WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

Briefing paper

Employment Law and AI-Based Recruitment:
A Close Examination of Existing Regulatory Gaps and the Path Forward

May 2020

World Economic Forum
91-93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland
Tel.: +41 (0)22 869 1212
Fax: +41 (0)22 786 2744
Email: contact@weforum.org
www.weforum.org

Contents

Executive summary

There has been an explosion in recent years of artificial intelligence (**AI**)-based tools for human resources (HR) applications. These tools are designed to assist with or perform key HR tasks including hiring, retaining talent, training, and managing benefits and employee satisfaction. If developed and deployed correctly, AI-based tools have the potential to boost employee productivity, save HR departments time and money, and improve fairness and diversity outcomes.

At the same time, many observers and commentators both inside and outside of government have raised legal and ethical concerns regarding the use of these technologies. In a white paper on AI (*EU White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*, the "**EU White Paper**"), the European Commission identified the use of AI applications for recruitment processes as well as in situations affecting workers' rights as "high risk". There are concerns about AI algorithms encoding bias and discrimination. The opaqueness of certain decision-making functions can undermine employee trust, leading to lower productivity and job satisfaction. In addition, unique aspects of the human resources setting, including small datasets, complex social interactions and data privacy concerns, pose challenges to developing effective algorithms.

The EU White Paper lays the foundation for a new legal framework, including legislation and *ex-ante* assessments, to regulate the use of AI in (among others) the field of HR. In the United States, new legislation has started to emerge to address some of the risks associated with the use of AI in this domain. Most recently, New York City has been considering a bill (Intro No. 1894–2020) that would prohibit the sale of an AI hiring tool unless it had been audited for bias in the year prior to sale and includes an annual bias audit.

This briefing paper looks into the most pressing concerns, first by identifying four major areas relevant to AI in HR where regulatory gaps currently exist (transparency and explainability; bias and fairness; data privacy; and liability), and then by examining, critiquing and proposing modifications to the framework proposed in the EU White Paper.

In terms of regulatory gaps, there is general agreement that *transparency and explainability* are key principles for developing AI systems that make legally significant decisions, such as those in employment settings. However, no comparable consensus exists regarding how these principles should be translated into legal mandates. Indeed, most US states and other non-EU jurisdictions have no specific laws requiring employers to disclose even the existence of automated decision-making in the employment process.

Laws prohibiting discrimination make the legal landscape for *bias and fairness* somewhat better defined. But anti-discrimination laws take many forms, both in terms of the protected groups they cover and the types of decisions and outcomes they prohibit, and can often pull employers in different directions when implementing new methods of employee selection. These laws have not yet been tested in the context of AI-based tools, making it difficult for employers to determine what steps they must take to ensure legal compliance.

The landscape for *data privacy* is evolving rapidly with the EU's adoption of the General Data Protection Regulation ("**GDPR**") and a growing movement for similar legislation in other jurisdictions. Commentators and policy-makers are debating how to strike the right balance between individual privacy and labour-market innovation, and even the GDPR is less than definitive, requiring employers to consider numerous factors when deploying HR automation without providing much guidance on which factors will be determinative. As a result of these and other regulatory gaps, there is considerable uncertainty on the issue of *liability*, including the allocation of damages between different companies involved in the development, deployment and use of HR tech.

Section 2 of this paper then proposes measures to enhance opportunities and mitigate risks, taking into account the policy options proposed by the European Commission and elsewhere.

On transparency and explainability, we recommend that policy-makers adopt policies that encourage employers to strike a thoughtful balance between completeness and interpretability, with regular auditing to ensure that employers do not manipulate these requirements to obfuscate the nature of the tools they use. We similarly recommend a nuanced approach to bias and fairness that takes into account both legitimate qualifications for a position and the need to guard against (and, where necessary, correct) systemic biases that may have disadvantaged members of certain social and demographic groups. Employers should continue to have the ability to use tools that have a statistical disparate impact so long as they are able to justify the differences through validation and conduct regular audits to ensure continued compliance.

In terms of data privacy, we recommend that policy-makers encourage companies to find ways to exchange data in a secure and individually anonymized manner so that companies can develop innovative new data-powered tools without infringing on legitimate individual privacy interests. Finally, on the subject of liability, we encourage policy-makers to consider adopting joint and several liability rules, under which the developers and vendors of AI tools would share some of the employer's burden if use of an AI tool foreseeably results in unlawful acts. This would allow a wronged party to recover damages from any party that contributed to the unlawful act, while allowing that party to seek indemnity from other involved entities.

1. AI's (key) features and pressing concerns

To better understand the most pressing concerns people (may) have with the use of AI in HR and to identify regulatory gaps, it is important to define AI's key features. The most talked-about subset of AI is "machine learning", a broad category of techniques that use statistical approaches and data to "teach" computers how to identify the factors important to completing a task. Machine learning can train a system to identify patterns and combinations of conditions that are relevant – sometimes even when a human expert is not conscious of them. Deep learning is a particularly complex form of machine learning that can identify subtleties in data that humans cannot discern – and, in some cases, cannot even explain. While these capabilities can be powerful, they also make deep-learning systems opaque, raising concerns around transparency and explainability, and the outputs of deep-learning systems can be difficult to verify, interpret or reproduce.

