday life [...]" (27)<br>"inform them of our work, and address their questions" (13)<br>"Provide a mechanism for a safe feedback from the audience to which AI is delivered."(21)<br>"We will engage in AI and ethics education."(10) | (10,11,13,14,21,25,27) |
| Policy makers, governments, enforcement agencies | "offer our know-how to policy makers and education providers to jointly tackle the challenges ahead"(10)<br>"Educating government, lawmakers, and enforcement agencies surrounding these issues so citizens work collaboratively with them to avoid fear or confusion (e.g., in the same way police officers have given public safety lectures in schools for years; in the near future they could provide workshops on safe A/IS)." (14) | (10,14,25) |
| NGOs | "Build civil society coalitions and expertise networks: It is important to emphasise the need to develop knowledge-exchange programs and facilitate joint-strategy development between civil society organisations." (20) | (19,20) |
| Scientists and engineers, Academia | No specific mechanism | (10,13,19) |
| Workers, employees | "We will engage in AI and ethics education."(10) "4 [...] Workers should have the right to access, manage and control the data AI systems generate, given said systems' power to analyse and utilize that data" (24) | Deutsche Telekom, UNI (10,24) |

### 4.3.4. Governance structures and accountability

Many guidelines recommend improving ethical outcomes through enhanced accountability. Unfortunately, most guidelines remain quite generic about the nature of the (new?) governance systems that ought to be put in place for this purpose. For example:

> *Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular A/IS.* (14)

> *Organizations should publicly describe the model governance programs they have in place to detect and remedy any possible discriminatory effects of the data and models they use, including the standards they use to determine whether and how to modify algorithms to be fairer."* (26)

> *If accidents occur, the AI will need to be transparent and accountable to an accident investigator, so the internal process that led to the accident can be understood.* (24)

> *"Continued attention and vigilance, as well as accountability, for the potential effects and consequences of, artificial intelligence systems should be ensured, in particular by: […] establishing demonstrable governance processes for all relevant actors […]"* (16)

The few concrete ideas that are mentioned are: to "rely[…] on trusted third parties or the setting up of independent ethics committees"(16) and to "[m]ake available API to query algorithm, allow the research community to perform automated auditing, plan for outside parties" (12) .

One somewhat more specific recommendation about the nature of such responsibility requires that AI is not used as a smokescreen to hide the responsibility of managers responsible for the decisions. That is to say, for accountability, there is always one or more humans behind the AI:

> *For the foreseeable future, A/IS should not be granted rights and privileges equal to human rights: A/IS should always be subordinate to human judgment and control.* (14)

> *Legal accountability has to be ensured when human agency is replaced by the decisions of AI agents. (28)*

> *The true operator of an AI system must be made known to the public.* (15)

### 4.3.5. Relevance to HR analytics

Many guidelines analyzed here can be interpreted as requiring that organizations, which implement AI technology to assist their HR decisions, should implement a system for monitoring its effects on employees and a purposely designed ethical governance system. One might summarize the more specific recommendations about such monitoring and governance found in the twenty guidelines analyzed here by the following list:

a) make available an API to allow the research community to query the algorithm used for predictive and prescriptive HR decisions;

b) set up an independent ethics governance process;

c) develop a process by which employees can correct errors in input data or in output decisions;

d) set up a mechanism allowing employees to challenge decisions based solely on automated processing of their information;

e) make visible avenues for redress for those affected by disparate impact and other discriminatory outputs;

f) keep records of the role of algorithmic recommendations and predictions in HR decisions.

The first recommendation may appear unrealistic. Organizations are unlikely to create automatic means to facilitate the job of researchers who may publicly criticize the algorithms they use as discriminatory, generating a steady risk of reputational losses. The second recommendation could be acceptable in some contexts in which AI is used (e.g. medicine) but is quite radical in the context of HR. Most likely it would make AI in HR too expensive and hence not worth the investment. More plausibly, large companies are well advised to build an *internal* ethics committee to review the introduction of AI in industry processes with significant impact on their employees. Of course, the feasibility of such ideas even for a large company depends on how such ethics committees are expected to operate. Should the ethics committee review all the HR decisions taken with the help of AI? And if so, how would such a review differ from the ordinary work of HR experts within the company? Rather, the idea of an ethics board is more plausible if it is conceived as a board that meets in order to plan the introduction of AI in HR, in particular what needs to be done to make it fair and intelligible. The ethics board should also clearly specify what steps are to be taken in order to monitor the behavior of the AI in operation.

