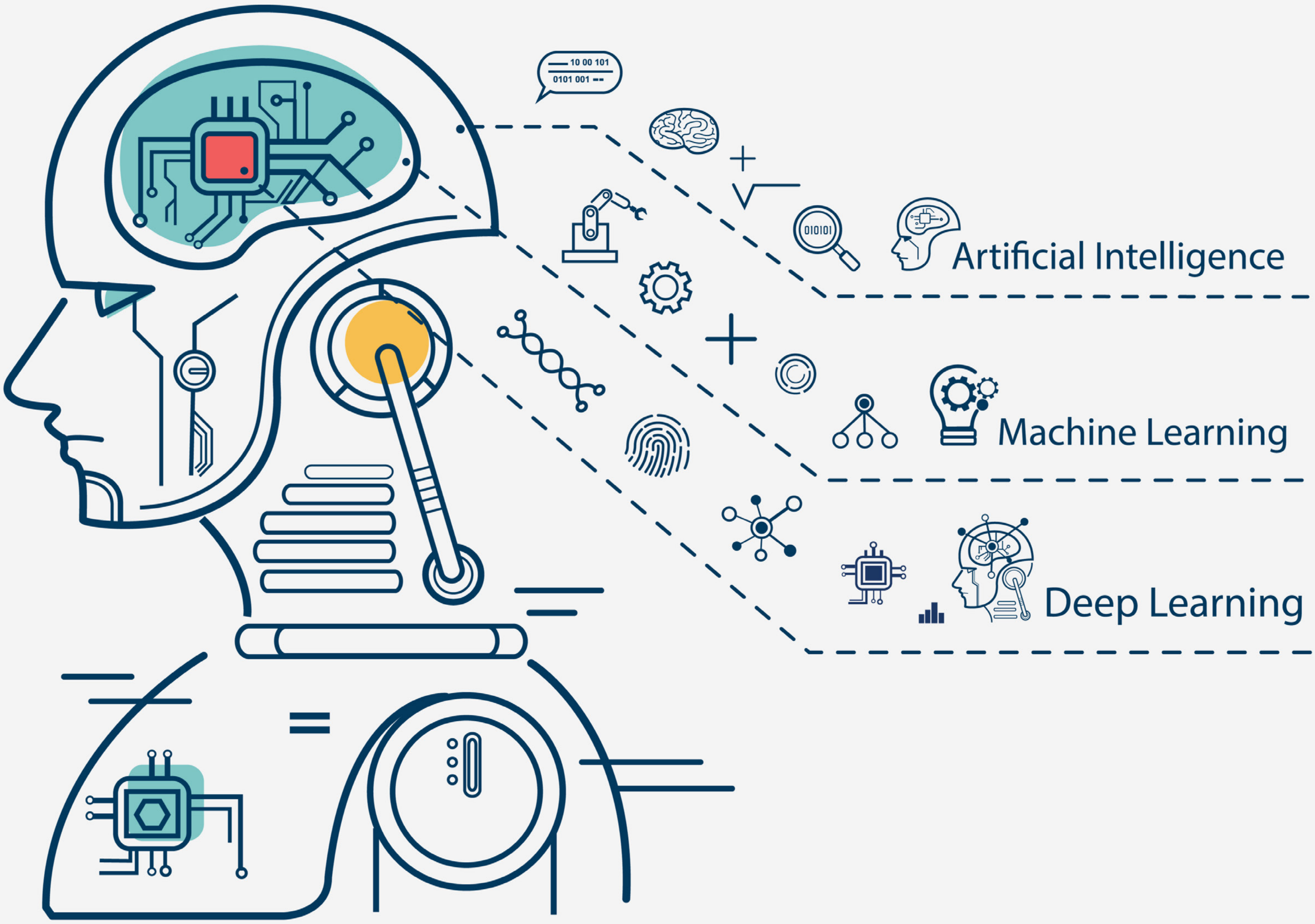
ELI
EUROPEAN
LAW
INSTITUTE

Guiding Principles for Automated Decision-Making in the EU

ELI Innovation Paper



EUROPEAN
LAW
INSTITUTE

ELI Innovation Paper

Guiding Principles for Automated Decision-Making in the EU

Professor Teresa Rodríguez de las Heras Ballell
(University Carlos III of Madrid and ELI Executive Committee Member)

