Ref. Ares(2020)3360001 - 26/06/2020

Policy Brief 2019-01

Artificial Intelligence: Humans must stay in command

Lack of accountability, potential misuse in HR processes and digital data monopolies must be regulated – and social consequences anticipated

Artificial Intelligence (AI) may improve the efficiency and reliability of industrial processes. It could thereby support the market position of European companies and thus sustain high-quality employment in a globally competitive world. However, it raises a number of major concerns for European workers in industry: (1) the capacity of machine-learning systems to supervise workers systematically and permanently; (2) the unexplainable nature of decisions or recommendations made by these systems; (3) their capacity to guess or to anticipate sensitive personal data of workers; (4) the rules to access industrial data, which can lead to digital monopolies and (5) the volume of employment and the qualification of tasks remaining for humans. Additional concerns relate to: (6) the inherent conservatism that algorithms based exclusively on past experience entail; (7) the loss of control on self-learning systems after delivery by the producer and (8) the unreliability of a system that can use its own output as teaching material. For each of these concerns, industriAll Europe makes suggestions for policy.

Preliminary remarks

"Human in command" approach

Following the ETUC[1] and the EESC[2], industriAll Europe believes that "*the **primacy of humans** on machines and AI (human-in-command approach) must be established as a fundamental ethical principle underpinning any future initiative aiming at regulating robotics and AI applications*" and that humans should "*never become the underlings of machines*". We therefore strongly support the development of **European ethical guidelines** on AI.

Potential benefits for the quality and cost-based competitiveness of European industry

The improvements brought by AI in speed, quality and reliability of professional processes mean that they will be implemented, and legitimately so, in order to make production more **efficient** in the usage of all resources. E.g. the waste rate in a steel rolling mill with a skilled workforce (or which would be driven by an AI system) is around 3%, while it reaches levels of 25% with an unskilled workforce. In that sense, the investment efforts announced by the Commission[3] (consolidated public-private investment of EUR 20bn over 2018-20, and EUR 20bn/year beyond) are welcome to maintain the competitive position of European companies in global markets.

Scope: "weak" AI

The scope of this reflection is "weak" AI, which is the type of AI that works and is operational in 2018. This form of AI performs strictly limited tasks, based on "machine learning": computers extract the information embedded in large amounts of unstructured data and develop a capacity to take decisions / make recommendations on cases not yet seen, based on

[1] ETUC "Resolution on tackling new digital challenges to the world of labour, in particular crowdwork" (October 2017)

[2] EESC Own-Initiative Opinion INT/845 "Artificial intelligence: anticipating its impact on work to ensure a fair transition" (September 2018)

[3] In its "Communication on Artificial Intelligence for Europe" COM(2018) 237 final of April 2018 .

the experience from the past gathered in the teaching data. The underlying software technique is called "neural networks", because it mimics the structure of the brain.

It does not cover the more futuristic prospects of "general AI", where a single software would be able to engage in a great variety of tasks.

Implementation of the proposed policies

This document suggests policies / regulation to address the issues that it identifies. IndustriAll Europe believes that the general framework regulating AI should be defined by legislation. Considering the broad diversity of labour relations in Europe, it leaves deliberately to all relevant parties the choice for the detailed implementation of these suggestions, within the general framework thus defined (via collective agreements at the appropriate scale, social dialogue, co-determination, legislation or any other suitable means).

Machine-learning systems can contribute to the systematic and permanent automated supervision of workers

Traditionally, the supervision of workers by management was technically and economically restricted by the difficulty of having a person looking permanently at the work performed by another to detect non-compliance with prescriptions (regarding speed, quality or safety). Even with cameras, it was difficult for a single person to supervise many, so that this supervision remained costly.

With AI systems, it becomes technically and economically possible to supervise all workers, permanently, and to detect all occasions of non-compliance with prescriptions, in real time. This has the potential to significantly reduce the space of worker autonomy and workers' contributions to innovation based on their professional skills and experiences – to the detriment of the quality of their life at work and of their motivation and long-term performance.

Suggestion for policy / regulation: The collective voice of workers must play a determining role in ensuring AI and machine-learning systems are used in their interest in a balance with that of employers.