Following most recommendations examined here, such a board should be transdisciplinary and include representatives of different departments, e.g. management, data science, and all the various functions directly or indirectly affected by the innovation. Ideally it may also involve an external expert of AI ethics. The idea of a process allowing workers to correct errors in data (the data which are used to make HR predictions and decisions about them) is implied by data protection law. The idea that employees should be able to challenge AI decisions may be understood as the idea that *operators of AI* in HR should be able to challenge AI prescriptions. This is not only plausible but most likely part of the way in which AI in HR analytics will be implemented everywhere. It is difficult to conceive of a concrete use of AI in HR analytics, in particular for prescriptive analytics, that would lead to the full automation of HR decisions. The idea that workers subjected to AI decisions should be able to challenge HR decisions made about them is not only meaningful, but basically presupposed by labor law, at least in most EU countries.

Theoretically, a redress mechanism for the mistreatment of employees due to the use of AI in HR analytics is implied by labor law protections (of, at least, formal employees), in most EU countries. But it may not be easy for employees to *obtain* redress when they deserve it. E.g. workers should have the legal right to challenge and overturn an unfair decision, irrespective of the role played by AI in the underlying motivation. This may be followed by some form of compensation for the worker. It is unclear, however, how effectively people affected by algorithmic decisions will be able to obtain redress as, for example, algorithmic discrimination may be hard to demonstrate.

Finally, the recommendation to keep records of algorithmic recommendations, including the employee data and software leading to them seems a realistic one for enhancing the accountability of the individuals behind the AI. Such records may enable a proper evaluation of individual unfair decisions, and also of the general flaws in the underlying software.

# 5. Conclusion

This conclusion is based on key ideas found in guidelines, expressed in a general form in the preceding chapters. We apply the recommendations of the twenty guidelines on AI and algorithms considered here to the domain of HR analytics, coming up with three different general recommendations:

1. GDPR+: Rules for data collection for HR analytics should go beyond GDPR

2. The development of data-driven (AI) HR tools needs adequate technical competence to generate knowledge about the algorithm

3. The impact of using the tool on employees should be carefully monitored

4. Adequate transparency about algorithmic decisions shall be identified and implemented.

I explicate each of these recommendations in three steps:

1. By extracting key implementation questions,

2. by developing practical recommendations that, when followed, would make the use of AI in HR more ethical,

3. by mapping how these recommendations relate to the key (substantive) ethical values of beneficence, non-maleficence, justice and fairness, and autonomy.

The summary follows the three-part division of topics, into data collection/storage/access, the development of algorithmic tools and the assessment of their impact.

## 5.1. GDPR+: Rules for data collection for HR analytics should go beyond GDPR

The first requirement of ethical AI is to achieve the highest level of data protection. This concerns the employee data used to train AIs and the individual employee data on the basis of which the AI makes a recommendation concerning her in particular. There is a clear overlap between AI guidelines concerning data collection and the legal principles of data protection. The GDPR legislation by the EU, being a recent and comprehensive one, could provide an adequate legal standard for many organizations to follow, which also takes into consideration the trade-offs between different values and goals pursued by organizations. But the GDPR is not sufficient. A GDPR+ approach could be considered as an ethical approach that complements respect for data protection with the principles of stakeholder engagement (4.3.3) and governance structures to promote accountability (4.3.4 and 4.3.5).

### Key implementation questions

The key ethical questions to be answered before implementing data collection are:

— *Are we guaranteeing the strictest level of privacy protection for the employees compatible with the goals of such analytics?*

— *Is the purpose for which we ask for employee data one that we can justify and coherent with our mission as an organization?*

— *Is it possible to engage employees as actors of data-driven algorithmic governance, not as passive recipients of algorithmic decisions?*

**Key implementation steps**

Practical recommendations to implement ethical process improvements are:

1. *Build an internal or independent ethics board* to carefully assess the *purposes* of data collection from an ethical point of view. An internal ethics board should include all the key competences in the organization (e.g. including, but not limited to, HR, data protection, and compliance). An independent ethics board should include stakeholders able to provide different perspectives.