Machine-learning systems usually require large datasets for training. The degree of access to (large amounts of) data and the quality of this data have a direct impact on the effectiveness and fairness of machine-learning systems, which could (potentially) create issues in terms of: (1) bias and fairness; and (2) data privacy.[1]

From a regulatory perspective, a final feature to note is that the performance of machine-learning algorithms can change in unpredictable ways when they are deployed in new or evolving settings that differ from the initial training data. It is possible to continuously feed in new training data to enable the algorithm to improve and adapt. Either of these approaches, though, could generate liability issues whereby the system's performance changes after a change in ownership.

1.1 Transparency and explainability

Importance of transparent and explainable AI

In many jurisdictions, employment decisions require some form of substantiation. Employers must base their decisions on objective grounds and, if questioned, they should be able to demonstrate the reasonableness of a particular decision. In some circumstances, statutory law or general duty of care provisions could imply a further right to explanation, but the level of detail required varies. The same set of rules applies *mutatis mutandis* to employment decisions based on machine learning. However, as machine-learning models become more complex, it becomes more difficult for humans to retrieve and decipher the rationale of how the algorithm reached its recommendation or decision. These models can become a "black box" where designers cannot explain why the model takes a particular decision or which features it considers while making a decision.

To ensure that employment decisions are not infected by irrelevant or unfair factors, machine-learning tools should be able to explain themselves. Understanding why and how the model works will also enable developers to debug, improve and optimize their models. It could even generate novel scientific hypotheses, which could result in new scientific discoveries. Lastly, explainability will boost confidence in the reliability of the models, which could serve as a catalyst for building trust.

Legal context

As legislators have started to acknowledge the importance of transparency and explainability in machine learning, transparency requirements have become an important part of the legislative initiatives regarding the use of AI. The European Commission's General Data Protection Regulation (**GDPR**) created a "right to explanation" requiring that automated decisions be subject to suitable safeguards.[2] Articles 13–15 of the GDPR require companies to inform data subjects that they are using automated decision-making and to provide *meaningful information* about the logic involved, as well as to explain the significance and envisaged consequences of such processing for the data subject.

In the US, Illinois' Artificial Intelligence Video Interview Act (**AIVIA**) requires employers who use AI analysis of video interviews to inform applicants how the AI works and what general characteristics it uses to evaluate them. Other US jurisdictions are considering similar legislation that would require employers to inform job applicants if they use AI in the hiring process.[3]

Regulatory gaps

Existing legal transparency requirements generally envisage a limited right for data subjects to understand and verify the basic functionality of certain automated decision-making systems. Beyond that minimum threshold, however, the precise contours of the "right to explanation" have been a matter of debate.[4]

The A29 WP Guidelines on Automated Individual Decision-Making and Profiling ("**AIDM Guidelines**"), which the European Data Protection Board has endorsed, give the following examples of *meaningful information*:

– The categories of data that have been or will be used in the profiling or decision-making process

– Why these categories are considered pertinent

– How any data subject's profile is built, including any statistics used in the analysis

– Why this profile is relevant to the automated decision-making process, and

– How the profile is used for a decision concerning the data subject[5]

The AIDM Guidelines stipulate that the data controller should find simple ways to tell the data subject about the rationale behind or the criteria relied on in reaching the decision. The information provided should be sufficiently comprehensive for the data subject to understand the reasons for the decision. This does not necessarily mean a complex mathematical explanation of the algorithms used or disclosure of the full algorithm; but presenting a simplified description of a complex system may be unethical if the target audience cannot understand the limitations of the simplified description (i.e. it is by definition incomplete and thus may lack important details, risking misinterpretation and wrong conclusions).[6]

Also, providing data subjects with information sufficiently comprehensive for them to understand the reasons for a particular decision is harder than it sounds. Model-agnostic or model-specific interpretation methods, which generate a second model that examines the existing model to provide explanations, make it possible to interpret certain complex machine-learning models. However, despite the fact that these methods show great promise, they can be difficult to implement.[7]

Further, identifying the most important feature in the decision-making process does not explain *why* that feature might have predictive value. Commercially available machine-learning algorithms cannot distinguish between causation and correlation. This means that an algorithm might be driven by chance correlations that exist in the training data but not in the world at large. Additionally, an algorithm may discover predictors that *are* correlated with job performance in the real world – but only because they are associated with demographic groups that predominate in the job at question, and not because they actually reflect ability to perform that job.[8] It may therefore not always be sufficient to know *how* a models works. Ideally, a model affecting workers should also be able to explain *why* it performs the way it does.

Another challenge in bridging the gap between interpretability and explainability is that developers, users, data subjects, regulators and auditors will all have different perspectives and demands in terms of the level of explainability. The AIDM guidelines merely focus on data subjects as a target audience, but users and especially regulators and auditors need other (and more detailed) information. The explanation must be relevant and understandable for the specific target audience, which makes a one-size-fits-all transparency approach impracticable.

Finally, when making an AI system more transparent and explainable, one can unintentionally make it less secure, which entails the risks of leaking private data and manipulation.

1.2 Bias and fairness

What is fair?