Workers should be **informed** and **consulted** regarding all **automated tools** used by management to (1) **supervise** work, (2) **manage** the workforce in Human Resources (HR) processes or (3) **profile** workers, i.e. anticipate their performance or reliability at work.

Management should thus report to, consult and reach agreement with trade unions or works councils on:

- the nature of the **data** being collected on workers, the frequency of its collection and the duration of its storage;
- the explicit **algorithms** or the **machine-learning system** used to process this data;
- the **metrics** used to evaluate work and the **performance values** required from workers;
- the **teaching data**, its **biases** and the means implemented to overcome them;
- the **reliability** and accuracy statistics of any implemented machine-learning system (error rates = "false positives" and non-detection);
- the **acceptable** means to **supervise** work and to detect, store and process circumstances of non-compliance with work prescriptions;
- the procedures for workers or their representatives to detect **errors** or **unfair treatment** in this automated processing, report them and gain **redress**.

Works councils should be provided the means to hire the competencies of software engineers or data scientists to support them in these discussions.

Policy should mandate the creation of a position of **data accountant** in companies, whose duty is to control and report annually on the use of AI

systems, in the way a financial accountant controls and reports on the financial situation.

Policy should refuse to consider **individual** consent as enough to ensure that it is "*freely given*" (GDPR, Art.4(11)), when in a situation of employment or of dependent work[4]. **Consent** to the processing of worker-related data and to profiling based on machine learning should only be given **collectively**.

Neural networks are currently un-explainable

In the current state of science, "machine learning" systems can neither **explain** nor **justify** their decisions/recommendations. Contrary to explicit algorithms which can be followed by a human (provided the source code and the supportive data are public), and where all steps having led to the decision / the recommendation are explicit, neural networks are a complete "black box". No scientist is able today to track back, from the teaching data and the learning algorithm, what led to a decision / recommendation.

This is problematic in general, because it **weakens** further the capacity of humans to **influence decisions**, when they receive "recommendations" from an AI system, along an argument akin to "nobody was fired for choosing IBM" in the 1970s: a human taking a decision contrary to the recommendation of an AI system will make mistakes, and be sanctioned for having done so, whereas they will not for having followed the recommendation of the machine, even if this decision ultimately proves to be wrong. This leads, statistically, to a situation where *de facto* almost all decisions are taken by the machine.

This is even more problematic in situations involving the **management of workers**, on decisions impacting their professional development (e.g. promotion, dismissal, training). "Employee assessment" software can now predict the professional development potential of a worker and make recommendations for the management of their career. The manager receiving such a recommendation would not be able to explain / justify it, other than with the "argument from authority": so, there is no justification given other than an opaque reliance on experience. This could deprive the worker from any possibility to discuss, present arguments to support their case and gain redress. This deprival of a human interaction, and of a fair judgement, is very problematic for workers.

This feature of being unexplainable is also problematic when considering **liability** and **improvement paths** in case of failure or accident – for autonomous cars, production machines, and even more for airplanes or nuclear power plants (i.e. safety-critical artefacts). Finding the cause of the failure or accident is important to determine who must pay for the damage. It is also important in order to **improve** the system and ensure this failure or accident of the system does not reoccur. If the command and control system is based on AI, and is thus unexplainable, then liability cannot be determined, and no improvement is possible.

Suggestions for policy / regulation:

Mandate that any machine learning software taking decisions regarding **humans** and specifically workers (e.g. regarding health or HR management) or embedded in a **safety-critical** system (e.g. rail equipment, civil aeronautics, nuclear power), be **explainable** – and **prohibit** its use if not the case.

Minimise the "unexplained" fraction of the predictive software based on neural networks. Many phenomena have been studied by science. They can be the purpose of **explicit modelling**,

[4] whereby following the "Guidelines on Consent under Regulation 2016/679" published by the Working Party of EU Data Protection regulators

involving known equations and explicit calculations using known parameters, or parameters that can be explicitly estimated using standard statistical tools. Thereby, instead of having one single "black box" modelling the whole system, the behaviour of which would need to be anticipated by the AI software, the idea would be to model explicitly all that can be modelled, leaving for AI, and for the unexplainable "black box", only a small fraction of the modelling. This would have the additional advantage of requiring much fewer teaching data.