2. *Engage employees to provide their opinion.* The voice of employees considering the purposes for which their data are analyzed should not be ignored. An employee discussion panel, which can also involve trade-union representatives as facilitators, can help identify ways of using data in HR that are regarded as more problematic and those that are regarded as more desirable.

**Key ethical values**

From the point of view of the substantive values considered here (3.1.5), the substantive ethical goals and constraints of this stage of implementation of AI are the following:

*Beneficence:* Ideally, AI in HR analytics should benefit all employees, not only their employers. The best way to collect data for AI is to show employees ways in which AI can make their workplace better *for them.* HR analytics should be based on data that employees provide because they are engaged to do so and committed to the goals for which AI is introduced. In order to engage employees, the goals of the HR analytics should be transparent and its benefits for employees should be clearly outlined. A mechanism should be in place to collect not only employee data, but also the employee's opinion about legitimate and illegitimate ways of using them in HR.

*Non maleficence*: HR analytics should not be used to harm employees. Strong privacy and cyber-security protection ought to be in place to provide a strong assurance that 1) the data will only be used for the purposes that have been declared; 2) misuses of the data and data breaches can be avoided. HR analytics should not be used to impose serious punishment and penalties (e.g. dismissals, salary reductions) on the basis of automated decisions or decisions where the human input is merely "formal", i.e. limited to acknowledging the recommendation of the software.

*Justice and Fairness:* Employees who are not willing to be engaged as data subjects for HR analytics should be treated fairly. While they may be excluded from some advantages intrinsically related to the willingness to provide data for analytics (e.g. receiving data-driven personalized feedback and predictions) any further (i.e. avoidable) disadvantage for people who are less comfortable sharing their data must be avoided.

*Autonomy*: Steps must be taken to share data-driven insights and predictions with employees. Employees should always be put in a position in which they a) can learn something useful for themselves from data-driven insights and predictions, b) can respond to a data-driven prediction or assessment by positively changing their behavior. The overall ethical goal, related to autonomy, is to ensure that employees do not become *passive* subjects of algorithmic governance, but can actively contribute to enhancing the organization's performance by taking advantage of data-driven insights that AI in HR may produce.

## 5.2. The development of data-driven (AI) HR tools needs adequate technical competence to generate knowledge about the algorithm

Generating adequate models learned from data *ethically* most fundamentally implies dealing with the challenge of explainability and fairness of the algorithms. To a significant extent, the intelligibility and fairness of a model can be understood, known, documented, and improved through technical methods. The required technical expertise should be complemented with less formalized insights, e.g. about what is morally adequate *in the context,* derived from stakeholders, that may need to be engaged for this purpose.

One challenge for companies is to improve their ability to understand the impact of AI decisions on specific groups, or the data that may be indirectly discriminatory. It is possible that more ethnically diverse and gender balanced research teams have a higher sensitivity to these issues. Even more crucially, there ought to be cognitive diversity, in particular there should be a need for experts to "think different(ly)" compared to most computer scientists (which includes computer scientists who have acquired competences in ethics). Research will hopefully produce insights on the effectiveness of specific hiring and diversity strategies for AI.

### Key implementation questions

The key ethical questions that should be answered before this stage of data processing are:

- *Do the human resources managers in the organization possess the technical skills required to adequately assess the explainability and fairness of AI tools?*
- *Do data scientists collect sufficient information about the training of AIs to be able to trouble-shoot them if they do not work appropriately? Are the potential flaws and limits of the training data sets adequately understood? Are the technical steps for assessing and documenting fairness and intelligibility already in place?*
- *Is there sufficient cognitive and moral/political diversity among the evaluators of AI to assess fairness and transparency issues in a critical way?*
- *Is the organization equipped to collect inputs and feedback from experts outside the organization, that are relevant to assess the fairness and intelligibility of AI tools?*

### Key implementation steps

Practical recommendations to implement ethical improvements are:

1. *Ensure you have adequate competences to build and implement AI ethically.* Organizations that aim to produce and implement AI tools in HR must recruit experts with the range of skills (and informal knowledge, and cognitive styles) required to evaluate an algorithm's intelligibility and fairness.

2. *Involve civil society organizations and expertise from academia.* Seek external validation that the technical methods you employ to train your models are scientifically sound and ethically defensible. Develop awareness of possible ethical criticism by collecting and attracting feedback from outside the organization. If possible, make available (anonymized) datasets and your algorithms for researchers to independently audit your tool.