Employers often assume that AI tools are objective and thus could drive a decision-making process that is free of the biases that affect human judgements. This assumes, however, that the developers of these algorithms, the data on which these tools are built and the organizations in which they are applied are unbiased. In reality, this utopian assumption does not hold. Bias, including bias based on legally protected characteristics such as race and gender, is embedded in human societies. Because AI tools are driven by data extracted from human society, the risk that AI tools will encode and exacerbate existing biases is difficult, if not impossible, to avoid.

In some respects, any bias in AI-driven decision-making processes can be mitigated in ways that are not usually possible with human judgement. A human decision generally does not record all of the various inputs and reasoning processes that go into that decision; in fact, humans might not even consciously *know* all of the inputs that went into an ultimate decision. But algorithmic decision-making processes inherently depend on making any decisional criteria formal and explicit, which creates the potential to detect and remove sources of bias.

To detect (and possibly modify) a biased algorithm, it is first necessary to define what a biased algorithm is. A biased algorithm is one that *unfairly* results in different outcomes for people from different social groups. But that begs the question: What is (un)fairness? There is no simple answer to this question. There are several possible definitions of fairness, and the appropriate definition may depend upon the particular use of the algorithm, the identified goals of the organization using the algorithm and the culture in which it is deployed.

Some definitions of fairness are based on a measure of outcome parity between people from different social groups, usually by requiring the same or similar average outcomes between different social groups (*statistical parity*). Other conceptions of fairness focus on equality in opportunity rather than equality in outcomes. Under this conception, an outcome is fair if the best-performing individuals on certain metrics or tests are picked, even if using those metrics or tests tends to result in statistical disparities between social or demographic groups.

Legal context

In the context of regulating employment decisions, the concept of fairness concentrates on discrimination, which many treaties and constitutions around the world prohibit. In general, anti-discrimination laws proscribe two types of discrimination: (1) direct discrimination (or adverse treatment); and (2) indirect discrimination (or adverse impact). Direct discrimination occurs when one person is treated less favourably than others because of certain protected characteristics (e.g. sex, religion, race etc.). Indirect discrimination occurs when a process results in differential effects or impacts by social group membership, even if these differential effects are unintentional and even if the process does not explicitly consider social group membership.

Scholars have linked these two conceptions of discrimination to the principles of *anti-classification* and *anti-subordination*, respectively. Under an anti-classification view, classifying or differentiating between individuals on the basis of protected characteristics constitutes discrimination. The anti-subordination view holds that the purpose of anti-discrimination laws is to "prohibit practices that enforce the social status of oppressed groups and allow practices that challenge oppression".[9] Laws prohibiting indirect discrimination reflect an anti-subordination perspective because they target policies, procedures or rules that apply equally to everyone but that have the effect of disadvantaging people with certain protected characteristics (typically, characteristics associated with historical patterns of direct discrimination).

Different legal authorities have adopted different standards for determining whether a particular adverse impact is large enough to constitute presumptive discrimination. Some rely on tests of statistical significance, although these have become less meaningful in the age of big data; even small differences between groups will be statistically significant given a sufficiently large dataset. Other authorities, particularly in the United States, have looked to indicia of "practical" significance such as the raw magnitude of the disparities between groups in addition to (or in place of) statistical significance tests. For example, federal agencies use the Uniform Guidelines on Employee Selection Procedures' "four-fifths" rule, under which a selection rate for any race, sex or ethnic group that is less than four-fifths of the rate for the group with the highest rate will generally be regarded as evidence of indirect discrimination.

Relying on such measures of statistical parity could lead to "unfair" situations from an anti-classification perspective if one demographic group, on average, holds higher qualifications or has had higher historical performance outcomes than another group. In such cases, the algorithmic classifier will disadvantage members of the groups that perform better (on average) on the chosen metrics to maintain similar average outcomes. Other fairness definitions therefore (also) take into account a candidate's qualifications and abilities for the position in question, and view fairness through a more individualized lens.

In this view, a fair outcome is one that results in an employer selecting the candidate with the best combination of qualifications, skills and other characteristics relevant to the position in question. Such an approach depends, however, on an employer using adequate and representative criteria for the position in question. "Qualifications"-based approaches also risk reinforcing systemic inequalities and repeating patterns of biased demographic representation. For example, a female candidate might have advanced more slowly in her prior jobs not because of a lack of ability, but because of discrimination or structural barriers to female advancement. Such a candidate might therefore have less impressive measurable qualifications for a

particular job when compared to her male counterparts, even if she would ultimately have proved equally or more capable of performing the job in practice.

Another problem that is often raised with regard to fairness measures is that people may be at the intersection of multiple social groups.[10] Additionally, groups that are not (yet) defined in anti-discrimination law but which may need protecting could fall through the cracks.[11] These contradictory definitions of fairness show the difficulty of solving this problem.[12]

In most jurisdictions, employers can rebut a showing of indirect discrimination by demonstrating that the decision was objectively justified by a legitimate aim, provided that the means of achieving that aim are proportionate and effective.[13] In the context of hiring, the aim of recruiting the best candidates would, in principle, be considered a legitimate aim. To be able to demonstrate that a selection procedure is "effective", an employer should be able to demonstrate that there is a connection between the selection procedure and specific, key aspects of job performance,[14] a process known as *validation*. Validation requires a thorough job analysis, the selection of specific and reasonably objective markers of job performance, and some theoretical or logical relationship between the content of the selection procedure and those markers of job performance.[15] Validation could turn out to be challenging for algorithmic tools, which rely on correlation rather than causation, and particularly challenging for algorithmic selection tools that make use of deep neural networks. The opaqueness of these systems will make it difficult to comply with modern standards of validation.