The European Law Institute

The European Law Institute (ELI) is an independent non-profit organisation established to initiate, conduct and facilitate research, make recommendations and provide practical guidance in the field of European legal development. Building on the wealth of diverse legal traditions, its mission is the quest for better law-making in Europe and the enhancement of European legal integration. By its endeavours, ELI seeks to contribute to the formation of a more vigorous European legal community, integrating the achievements of the various legal cultures, endorsing the value of comparative knowledge, and taking a genuinely pan-European perspective. As such, its work covers all branches of the law: substantive and procedural; private and public.

ELI is committed to the principles of comprehensiveness and collaborative working, thus striving to bridge the oft-perceived gap between the different legal cultures, between public and private law, as well as between scholarship and practice. To further that commitment it seeks to involve a diverse range of personalities, reflecting the richness of the legal traditions, legal disciplines and vocational frameworks found throughout Europe. ELI is also open to the use of different methodological approaches and to canvassing insights and perspectives from as wide an audience as possible of those who share its vision.

European Law Institute Secretariat
Schottenring 16/175
1010 Vienna
Austria
Tel.: + 43 1 4277 22101
Mail: secretariat@europeanlawinstitute.eu
Website: www.europeanlawinstitute.eu

ISBN: 978-3-9505192-7-3

Cover image: Shutterstock

This publication was co-funded by the European Union's Justice Programme. Acknowledgment is also due to the University of Vienna which has generously hosted the ELI Secretariat under successive Framework Cooperation Agreements since 2011.

Supported by
the European Union

universität
wien

Table of Contents

I. Background

The intensive and extensive use of algorithms has pervaded an immense and growing variety of tasks, activities, and decision-making processes in the digital economy. In an over-informed society, automation is key to managing complexity, curbing uncertainty, performing mass activities at an affordable cost and to ensuring effectiveness in the processing of data, information, and digital content. From basic tasks (searching, comparing, ordering, prioritising), to more sophisticated added-value services (profiling, personalising, recommending, multi-attribute rating, filtering, content moderation, algorithmic management, complaint handling), all are performed by algorithm-driven systems.

Algorithmic automation provides efficiency, dramatically reduces transaction costs, streamlines processes, and assists decision-making in complex contexts. Rating, ranking, recommender systems or comparators are extremely helpful tools, which assist users in making informed decisions. Profiling, personalising or contextualising solutions enable companies to successfully reach their prospective customers with targeted communications, personalised offers, and customised services. Algorithms are instrumental in rendering flagging, filtering, content moderation or content removal feasible, affordable, and effective. Algorithms are vital to managing complexity, uncertainty, and virality in contemporary societies.

However, at the same time, significant risks and undesired effects of algorithm-driven systems for society are becoming increasingly perceptible. Algorithmic logic may perpetuate choices and preferences, radicalise speech, polarise public opinion in echo chambers and ideological silos, reduce diversity, enlarge bias and discrimination divides, standardise behaviour on the basis of stereotypes, lead to opaque decisions that leave victims undefended, stoke the virality of fake news, encroach upon free speech, or distort consumers' choices with misleading ratings, rankings, dark patterns, or recommendations.

The potential, as well as the inherent risks, of automation for the future of the digital society have not gone unnoticed in the European Union (EU). On the contrary, the principles of transparency, explainability, risk assessment, and human oversight of algorithm-driven systems are crystallising in the EU legislative initiatives adopted in the last few years and in those proposed more recently.