3. *Engage in AI and ethics education.*

   a. Educate potential end-users (e.g. HR professionals) to ensure that they have the know-how and skills necessary to operate AIs in HR correctly. End-users should not have blind faith in AI tools, but the adequate level of trust combined with critical attitudes. Engage intended users of your tool to ensure that they know enough about the tool, its limits, and its proper domain of application. Promote educational resources that help the personnel in HR departments to avoid misconceptions about AI in HR.[47]

   b. Educate employees potentially affected by AI or their representatives (e.g. labor councils) to ensure that they have the know-how and skills necessary to correctly understand how AI is used in HR. They should not have blind faith in AI tools, but the adequate level of trust combined with critical attitudes. Engage intended affected individuals or their representatives of your tool to ensure that they know enough about the tool, its limits, and its proper domain of application. Promote educational resources that help them avoid misconceptions about AI in HR.

47 For example, data scientists may provide analyses and examples of situations in which the tool is misused.

**Key (substantive) ethical values**

From the point of view of the values considered here (3.1.5), the substantive ethical goals and constraints of this stage of implementation of AI are the following:

*Beneficence:* The tools produced should provide value to the organization and to its employees who are willing to be assessed based on their personal data.

*Non-maleficence*: The risks deriving from model inaccuracies (e.g. decisions based on wrong predictions) should be carefully evaluated. Once the risks are carefully understood, they should be weighted with the potential benefits. In particular, one should also weigh the benefits from deploying a mechanism, which may perform better (both with respect to fairness and with respect to efficiency) than the procedure already in place. Employees should be engaged to better understand their views about the risks and benefits of such tools.

*Justice and fairness:* The fairness of the tool should be assessed by using state of art methodologies and an adequate mix of technical (statistics based) and non-technical (psychological or philosophical) approaches.

*Autonomy*: In order to contribute to human autonomy, the logic behind the tool used for HR assessments and recommendations must be understood. This is a mix of proper scientific procedures that allow one to reconstruct how the tool learned to make recommendations the way it does. For example, in the case of statistical learning, the data, training method, and specification of the utility/loss function of the algorithm should be noted down. Moreover, it should be possible to provide explanations that allow a meaningful and constructive debate between data scientists and HR experts with different forms of domain knowledge. For example, it may be useful to have a way to explain individual recommendations which have been made by an AI. Such explanations may be less precise than those used by data scientists, but they play an important role nonetheless.

## 5.3. The impact of using the tool on employees should be carefully monitored

It should be possible to monitor and document the actual effect of using AI to assist HR decisions. This can be achieved by implementing technical procedures (e.g. automatically collecting data about decisions taken with the inputs from AIs) and by implementing social processes, e.g. the possibility for HR employees to provide feedback and discuss the AI's outputs, also with its creators.

**Key implementation questions**

The key ethical questions that should be answered before this stage of data processing are:

- *Do we have mechanisms in place to ensure that the decisions taken algorithmically can be assessed and corrected if anything goes wrong?*
- *Are our procedures to document the decisions taken by algorithms adequate? Can we make them so without compromising the privacy of our employees?*
- *How can we establish a mechanism for safe feedback, which does not hinder the functioning of the HR tool, but allows the data scientists to improve it step-by-step?*
- *Is there a mechanism for compensating individuals who are treated unjustly due to an inaccurate algorithmic assessment or prediction?*

**Key ethics implementation steps**

The key organizational steps that should be in place to implement the guidelines at this stage of data processing, according to the twenty guidelines analyzed here, are the following:

1. *Develop an (adequately privacy protected) mechanism to record high-stake decisions about employees that are made with the help of algorithmic recommendations or predictions.*

2. *Develop an (adequately privacy protected) mechanism to determine if particular populations (e.g. non-native speakers, pregnant women, minority members, etc.) are negatively or positively affected by algorithmic decisions.*

3. *Develop a process by which employees can correct errors in input data or outputs.*

4. *Develop a process by which employees can challenge decisions that are fully automated (if any).*

5. *If a subgroup of your employees (e.g. non-native speakers, parents with kids, members of religious groups, etc.) appears to be systematically disadvantaged by the introduction of algorithmic decisions, set up a procedure of redress, or even better, improve the outcome for them in the long term.*

**Key (substantive) ethical values**

From the point of view of the substantive values considered here (3.1.5), the ethical purposes of this stage can be characterized as follows:

*Beneficence:* It should be possible to support any claim about the positive impact of AI in HR analytics in the workplace with data. When the data do not reveal any benefit from the use of such tool, the use of the tool should be reconsidered. When the data necessary to make such an assessment do not exist, they should be produced.