To be able to demonstrate that a certain selection procedure is proportionate, an employer generally must be able to show that the method chosen was reasonable in light of the available alternatives. If an employee can demonstrate that there was an alternative employment practice available that could have accomplished the same business objective with less discriminatory impact, the selection procedure would be deemed not proportionate. Because most algorithmic selection tools make use of de-biasing techniques (more on that later), it will be relatively easy for a plaintiff to demonstrate that a less discriminatory alternative exists in cases where an algorithmic selection tool does not make use of any de-biasing techniques but does have an indirect discriminatory effect.[16]

Finally, apart from general discrimination laws (as discussed in the foregoing), the GDPR-and AI-specific legislation, such as the New York City bill (Int. 1894–2020) discussed above, includes regulations on bias and fairness. Under the GDPR, personal data must be processed lawfully, *fairly* and in a transparent manner. Recital 71 of GDPR requires the data controller to prevent, inter alia, discriminatory effects on natural persons on the basis of protected characteristics, or processing that results in measures having such an effect. Fairness considerations will have to be part of data privacy impact assessments (**DPIA**, as discussed further in the data privacy section), and if a machine-learning model is *solely* based on automatic decision-making processes or profiling and produces legal or similarly significant effects, article 22 of GDPR requires the implementation of suitable measures to safeguard the data subject's rights, freedoms and legitimate interests. Pursuant to the AIDM Guidelines, these safety measures should include regular quality assurance checks of systems to ensure individuals are being treated fairly and not discriminated against, whether on the basis of special categories of personal data or otherwise.

Regulatory gaps

To improve fairness (and avoid liability), most vendors of algorithmic selection tools test their algorithms for bias. They use testing data[17] and focus on equality of outcomes (i.e. statistical parity) against various social groups. One commonly used de-biasing method is removing or down-weighting features found to be highly correlated with the protected characteristic in question. However, if these features have substantial predictive value, this would ordinarily lead to less consistent outcomes.[18] In addition, because most vendors focus solely on compliance with the fourth-fifths rule,[19] other relevant definitions of fairness might become underexposed.

Another concern regarding de-biasing is that it could trigger claims of "reverse" discrimination. In the United States, courts currently interpret the prohibition against direct discrimination as protecting traditionally dominant demographic groups in addition to the historically disadvantaged groups (such as women and racial/ethnic minorities) that the laws were originally intended to protect. As a result, courts have held that there are limits on the extent to which employers can modify or disregard the results of employee selection procedures to eliminate adverse impacts.[20] It is not clear to what degree employers can use algorithmic de-biasing techniques that make explicit adjustments to the tool's parameters based on protected-class status.

Finally, another (and possibly underexposed) problem of using machine-learning tools for recruitment purposes is that there is a risk that candidates with a wide range of disabilities stand little chance of getting through the recruitment process. In some cases (e.g. video interviews), it might be impossible for disabled people to comply with the standards that the algorithm requires to function properly. In other cases, where there are no direct practical burdens, they will often have a lower chance of being selected by the AI creator as potential top-performers because machine-learning algorithms make predictions based upon common patterns in the training data and candidates with disabilities may be less likely to fit these common patterns.

1.3 Data privacy

Conflicting interests

The use of machine-learning tools could potentially lead to breaches of fundamental rights, including the right to protection of personal data and private life. The fact that the effectiveness and fairness of machine-learning models largely depends on the quality and quantity of data brings challenges for (data) privacy and data protection. A model's accuracy depends on huge amounts of data; this means that building good models might be at odds with one of the key principles of data privacy: "data minimization", which calls for using only information that is relevant or necessary that is relevant or necessary to achieve specific purposes. The biggest challenge for any data privacy legislation is therefore to find a balance between the conflicting interests of innovation and privacy.

Legal context

In regulations across the world, legislators have introduced a wide range of data privacy laws to safeguard the rights of individuals with regard to their personal data. These laws all impose requirements and restrictions on the collection, use and disclosure of "personal information" but have different levels of strictness, with the GDPR likely the most comprehensive and far-reaching of them all. It has an extraterritorial reach, includes significant penalties for non-compliance[21] and gives data subjects more control over their personal data. Moreover, the GDPR fleshes out its automatic processing requirements more than other legislation.

Under the GDPR, the processing of "personal data" is lawful only under specific circumstances (e.g. when processing is necessary for the performance of a contract or for the purposes of the legitimate interests pursued by the controller).[22] Additional restrictions apply to special categories of personal data.[23] Data subjects' rights include the right to information and explanation (as discussed earlier in the transparency section), the right to access their personal data and the right to rectify or erase personal data and object to its processing. The rights to rectification and erasure apply to both the personal data used to create the profile and the profile itself.[24] The right to object should be brought explicitly to the data subject's attention. The data subject can object to processing on grounds relating to his or her particular situation. Once the data subject exercises this right, the controller must interrupt (or avoid starting) the processing process unless it can demonstrate compelling legitimate grounds that override the interests, rights and freedoms of the data subject.