Machine-learning systems can guess or anticipate sensitive personal data of workers

Machine-learning systems can be used (and are already used) in HR processes: upon hiring a person, and when managing their career (promotion, training, dismissal).

In these processes, machine-learning systems can detect or **predict** professional behaviour and performance with a level of accuracy close to that of a human, at a fraction of the cost, in a process called "**profiling**" described in the General Data Protection Regulation – GDPR[5]. These evolutions are regrettable but difficult to stop. They should however be regulated, with a strong involvement of workers.

Profiling with machine-learning systems could also be used to detect or predict **sensitive personal data** ("*special categories of personal data*" as defined by Art.9, GDPR), such as "*political opinions, religious or philosophical beliefs, trade union membership, […] health, […] sex life or sexual orientation*".

The processing of data that explicitly describes the current situation of a person regarding these sensitive issues is of course currently prohibited by the GDPR (Art.9). This Regulation does not however prohibit guessing or detecting this sensitive information using indirect data sources (e.g. trade union membership from the person's contacts with the shop steward) or predicting the person's evolution in these matters (e.g. burn-out from repeated absences and conflicts with colleagues). It only provides (Art.22) that a person has "*the right not to be subject to a decision based solely on automated processing, including profiling*", and authorises this profiling if the person gives their "*explicit consent*" on an individual basis (Art.22 (2)c).

Therefore, there is a risk that workers be required upon recruitment to individually sign their consent to be subject to a general-purpose profiling, based on machine-learning, which would anticipate or detect their status regarding "*special categories of personal data*", including trade union membership. Provided thereafter that any final decision regarding HR (recruitment, training, promotion, dismissal) involves a human being (and is thus not "solely based" on automated profiling), it could be allowed.

A means for employers to justify such profiling could be to anticipate the **health condition** of a worker, and specifically the appearance of **chronic diseases**. This can indeed be used positively, in order to engage in **preventive measures**. It can also be used in a more malevolent way by the employer: by knowing in advance that the person would develop a chronic disease (or have a high probability of doing so), the employer could dismiss the person beforehand, and thus evade its responsibility.

Suggestions for policy / regulation:

Following the suggestion made in the previous paragraph, the management of companies should also report to, consult and reach agreement with

[5] "Profiling" in the GDPR means: an automated processing of data to "***predict*** *aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements*" (Art.4(4)).

trade unions or works councils on the nature of any automated **profiling** being performed on workers and on the **information** given to each worker on their profile(s).

Prohibit the usage by employers of machine learning systems that anticipate or detect the health status of workers or any "*special category of personal data*". The only person entitled to anticipate the health condition of a worker, if at all, should be the medical doctor responsible, under strict confidentiality clauses.

Access to data: risk of digital monopolies

The development of an AI system based on "machine learning" relies on the availability of **teaching data**. Without such teaching data, the machine cannot learn, and thus cannot be implemented. This leads to the issue of **access rights** to personal or industrial **data**.

In the current state of the art, collecting large amounts of data requires no inventiveness, and almost no investment, because of the very low cost of sensors, and of data transmission and storage. There is thus no legal or moral basis for defining any form of "ownership" over such machine-collected data (be it on private persons, on workers or on objects / machines). A private capture of machine-generated data in a professional environment would be particularly damaging, because this data embeds the **professional experience** of workers, so that the data monopolist would *de facto* capture this experience. On the other hand, machine-collected data can find many socially and economically beneficial usages with different players, such as improving the process (in the operating firm), the maintenance procedure (in the maintenance service firm) or the machine itself (with the manufacturer / designer of the machine).

Suggestion for policy / regulation: Considering the collective advantage of broadly sharing access to such data, and the risks of unjustified rents associated with monopolistic access, there is a rationale for a regime of **mandatory non-exclusive licensing of machine-collected data**[6].

The impact of AI on employment and qualifications must be managed responsibly

AI performs and will perform tasks that humans currently do. It will often perform them better, at lower cost and with greater reliability than humans. As such, this technology is yet another one **increasing** the **productivity of human labour**, like many other technologies in history. One important difference is that the tasks AI can perform are those relying on human **experience**, and are thus often very **qualified**, e.g. of salespersons, workers driving complex machinery, maintenance workers diagnosing failures, lawyers, medical doctors.