*Non-maleficence*: There must be procedures in place to detect problems caused by algorithmic decisions, which can be based on inaccurate decisions, or, on fear and rejection of such tools (that may also be irrational). One way to achieve this is to have a procedure to enable employees to contest and criticize HR decisions taken with the aid of algorithms, and receive explanations about them.

*Justice and fairness:* Besides technical evaluations routinely made when training a tool, the fairness of AI must be assessed *in relation to real-world outcomes.* For this purpose, it is important to identify those groups of employees which may be adversely affected by such a tool. The impact on employees of different groups should be measured and assessed. Groups that are suffering from disadvantages, or are not benefiting from the introduction of such tools, as other groups, should be given ways to improve their situation.

*Autonomy*: Fully automated decisions in the field of HR should be avoided as a rule. If they exist, a procedure for contestation of such decisions has to be in place.

## 5.4. HR and management should guarantee adequate transparency about the data-driven tools used in HR

The value of transparency, as the preceding analysis shows, is considered by all the guidelines examined here about AI. However, transparency will only be beneficial if it is designed in the right way, otherwise it will backfire. It will be beneficial when a better understanding of the logic of the algorithm provides workers with incentives to work better. This result may be possible, but it is not guaranteed to be obtained simply because the HR algorithm is based on statistically accurate statistical models. Ill-designed transparency can backfire in two ways: firstly, employees may exploit transparency to game the metrics that are used to make decisions about them, in order to gain personal advantage and in ways that are not good for the company; and secondly, there is risk of a perversion of the goals pursued by employees, if employees maximize their scores based on proxies of performance and excellence, instead of aiming for authentic improvement which produces good scores as side-effects.[48] More generally, if AI derives from machine learning models which are blind to strategic consideration, its predictive power may be undermined, as knowledge of the algorithm affects workers' incentives and thus leads to new

48 As discussed in the section "Open challenges for algorithmic transparency and accountability", above.

patterns of behaviors (different from those producing the data used in training the AI). It should be emphasized, however, that employees and managers may attempt to game even non-transparent AI, based on guesswork or "company myths" about the way the inscrutable AI works. Hence, non-transparent AI may also generate distortive effects, undermining predictions, and it is unclear that more transparent AI will always lead to worse attempts to game the system. One scientific way to avoid such distortions is to design *strategy-proof* AI (i.e. decision rules that are dominant-strategy-incentive-compatible in the game-theoretic sense). When subjected to a strategy-proof known decision rule, then, by definition, every worker's selfish best interest is best served by reporting truthfully the information concerning him or her. This may require more advanced machine learning models which are based not only based on the principles of statistics, but also on principles of economics, in particular game theory (77).

Indeed, because of its complexity, an adequate transparency strategy needs to be planned, designed, and achieved with different solutions tailored for different contexts. Different kinds of transparency should be addressed to the right kind of stakeholders (e.g. workers' representatives and HR managers in the case of HR), by engaging management in designing a transparency strategy side-by-side with the introduction of AIs. It is also possible that different forms of transparency should be used simultaneously. For example, the documentation of the machine learning process may be sufficient for some purposes and some stakeholders, but useless for others; detailed knowledge of the features and how they affect the outcome may be fair, useful and possible in some contexts (e.g. where the algorithms are not "black boxes" and knowledge of them does not deliver perverse objectives), but a higher-level explanation of the logic of the decisions involved, and full knowledge about the features, may be more appropriate in other cases (e.g. where detailed knowledge would be used to game the system).

# 6. References

1. Peck D. They're Watching You at Work. The Atlantic. 2013 Nov 20 [accessed 2019 Feb 20]; Available from: https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/

2. Walker J. Meet the New Boss: Big Data. Wall Street Journal. 2012 Sep 20 [accessed 2019 Sep 4]; Available from: https://www.wsj.com/articles/SB10000872396390443890304578006252019616768

3. Sharp R. Virtual career assistants can solve coaching challenges. Hrmagazine. [accessed 2019 Sep 4]. Available from: http://www.hrmagazine.co.uk/article-details/virtual-career-assistants-can-solve-coaching-challenges

4. Pape T. Prioritising data items for business analytics: Framework and application to human resources. European Journal of Operational Research. 2016 Jul 16;252(2):687–98.