Further, where a type of processing is likely to result in a high risk to the rights and freedoms of natural persons, the controller must carry out an assessment, prior to the processing, of its impact on the protection of personal data. A DPIA is required in cases (among others) that involve systematic and extensive evaluation of the personal data of natural persons that has been derived from automated processing and produces legal or similarly significant effects concerning a natural person. Because GDPR article 35(3)(a) applies to all decisions "based on" automated processing (as opposed to "solely" automated decision-making, which is discussed below), it is assumed that a DPIA is required even for partially automated decision-making with legal or similarly significant effects.

A DPIA includes (at least) a description of the envisaged processing operations and the purposes of the processing, an assessment of the necessity and proportionality of the processing, an assessment of the risks to the rights and freedoms of data subjects, and the measures envisaged to address the risks and demonstrate compliance with the GDPR. As a matter of good practice, a DPIA should be kept up to date in terms of existing processing activities (and amended, if changes are required). However, it should then be reassessed after three years, perhaps sooner, depending on the nature of the processing and the rate of change in the processing operation and general circumstances.

For machine-learning tools that are *solely* based on automatic decision-making processes or profiling and that produce legal or similarly significant effects, additional requirements apply. Pursuant to article 22 of the GDPR, a data subject has the right not to be subject to a decision based *solely* on automated processing, including profiling,[25] that produces legal effects concerning the subject or similarly significantly affects the subject.[26] The AIDM Guidelines stipulate that the controller cannot avoid the article 22 provisions by employing token human involvement. To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful and the human has the authority and competence to change the decision.

The article 22 prohibition rule does not apply when such automated processing: (1) would be necessary for entering into, or performance of, a contract between the data subject and a data controller; or (2) is authorized by European Union (EU) or EU member state law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (3) is based on the data subject's explicit consent.

According to the AIDM Guidelines, the imbalance of power in an employer-employee relationship means that employee consent is free and adequate only when giving or withholding consent will not lead to any adverse consequences.[27] This implies that employers must ensure there is an alternative selection procedure available to employees or applicants who refuse to give consent. As this would remove all of the advantages of such a tool, it is not likely that employers will ever rely on consent with regard to fully automatic decision-making for recruitment purposes.

The AIDM Guidelines also indicate that the contract-related "necessity" exception should be interpreted narrowly. They explicitly mention that using fully automated means to sift out irrelevant applications and to identify fitting candidates with the intention of entering into a contract with a data subject might qualify as "necessary" in situations where there are exceptionally high volumes of applications.[28] It is

not clear, however, what would count as "exceptionally high volumes of applications". In addition, "sifting out irrelevant applications" implies a first screening, which permits the inference that a final screening of fitting candidates would still have to be carried out by a human being. Also, there is always a risk that another effective and less intrusive way to achieve the same goal exists, which would then undercut a claim that the automatic processing was "necessary" for the performance of a contract. These uncertainties make it extremely risky for employers to rely on contract necessity to validate their automated decision-making processing.

Consequently, machine-learning tools for recruitment purposes based *solely* on automatic decision-making processes that produce legal or similarly significant effects are likely to be impracticable and a certain level of human oversight will always be required.

For all decisions based solely on automated decision-making, suitable measures to safeguard the data subject's rights, freedoms and legitimate interests must be in place. The AIDM Guidelines provide the following non-exhaustive list of measures:

- Regular quality assurance checks of systems to make sure individuals are being treated fairly and not discriminated against, whether on the basis of special categories of personal data or otherwise
- Algorithmic auditing – testing the algorithms used and developed by machine-learning systems to prove they are actually performing as intended, and not producing discriminatory, erroneous or unjustified results
- For independent "third-party" auditing (where decision-making based on profiling has a high impact on individuals), providing the auditor with all necessary information about how the algorithm or machine-learning system works
- Obtaining contractual assurances for third-party algorithms that auditing and testing have been carried out and the algorithm is compliant with agreed standards
- Specific measures for data minimization to incorporate clear retention periods for profiles and for any personal data used when creating or applying the profiles
- Using anonymization or pseudonymization techniques in the context of profiling
- Ways to allow the data subject to express his or her point of view and contest the decision
- A mechanism for human intervention in defined cases; for example, providing a link to an appeals process at the point when the automated decision is delivered to the data subject, with agreed timescales for the review and a named contact point for any queries[29]

Regulatory gaps

Data subjects are well protected under the GDPR and the consensus is that data privacy issues will also be addressed more thoroughly in other parts of the world in the years to come. Indeed, the California Consumer Privacy Act (CCPA), enacted in 2018, provides the strongest data privacy protections of any US law passed to date, although it is not as comprehensive as the GDPR. Other US states are likely to enact similar, and perhaps even broader, data privacy protection in the coming years.

Securing personal data can, however, be challenging in the world of AI. Anonymized personal data can sometimes be easily deanonymized by AI (or other technology) and combinations of non-personal data can lead to the identification of persons and/or other sensitive information. In addition, because some machine-learning tools for HR integrate with third-party services, there can be a higher risk of unintentional data leakage.

1.4 Liability

Who will courts and agencies hold liable when an employer uses an AI system in their HR processes that violates some law? This is a major area of legal uncertainty and legal risk both for employers looking to integrate AI into their HR processes and for the vendors who design, develop and sell the underlying technology.