Inter alia, article 22 of the General Data Protection Regulation (GDPR)[1] on decisions based solely on automated processing, including profiling, has long been the centerpiece of the EU's legal approach to automated decision-making (ADM) and embodies its main policy goals. The Platform-to-Business Regulation (P2B Regulation)[2] confirms these policy goals with transparency requirements in the provision of ranking services. Likewise, algorithmic accountability and transparency also bolster some obligations laid down in the proposed Digital Services Act (DSA)[3] – with respect to recommender systems, terms and conditions, and content moderation. Risks arising from algorithmic decisions are acknowledged throughout the proposal and, accordingly, included in the risk assessment and subject to risk-mitigating measures that apply to very large online platforms. In addition, the proposed Artificial Intelligence Act (AI Act)[4] represents a risk-based approach to AI systems and the consolidation of certain principles, which aim

[1] Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.

[2] Regulation (EU) 2019/1150 of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services [2019] OJ L186/57.

[3] Commission, 'Proposal for a Regulation of the European Parliament and the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC' COM (2020) 825 final.

[4] Commission, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts' COM (2021) 206 final.

at providing guidance as to the placing on the market and the use of AI on the basis of its intended purpose. The recent Proposal for a Directive on improving working conditions in platform work[5] (Directive on Platform Work) devotes its Chapter III to algorithmic management under the principles of transparency, human monitoring, and human review of significant decisions.

All the above references show that automated decision-making systems are attracting regulatory attention in the EU. Nevertheless, rules related to automated processes are scattered in different pieces of legislation, are partial in their scope, and are unharmonised. Some rules are sector-specific, while others apply solely to certain types of automated systems (rating, recommender systems, algorithmic management). Besides, despite the fact that the main policy goals are enshrined in a number of legal provisions (relating to transparency, explainability, and human monitoring), their implementation is still uncertain in practice, may be unfeasible or too costly, or may become significantly complex.

Hence, even if useful rules on algorithm-based systems and automated means can already be found in the EU legislative scene, they do not form a consistent, coherent, and all-embracing body of principles/rules governing automated decision-making systems.

[5] Commission, 'Proposal for a Directive of the European Parliament and of the Council on improving working conditions in platform work' COM (2021) 762 final.

II. Conceptualising ADM

Automation and algorithmic processes have intensively permeated EU regulation. Automation is explicitly mentioned, or is implicitly referred to, in several provisions of the most recent pieces of EU legislation. Legislation and legal proposals such as the GDPR, DSA, the proposed Digital Markets Act (DMA),[6] and P2B Regulation refer to algorithmic rating, algorithmic decision-making, algorithmic recommender systems, algorithmic content moderation, algorithmic structures, automated profiling, or a variety of activities and actions conducted by automated means. Nevertheless, there is neither a unified concept of ADM nor harmonised terminology to describe such a wide variety of automated processes. The definition of 'AI systems',[7] for the purposes of the AI Act,[8] does, however, provide key elements to enable the formulation of a working definition of ADM:

1. **inputs** (these can be human-based inputs, machine-generated data, or interactions with the environment);
2. **pre-defined objectives**;
3. **techniques and approaches to achieve those objectives**; and
4. **outputs** (deliver, enable or support content moderation; ratings, rankings, predictions or recommendations; online advertising; complaint handling and dispute resolution; traceable traders (Know Your Business User approach); algorithmic management in platforms; credit scoring; pricing, trading and investing, or compliance).

Based on the above-mentioned elements, for the purposes of this Innovation Paper, the following definition of ADM is proposed:

> **ADM is a (computational) process, including AI techniques and approaches, that, fed by inputs and data received or collected from the environment, can generate, given a set of pre-defined objectives, outputs in a wide variety of forms (content, ratings, recommendations, decisions, predictions, etc).[9]**

The above-proposed definition of ADM has two consequences.

[6] Commission, 'Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act)' COM (2020) 842 final.

[7] Art 3(1) AI Act:

> 'Artificial intelligence system (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with'.