5. Moore P, Robinson A. The quantified self: What counts in the neoliberal workplace. New Media & Society. 2016 Dec 1;18(11):2774–92.

6. Davenport TH, Harris J, Shapiro J. Competing on talent analytics. Harvard business review. 2010;88(10):52–58.

7. Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L. The Ethics of Algorithms: Mapping the Debate. Big Data & Society. 2016 Nov 1;3(2):2053951716679679.

8. Jobin A, Ienca M, Vayena E. Artificial Intelligence: the global landscape of ethics guidelines. Nat Mach Intell. 2019 Sep;1(9):389–99.

9. Loi M, Heitz C, Ferrario A, Schmid A, Christen M. Towards an Ethical Code for Data-Based Business. In: 2019 6th Swiss Conference on Data Science (SDS). Bern, Switzerland: IEEE; 2019 [accessed 2019 Sep 2]. p. 6–12. Available from: https://ieeexplore.ieee.org/document/8789855/

10. Deutsche Telekom. AI Guidelines. 2018. Available from: https://www.telekom.com/resource/blob/532446/f32ea4f5726ff3ed3902e97dd945fa14/dl-180710-ki-leitlinien-en-data.pdf

11. Women leading in AI. 10 Principles of responsible AI. [accessed 2019 Dec 10]. Available from: http://womenleadinginai.org/wp-content/uploads/2019/02/WLiAI-Report-2019.pdf

12. Fairness, Accountability, and Transparency in Machine Learning (FATML). Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. 2016. Available from: https://www.fatml.org/resources/principles-for-accountable-algorithms

13. Partnership on AI. Tenets. 2016. Available from: https://www.partnershiponai.org/tenets/

14. Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, version 2. 2017. Available from: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

15. The Public Voice. Universal Guidelines for Artificial Intelligence. 2018. Available from: https://epic.org/international/AIGuidleinesDRAFT20180910.pdf

16. ICDPPC. Declaration on ethics and data protection in Artificial Intelligence. 2018. Available from: https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf

17. W20. Artificial Intelligence: open questions about gender inclusion. 2018. Available from: http://webfoundation.org/docs/2018/06/AI-Gender.pdf

18. Leaders of the G7. Charlevoix Common Vision for the Future of Artificial Intelligence. 2018. Available from: https://www.mofa.go.jp/files/000373837.pdf

19. Access Now ; Amnesty International. The Toronto Declaration: Protecting the right to equality and nondiscrimination in machine learning systems. 2018. Available from: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

20. Privacy International & Article 19. Privacy and Freedom of Expression In the Age of Artificial Intelligence. 2018. Available from: https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf

21. WEF, Global Future Council on Human Rights 2016-2018. White Paper: How to Prevent Discriminatory Outcomes in Machine Learning. 2018. Available from: http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

22. Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:180207228 [cs]. 2018 Feb 20 [accessed 2019 Jun 17]; Available from: http://arxiv.org/abs/1802.07228

23. Villani C. For A Meaningful Artificial Intelligence: Towards A French And European Strategy. 2018 Mar. Available from: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

24. UNI Global Union. Top 10 Principles for Ethical Artificial Intelligence. 2017. Available from: http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

25. Information Technology Industry Council (ITI). ITI AI Policy Principles. 2017. Available from: https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf

26. Software & Information Industry Association (SIIA), Public Policy Division. Ethical Principles for Artificial Intelligence and Data Analytics. 2017. Available from: http://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990

27. COMEST/UNESCO. Report of COMEST on Robotics Ethics. 2017. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000253952

28. Internet Society. Artificial Intelligence and Machine Learning: Policy Paper. 2017. Available from: https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AI-Policy-Paper_2017-04-27_0.pdf

29. Independent High-Level Expert Group On Artificial Intelligence Set Up By The European Commission. Ethics guidelines for trustworthy AI. European Commission - Digital Single Market; 2019 [accessed 2019 Apr 9]. Available from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

30. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019 Sep;1(9):389–99.