Legal context

Under many relevant laws, such as laws prohibiting discrimination and data-privacy laws prohibiting certain uses of candidate data, the primary responsibility would likely fall on the employer. Such laws, where they apply, will place the onus squarely on employers to ensure that any AI products they use in their HR processes comply with all applicable laws.

A plaintiff might also attempt to invoke the principles of products liability to recover damages from designers or developers of an AI-powered HR tool. Products liability usually holds the designers and sellers of a defective product strictly liable for any harm resulting from the defect. In the United States, products liability is a creature of common law (that is, judge-made law) rather than statute. Courts and plaintiffs rarely invoke products liability in areas of law where statutes specify remedies available to plaintiffs, such as anti-discrimination laws.

Regulatory gaps

It is not clear how far principles of products liability extend to “products” built on machine learning. In some sense, any machine-learning tool that relies on training data provided by the end user is not truly a finished product until the end user supplies that data. As a result, the same algorithm might produce discriminatory results for Employer A but not for Employer B. In such circumstances, it is not clear whether or under what circumstances the courts would apply principles of products liability against the vendor. That said, products liability could provide a potential route to recovery for harms that are not specifically governed by any statute, such as violations of privacy in jurisdictions (including many US states) where there is no specific statute that specifies the privacy rights of employees and applicants.

2. EU White Paper and recommendations

2.1 General comments

Risk-based approach

The AI White Paper proposes a risk-based approach, whereby an AI system will be classified as either high-risk or low-risk, depending on a combination of two factors: sector; and specific use and effect. There may also be some high-risk AI systems irrespective of the sector; the use of AI systems for recruitment processes as well as in situations affecting workers is one example of this.

In our view, the attempt to bifurcate the AI ecosystem into two categories seems to be based on a rather naive view of what AI is. It will require arbitrary line-drawing between sectors and applications, and could lead to situations in which high-risk AI systems in low-risk sectors will not be captured. Also, certain low-risk AI systems used for recruitment purposes could fall under the scope of high-risk restrictions (e.g. the use of AI to select the date and time of candidate interviews, as opposed to deciding who to interview in the first place).

An all-or-nothing approach bears the risk of stifling innovation in certain sectors and leads to disorder in others. We believe that *if* the European Commission were to decide to implement a "risk assessment" framework on an *a priori* basis, such a system should be more gradated. Simply put, a moderate level of compliance requirements should apply to moderately risky AI systems.

However, we do not think that a reliable "risk assessment" framework can be imposed on an *a priori* basis, no matter how fine the gradations. If bright lines are drawn in advance to determine the projected risk level (and therefore the level of legal compliance required), companies will find loopholes that allow higher- risk uses to be treated as lower risk. Moreover, it will often be impossible to determine a risk level for new technology in advance. We endorse the idea that a new legal framework requires companies to take reasonable steps to develop trustworthy AI systems and to deploy them fairly. We also agree that these "reasonable" steps should depend on the level of risk the system imposes in the first place. We believe, however, that the risk level (and therefore the level of legal compliance required) should not be fixed by statute.

AI impact assessment

Instead, we believe that companies themselves should be responsible for determining and monitoring the risk level of the AI tools they develop and/or deploy. AI legislation should require companies to carry out AI impact assessments (**AIIA**s) for every tool that produces legal or similar effects concerning a natural person. AIIA's can be based on the DPIAs as established under the GDPR, on the understanding that the scope is broader and not limited to (data) privacy aspects. A new regulatory framework should provide companies with some flexibility to determine the precise structure and form of an AIIA, so that it fits with existing working practices. The law should, however, prescribe clear (ethical) guidelines and minimum requirements to ensure the adoption of ethical use of AI and to provide a clear basis for auditing.

AIIAs should identify the (possible) impact of the envisaged AI tools on the company, its workers and society, and could address concerns about bias and fairness, explainability and data privacy (as discussed above and as further set out below) at an early stage. Requiring AIIAs to be carried out on a regular basis ensures that other safety measures can be put in place, as soon as the risk level changes. They also provide a foundation for deeper audits of AI tools.

Audits

AIIAs alone cannot be completely effective in preventing adverse effects. AI legislation should therefore also require regular (third-party) audits that can evaluate a company's AI practices retrospectively and help match foresight with hindsight. An agile legal framework based on AIIAs and auditing makes it possible to keep up with rapid technological developments. Further, making the companies themselves responsible for determining the risk level will discourage them from finding loopholes and the "legal uncertainty" that comes with it, and will force them to err on the side of caution, provided that the punishments for failing to take sufficient steps to make the AI safe are sufficiently strong.

2.2 Transparency and explainability

The EU White Paper acknowledges that transparency is paramount. It describes the importance of proactively providing adequate information about the use of high-risk AI systems,[30] but lacks guidance on the level of detail required. We recommend that a machine-learning tool that produces legal or similar effects concerning a natural person should not only be interpretable, but also explainable. This means that identified patterns can be explained on the basis of scientific principles and that employers making use of hiring tools are able to demonstrate that there is a connection between the most relevant or decisive features and specific, key aspects of job performance. If it turns out that the tool in question does not meet the explainability requirements, the tool should not be used.