Estimates vary regarding the fraction of human tasks that could be replaced with AI. The same uncertainty reigns regarding the duration necessary for AI applications to penetrate and dominate the market. Whatever the exact figure, the impact will most probably be considerable.

Policy demands

Trade unions support strong policies to anticipate and manage the social consequences of industrial change, such as those potentially brought by AI. These demands address: (1) **anticipation of change**, in order to act before the restructuring takes place; (2) **reskilling** and **upskilling** of all workers, of all ages and qualification levels.

Regarding anticipation of change, industriAll Europe believes that workers must not be left with contributing exclusively to managing the social

[6] For more details on this topic, see our Policy Brief "Sharing the value added of industrial Big Data fairly".

consequences of AI but must proactively contribute to shaping a world of work in which AI would play a role. Timing is key. Workers must thus be informed and consulted at the earliest possible stage, both at national and European levels, should we want to ensure that the change brought by AI is anticipated and managed in the most socially responsible way. Workers must play an active role in the decision-making process including, where possible, at board level, concerning the introduction of AI[7].

On reskilling and upskilling of workers, industriAll Europe considers that forward planning of employment and skills, workforce planning, lifelong learning and upgrading of workers' skills are all cornerstones in terms of the anticipation of and preparation for changes within companies, as well as limiting any negative consequences they may have on employment. Access to continuous training must be guaranteed for all workers, irrespective of their age, profession or statute (employee, self-employed, platform worker, freelancer). This must be underpinned by an individual worker's right to training, preferably guaranteed by collective agreements, as called for by industriAll Europe's first Common Demand and by the ETUC[8]. This right should also include the right of validation and recognition of the training.

In addition, industriAll Europe supports the demands by ETUC and the EESC of a "**European transition fund**" to support those workers and regions negatively impacted by AI and more generally by the digitalisation of industry.

An additional area of trade union reflection relates to **working time**, which is the purpose of industriAll Europe's recently adopted Working Time Charter.

Additional concerns

Conservatism embedded in software

AI relies on the general idea that decisions impacting the future should be based on the past experience embedded in the teaching data, i.e. on the implicit assumption the **future** will be **identical to the past**. This is the very definition of **conservatism**. It leaves **no space** for **change** or **innovation**. AI thus risks reproducing the *status quo* forever – including any **discrimination bias** present in our societies, and thus in teaching data.

Suggestion for policy / regulation: Support research on means to introduce innovation, experimenting, change and creativity in the operations of machine learning systems.

Systems based on AI continue evolving after having been delivered, leading to a loss of control by humans and to potential liability evasion

Machine learning does not stop upon delivery of the product to the customer. It continues, based on the experience and teaching data accumulated while being used by the customer. This leads to a situation where the original manufacturer and the customer / user have **lost control** on the behaviour of the machine. This raises a significant question regarding the **liability** in case of accident, because no one had any means to fully anticipate the behaviour of the machine.

Suggestions for policy / regulation:

Define clearly the **liability** in accidents and incidents involving AI systems. The current general rule, whereby the employer is by default liable for any accident in the workplace (in the absence of any wrongdoing by the worker) should remain, and workers victims of such an accident should be compensated swiftly and without delay.

This solid liability regime should be maintained even when the behaviour of a machine-learning system keeps evolving after purchase. One way

[7] For further developments, please refer to industriAll Europe's resolution "Strengthening our capacity to anticipate and deal with change in national and multinational companies in the EU" (2015)

[8] ETUC resolution "EU priorities on education and training post 2020 – towards a European right to training for all" (March 2018)

forward could be to reapply and adapt the legal regime of animals.

Machine-learning systems that use their own output as teaching data lose any reliability over time

When humans use the output from machine-learning systems, they produce data. If this data is indiscriminately re-used as teaching data for a further cycle of machine learning, this leads to an unstable amplification of any error or bias in the initial machine-learning system. The most obvious example of such instability is provided by the Google Translate service. As more and more persons use this system to translate, and to publish their translated texts on the Web, the Google Translate service considers these texts as legitimate sources of teaching to modify its translation machine, and whereby deteriorates its quality, as the proportion of genuine, human-based translation in its teaching database diminishes.

Suggestion for policy / regulation: Mark the output of machine-learning system with an identifier signalling that this data should not be re-used as teaching data for the same system.