31. Schroeder M. Value Theory. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Fall 2016. Metaphysics Research Lab, Stanford University; 2016 [accessed 2019 Oct 11]. Available from: https://plato.stanford.edu/archives/fall2016/entries/value-theory/

32. Santoni de Sio F, Van den Hoven J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Front Robot AI. 2018 [accessed 2019 May 27];5. Available from: https://www.frontiersin.org/articles/10.3389/frobt.2018.00015/full

33. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. 6. ed. New York: Oxford University Press; 2008.

34. Wachter S, Mittelstadt BD. A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI. Columbia Business Law Review. 2018.

35. FAT ML. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. no date [accessed 2019 Jan 29]. Available from: http://www.fatml.org/resources/principles-for-accountable-algorithms

36. Lippert-Rasmussen K. Nothing Personal: On Statistical Discrimination*. Journal of Political Philosophy. 2007 Dec 1;15(4):385–403.

37. Custers B, Calders T, Schermer B, Zarsky T, editors. Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases. 2013 edition. New York: Springer; 2012.

38. Barocas S, Selbst AD. Big data's disparate impact. California Law Review. 2016;104(671):671–732.

39. Zafar MB, Valera I, Rodriguez MG, Gummadi KP. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. arXiv:161008452 [cs, stat]. 2017;1171–80.

40. Lipton ZC. The Mythos of Model Interpretability. arXiv:160603490 [cs, stat]. 2016 Jun 10 [accessed 2018 Sep 18]; Available from: http://arxiv.org/abs/1606.03490

41. Kroll JA. The fallacy of inscrutability. Phil Trans R Soc A. 2018 Nov 28;376(2133):20180084.

42. Kroll JA, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H. Accountable Algorithms. U Pa L Rev. 2016 2017;165:633.

43. Selbst AD, Barocas S. The intuitive appeal of explainable machines. Fordham L Rev. 2018;87:1085.

44. Loi M, Ferrario A, Viganò E. Transparency As Design Publicity: Explaining and Justifying Inscrutable Algorithms. Rochester, NY: Social Science Research Network; 2019 Jun [accessed 2019 Jul 5]. Report No.: ID 3404040. Available from: https://papers.ssrn.com/abstract=3404040

45. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:160204938 [cs, stat]. 2016 Feb 16 [accessed 2018 Sep 18]; Available from: http://arxiv.org/abs/1602.04938

46. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR.(2017). Harvard Journal of Law & Technology. 2017;31:841.

47. Pasquale F. The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press; 2015.

48. Lipton ZC, Chouldechova A, McAuley J. Does mitigating ML's impact disparity require treatment disparity? arXiv:171107076 [cs, stat]. 2017 Nov 19 [accessed 2018 Dec 14]; Available from: http://arxiv.org/abs/1711.07076

49. Ferrario A, Loi M, Viganò E. In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. Philos Technol. 2019 Oct 23 [accessed 2019 Oct 28]; Available from: https://doi.org/10.1007/s13347-019-00378-3

50. Ferrario A, Loi M, Viganò E. In AI we trust incrementally. A multi-layer model of trust to analyze human-artificial intelligence interactions. Philosophy and Technology. accepted;

51. Colquitt JA, Zipay KP. Justice, Fairness, and Employee Reactions. Annu Rev Organ Psychol Organ Behav. 2015 Apr 10;2(1):75–99.

52. Colquitt JA, Wesson MJ, Porter COLH, Conlon DE, Ng KY. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. Journal of Applied Psychology. 2001 01;86(3):425–45.

53. Colquitt JA, Scott BA, Rodell JB, Long DM, Zapata CP, Conlon DE, et al. Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives. Journal of Applied Psychology. 2013;98(2):199–236.

54. Thibaut JW, Walker L. Procedural Justice: A Psychological Analysis. L. Erlbaum Associates; 1975. 150 p.

55. Binns R, Kleek MV, Veale M, Lyngs U, Zhao J, Shadbolt N. "It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions. SocArXiv. 2018 Jan 31 [accessed 2018 Mar 29]; Available from: https://osf.io/preprints/socarxiv/9wqxr/

56. Hammarfelt B, de Rijcke S. Accountability in context: effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the faculty of Arts at Uppsala University. Res Eval. 2015 Jan 1;24(1):63–77.