Information provided to data subjects should be understandable (and can be simplified), but to avoid misinterpretations or misrepresentations, employers should look for a thoughtful balance between completeness and interpretability.[31]

Further, the EU White Paper states that explanatory information should be tailored to each particular context, but it does not distinguish between target audiences. The suggested record-keeping requirements (e.g. documentation on the programming of the algorithm, the data used to train high-risk AI systems and, in certain cases, the keeping of the data itself) seem to be only for testing or inspection by competent authorities.[32] We are of the opinion, however, that this information should also be made available to users to ensure that they can make well-informed decisions before purchasing these products.

Finally, we recommend an audit framework that would not only supervise the extent to which companies comply with the transparency requirements but also ensure that the interpretation methods themselves will be assessed to avoid, *inter alia*, manipulation of model explanations.

2.3. Bias and fairness

The EU White Paper acknowledges the risk of AI exacerbating biases. The suggested record-keeping requirements include documentation on testing and validation in respect of biases,[33] but the EU White Paper does not provide any guidance on *how* AI tools should be tested. We recommend that companies be encouraged to consider multiple definitions of fairness and that an AI tool will be tested on (at least) both disparate impact and differential validity to ensure an acceptable level of fairness. In addition, companies should be required to take into account how these tools might affect individuals with disabilities.

We further recommend that companies be required to perform and report (to users, auditors and possibly individuals) the results of these bias and fairness tests, but, consistent with current anti-discrimination laws, without requiring that employers immediately cease using a model that fails a statistical bias test. This regime allows for the mandating of tests that are not already written into law (e.g. tests for differential accuracy), but also enables employers to deploy a model if they reasonably believe that a certain disparity is "justified" and have taken steps to validate the selection tool in accordance with contemporary standards of industrial and organizational psychology.

Finally, to mitigate the risk that an AI model that scored well in the testing phase produces discriminatory results on deployment, the model should be audited on a regular basis. An annual auditing service funded by the vendor could achieve this goal.[34]

Requiring this level of transparency will incentivize the construction of models that perform well on the *ex-ante* audit, without imposing additional legal constraints that might be unduly burdensome. Of course, existing *ex-post* discrimination protections would have to continue unchanged.

2.4 Data privacy

In general, the GDPR deals well with the (data) privacy challenges that machine-learning tools bring and it seems unlikely that a European regulatory framework covering the use of AI will introduce new restrictions or guidelines on privacy and/or data governance.

Because the GDPR (and other data privacy laws) inhibit organizations from sharing (sensitive) data, society cannot benefit from AI's full potential. Governments should therefore continue to invest in finding ways to exchange data in a secure manner, and a new regulatory framework should include further incentives to build privacy protection collaborations into the systems.

Finally, we recommend that DPIAs include the company's considerations with regard to the alleged lawful basis for processing and, most importantly, take into account the balancing act between the company's interest and the interest of the data subject. This will force companies to observe the rights of the data subjects and will streamline auditing.

2.5 Liability

In some instances, the responsibility for unlawful employment actions stemming from AI-powered tools might be unclear and in others the acts of both developer and employer may be essential to the unlawful act, and the unlawful act might have been a foreseeable consequence of both entities' actions in developing and deploying the tool that caused it. To ensure that these circumstances do not hamper an affected individual claiming damages, we recommend implementing the principle of joint and several liability to allocate damages. Under joint and several liability, a plaintiff might recover its full damages from any party found even partially liable for the harm. That party must then seek indemnification from any other entities that may have contributed to the unlawful acts underlying the judgement.

In any case, an employer may seek indemnity from vendors selling machine-learning-powered tools that they incorrectly claim to be legally compliant. Laws against deceptive business practices might provide employers with a statutory route to indemnification. Where such laws are absent or insufficient, employers can negotiate indemnification clauses that fairly allocate responsibility if use of the algorithm ultimately results in violations of the law.

Contributors

Roderick Beudeker, Senior Associate, Baker McKenzie; World Economic Forum Fellow

With acknowledgement of contributions from Matthew U. Scherer, Associate, Littler

World Economic Forum

Kay Firth-Butterfield, Head of Artificial Intelligence and Machine Learning; Member of the Executive Committee, World Economic Forum

Anne Flanagan, Project Lead, Data Policy, World Economic Forum

Matissa Hollister, Assistant Professor of Organizational Behaviour, McGill University; World Economic Forum Fellow

Baker McKenzie

Celeste Ang, Principal

Carlos Dodds, Partner

Mihoko Ida, Partner

Jonathan Isaacs, Partner

Victor Estanislao Marina, Paralegal

Bradford Newman, Partner

Tracy Robbins, Associate

Daryl Yang, Practice Trainee

Pymetrics

Frida Polli, Chief Executive Officer

Kelly Trindel, Head of Policy + I/O Science

Additional Contributions

Garry G. Mathiason, Senior Partner, Littler

Manish Raghavan, PhD Student, Cornell University

Susan Scott-Parker, Chief Executive Officer, Business Disability International

Endnotes

1. The dependence on data also gives a major advantage to the biggest companies with access to the largest amounts of data, but that is beyond the scope of this paper.