The compromise text unveiled at the end of November 2021 by the Slovenian Presidency of the European Council ('joint compromise', Council of the European Union, Presidency compromise text, 29 November 2021, 2021/0106(COD), henceforth simply 'joint compromise'), <https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf> accessed on 27 April 2022, proposed some changes to this definition. In the preamble, the joint compromise clarifies that the proposed amendments are intended to make explicit that an AI system, unlike traditional software, should be capable of determining how to achieve a given set of human defined objectives by learning, reasoning, or modelling. The revised definition is the following: 'artificial intelligence system (AI system) means a system that:

> (i) receives machine and/or human-based data and inputs,
> (ii) infers how to achieve a given set of human-defined objectives using learning, reasoning or modelling implemented with the techniques and approaches listed in Annex I, and
> (iii) generates outputs in the form of content (generative AI systems), predictions, recommendations or decisions, which influence the environments it interacts with.'

[8] A similar definition can be found in the European Parliament Resolution with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)) [2020] OJ C404/107, that includes a Proposal for a Regulation of the European Parliament and the Council on liability for the operation of Artificial Intelligence-systems: 'AI-system means a system that is either software-based or embedded in hardware devices, and that displays behaviour simulating intelligence by, *inter alia*, collecting and processing data, analysing and interpreting its environment, and by taking action, with some degree of autonomy, to achieve specific goals'.

[9] The definition of ADM is largely aligned with the definition of algorithmic decision-making proposed by Art 2 of the ELI Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration. ELI, *ELI Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration*, <https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Model_Rules_on_Impact_Assessment_of_ADMSs_Used_by_Public_Administration.pdf> accessed 20 April 2022.

1). First, *the concept of ADM includes algorithmic decision-making as well as AI-driven decision-making*. This is relevant for EU legislation, insofar as, while AI-specific legislation explicitly refer to and define AI systems within their scope, other texts (GDPR, DSA, DMA, P2B Regulation, Directive on Platform Work) refer, without defining it, to algorithmic processes, or automated means. Both categories of legislation are relevant in the attempt to specify potential uses and applications of ADM, to infer principles and to formulate harmonised rules for ADM.

2). Second, ADM can produce or deliver a myriad of outputs from a rating to a credit granting decision, from a movie recommendation to the allocation of work assignments, from an estimation of an insurance premium to a decision to remove illegal digital content. *The variety of outputs is immense. Thus, the resultant legal consequences and the applicable legal regimes to ADM differ and are potentially multiple and varied.*

Assuming a diversity of outputs delivered by ADM, a relevant distinction is to be made between two categories of outputs. The classification is based on the position of the 'affected person' and their relationship with the ADM's output. To this end, the term 'affected person' denotes the person who interacts with the ADM, being the addressee of the ADM (person affected by the decision), or using or relying on its outputs for other purposes (relying on a recommendation or on a ranking to make a subsequent informed decision). The terminology used to describe the parties involved in the operation of ADM is further explained below (III).

> 2.a. The first category of possible outputs of ADM comprises ratings, rankings, predictions, recommendations, or content classification. The ADM lists, prioritises, classifies, rates, filters, ranks or recommends. The resulting output may be 'used' as an input to make subsequent decisions: what to buy, where to stay, which item to choose, whom to deal with. Thus, the output is not a decision which relates to the rights or the status of the person interacting with the ADM, or which has legal effects on them. Naturally, the person can also be indirectly affected by the ADM if the prediction is inaccurate, the ranking is misleading, the rating is based on self-preferring practices, or the recommendation is biased to intentionally promote the acquisition of operator-sponsored products. Consequently, the person may subsequently make a 'wrong' decision as a result of relying on the ADM. But still, in all these cases, the ADM's output is an 'input' of the subsequent decision taken by the person interacting with the ADM. On the other hand, third parties can be directly affected by the ADM: the downrated seller who loses customers; the demoted professional user whose reputation is harmed; the author of content classified as unreliable who alleges that the decision encroaches upon their freedom of expression, or the unrecommended product manufacturer who alleges unfair competition.
>
> 2.b. The second category of outputs covers decisions adopted by the ADM related to the affected person. In such cases, the person interacting with the ADM is directly affected by it. The ADM scores, grants, awards, settles a dispute, handles a complaint, removes digital content, assigns work, closes an account, dismisses an employee, refuses a request for credit, or decides whether an event notified by a user has to be compensated as per the insurance policy terms and conditions. The ADM produces legal effects concerning the affected person and/or significantly affects their rights, legal or contractual status, or interests. The affected person is selected or rejected, scored, awarded, demoted, expelled from the platform, dismissed or is somehow directly affected by the decision.