57. Sousa SB, Brennan JL. The UK Research Excellence Framework and the Transformation of Research Production. In: Musselin C, Teixeira PN, editors. Reforming Higher Education: Public Policy Design and Implementation. Dordrecht: Springer Netherlands; 2014 [accessed 2019 Oct 27]. p. 65–80. (Higher Education Dynamics). Available from: https://doi.org/10.1007/978-94-007-7028-7_4

58. Müller R, de Rijcke S. Thinking with indicators. Exploring the epistemic impacts of academic performance indicators in the life sciences. Res Eval. 2017 Jul 1;26(3):157–68.

59. Dalen HP van, Henkens K. Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. Journal of the American Society for Information Science and Technology. 2012;63(7):1282–93.

60. Baccini A, Nicolao GD, Petrovich E. Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. PLOS ONE. 2019 Sep 11;14(9):e0221212.

61. Azevedo AIRL, Santos MF. KDD, SEMMA and CRISP-DM: a parallel overview. In: Weghorn H, Abraham AP, editors. IADIS Proceedings of Informatics 2008 and Data Mining 2008. Amsterdam: IADIS; 2008. p. 182–5.

62. Bogen M. All the Ways Hiring Algorithms Can Introduce Bias. – Google Search. Harvard Business Review Digital Articles. :2–4.

63. Bogen M, Rieke A. Help wanted – an exploration of hiring algorithms, equity, and bias. Upturn; 2018.

64. Loukina A, Madnani N, Zechner K. The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Florence, Italy: Association for Computational Linguistics; 2019 [accessed 2019 Aug 21]. p. 1–10. Available from: https://www.aclweb.org/anthology/W19-4401

65. Gilbert DE. Luck, Bayesian Tensor Completion, and Fairness [A Dissertation in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy]. [Ithaca, NY 14850, United States]: Faculty of the Graduate School of Cornell University; 2019.

66. Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research. 2018 Jul 2;0049124118782533.

67. Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. arXiv:161002413 [cs]. 2016 Oct 7 [accessed 2017 Nov 13]; Available from: http://arxiv.org/abs/1610.02413

68. Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:161007524 [cs, stat]. 2016 Oct 24 [accessed 2018 Jan 17]; Available from: http://arxiv.org/abs/1610.07524

69. Brennan T, Dieterich W, Ehret B. Evaluating the predictive validity of the COMPAS risk and needs assessment system. Criminal Justice and Behavior. 2009;36(1):21–40.

70. Kleinberg J, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:160905807 [cs, stat]. 2016 Sep 19 [accessed 2017 Nov 13]; Available from: http://arxiv.org/abs/1609.05807

71. Angwin J, Larson J. Machine Bias. ProPublica. 2016 [accessed 2018 Mar 14]. Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

72. Heidari H, Loi M, Gummadi KP, Krause A. A Moral Framework for Understanding Fair ML Through Economic Models of Equality of Opportunity. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM; 2019 [accessed 2019 Feb 22]. p. 181–190. (FAT* '19). Available from: http://doi.acm.org/10.1145/3287560.3287584

73. Binns R. On the Apparent Conflict Between Individual and Group Fairness. arXiv:191206883 [cs, stat]. 2019 Dec 14 [accessed 2020 Jan 17]; Available from: http://arxiv.org/abs/1912.06883

74. Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, version 1. 2019. Available from: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents

75. Jamieson KH. Cyberwar: how Russian hackers and trolls helped elect a president what we don't, can't, and do know. New York, NY: Oxford University Press; 2018.

76. Beitz CR. The Idea of Human Rights. Oxford; New York: Oxford University Press; 2009.

# People Analytics must benefit the people. An ethical analysis of data-driven algorithmic systems in human resources management

Dr. Michele Loi,
2 March 2020

Publisher:
AW AlgorithmWatch gGmbH
Linienstraße 13
10178 Berlin
Registered Office: Bergstr. 22, 10115 Berlin

Contact: info@algorithmwatch.org

Proofreading:
Graham Holliday

Layout:
Beate Autering
Tiger Stangl
www.beworx.de

Published as a part of the research project
**Automatisiertes Personalmanagement und Mitbestimmung (Automated Human Resources Management and Labor Rights)**

Website: algorithmwatch.org/en/project/auto-hr/

Funded by

Hans **Böckler** **Stiftung**