2. Safeguards "should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision": Recital 71, GDPR.

3. For example, under the New York City bill regulating AI hiring tools (Int. 1894–2020), employers using automated employment decision tools to (among other tasks) screen a candidate for an employment decision are required to notify the candidate about the job qualifications or characteristics that such a tool uses to assess candidates.

4. Bryan Casey, Ashkon Farhangi and Roland Vogl, "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise", Berkeley Technology Law Journal (2019), p. 158.

5. A29 WP Guidelines on Automated Individual Decision-Making and Profiling, p. 31.

6. Idem, p. 25.

7. Limitations include the need for domain experts to evaluate explanations, the risk of manipulation or spurious correlations reflected in model explanations, the lack of causal intuition and the latency in computing and showing explanations in real time. Umang Bhatt et al., "Explainable Machine Learning in Deployment", Cornell University (2019), p. 9: arXiv:1909.06342 [cs.LG] (link as of 18 May 18, 2020) .

8. Both phenomena are examples of construct-irrelevant variance; that is, systematic differences in test results that are due to factors that do not actually affect or reflect a candidate's ability to perform the job. See American Educational Research Association et al., Standards for Educational and Psychological Testing (4th ed., 2014), p. 109.

9. Bradley A. Areheart, "The Anticlassification Turn in Employment Discrimination Law", 63 Alabama Law Review, 955 (2012).

10. Kimberle Crenshaw, "Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color", Stanford Law Review, 43 (6) (1991), pp. 1241–1299.

11. Sandra Wachter and Brent Mittelstadt, "A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI", Columbia Business Law Review (2019-1).

12. Nripsuta Ani Saxena, "Perceptions of Fairness". In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). ACM, New York, NY, USA, pp. 537–538. https://doi.org/10.1145/3306618.3314314 (link as of 18 May 18, 2020).

13. In the United States, this general concept, with some modification, is encoded in the "business necessity" or "job-relatedness" defence to claims of disparate impact on the basis of race, colour, religion, sex or national origin. This defence allows an employer to rebut a showing of disparate impact by demonstrating that the selection procedure was "job related for the position in question and consistent with business necessity": 42 U.S.C. § 2000e-2(k)(1)(A)(i) (link as of 18 May 18, 2020). See also Americans with Disabilities Act, 42 U.S.C. § 12112(b)(6) (link as of 18 May 18, 2020) (prohibiting employers from using selection procedures that "screen out or tend to screen out" individuals with disabilities unless the procedure is "job-related for the position in question and is consistent with business necessity").

14. See, e.g., Ernst v. City of Chicago 837 F.3d 788, 805 (7th Cir. 2016).

15. Matthew U. Scherer, Allan G. King and Marko J. Mrkonich, "Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms", South Carolina Law Review, 71 (2019), p. 22.

16. Given the mathematical principles upon which many algorithmic tools operate, however, it is often impossible for an employer to check every possible model to see which one provides the least discriminatory impact. Consequently, even the use of strong de-biasing techniques may not eliminate the possibility that a plaintiff might later be able to discover – even if only by chance – an equally effective and even less discriminatory alternative procedure. It is not clear how courts or enforcement agencies would handle a discrimination claim where a plaintiff can point to a theoretically "available" alternative procedure that the employer could not reasonably have been expected to find. See Scherer et al., pp. 62–63.

17. A reference group of people where protected characteristics are known.

18. Accuracy trade-offs may be reduced if: (1) the algorithm is learning a task that is not strongly linked to protected characteristics; (2) the raw data is job-relevant; and (3) standards are in place for both accuracy and fairness such that algorithms are selected only if they meet such standards.

19. Manish Raghavan et al., "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices", Cornell University (2019), 12, p. 9: arXiv: 1906.09208, 2019) (link as of 18 May 18, 2020), p. .

20. See, e.g., Ricci v. Destefano, 557 U.S. 557, 593 (2009) (holding that an employer engaged in unlawful disparate treatment discrimination when it discarded the results of an employment test that had a severe adverse impact on black and Hispanic candidates); 42 U.S.C. § 2000e-2(l) (employers may not "adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment related tests" on the basis of protected-class status).

21. Penalties of up to 4% of annual global turnover of the preceding financial year or €20 million, whichever is greater.

22. Article 6, GDPR.

23. Article 9, GDPR.

24. A29 WP Guidelines on Automated Individual Decision-Making and Profiling, p. 18.

25. "Profiling" means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements. Article 4 (4), GDPR.

26. Most employment decisions qualify as decisions producing legal effects or having a similarly significant impact. Pursuant to Recital 71, e-recruiting practices fall under this scope.

27. Guidelines 05/2020 on Consent under Regulation 2016/679, p. 8.

28. A29 WP Guidelines on Automated Individual Decision-Making and Profiling, p.23.

29. Idem, p. 32.

30. European Commission, EU White Paper on Artificial Intelligence – A European Approach to Excellence and Trust, p. 20. See https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (link as of 18 May 18, 2020).

31. Leilani H. Gilpin et al., "Explaining Explanations: An Overview of Interpretability of Machine Learning", IEEE 5th International Conference on Data Science and Advanced Analytics (2018), pp. 80–89.

32. Idem, pp. 19–20.

33. Idem, p. 19.

34. As suggested in the New York City bill regulating AI hiring tools (Int. 1894–2020).

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744

contact@weforum.org
www.weforum.org