The Guiding Principles apply to both types of outputs. However, the intensity of the legal effects in the second category of ADM requires closer scrutiny and greater legal control. Hence, certain Principles are only relevant for the second category, where the ADM makes decisions likely to produce legal effects for the affected person or to have a significant impact on their rights, status, or interests.

III. Relevant Parties Engaged in ADM

For the purposes of this Innovation Paper, there are two main parties involved in the provision and the performance of ADM: the operator and the affected person.

The **operator** employs, implements, or utilises the ADM in the course of a professional, or business activity. Purely personal, non-professional activities are excluded from the scope. However, as ADM can be used in the provision of public services, in the context of dispute resolution, or in the performance of public administrative functions, the following Principles do not, in principle, differentiate between private and public activities and may potentially apply to public authorities implementing ADM, without prejudice to the application of additional specific principles and rules relevant to public services, or public authorities. The provision of public services or the exercise of public functions likely to materially impact citizens' rights and liberties should be subject to special regulatory scrutiny.[10] Likewise, legislators may consider it unacceptable to admit fully automated dispute resolution as access to justice would thus be deprived of human intervention. Thus, although these Principles aspire to lay the foundation for a comprehensive set of rules for ADM, and to that end, public entities are not per se excluded as operators (ie providing ancillary services or certain decisions), specific rules and principles will prevail on the basis of the nature of the service delivered or the decision adopted or supported by the ADM.

The concept of operator would describe what the AI Act (article 3(4)) terms 'user'. In the terminology and within the scope of the DSA, DMA, or the P2B Regulation, examples of operators of ADM would be the 'provider of online intermediation services' (article 5 P2B Regulation) providing ranking functionalities, the 'very large online platforms' using recommender systems (article 29 DSA), the 'provider of hosting services' using automated means in content moderation (article 14(6) DSA), or the 'core platform service provider' (or, if designated as such, the gatekeeper) applying algorithms to the provision of a variety of services (DMA).

The definition of 'operator' (frontend operator) on the basis of control and benefit proposed in the Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies[11] by the Expert Group on Liability and New Technologies – New Technologies Formation, and followed by the cited European Parliament (EP) Resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence,[12] albeit formulated for the purposes of allocating civil liability, is aligned with the concept of 'operator' in this Innovation Paper. The operator is the person in control of the risks connected with the ADM and who benefits from its operation in the context of a particular activity. Therefore, the binomial control-and-benefit defines who is the operator of the ADM.

The **affected person**, as described above (II), is the natural or legal person interacting with the ADM, either being the person affected by the final decision, or the person using or relying on, for subsequent purposes, including subsequent decision-making, the output of the ADM (prediction, recommendation, rating, ranking). The affected person can be a consumer or a professional user. Additional consumer protection rules apply if the affected person is acting

[10] ELI (n 9).

[11] Expert Group on Liability and New Technologies – New Technologies Formation, 'Liability for Artificial Intelligence and Other Emerging Digital Technologies' (2019) <https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en> accessed 20 April 2022.

[12] European Parliament (n 8).

outside the scope of an economic activity. Certain outputs of ADM systems may have systemic effects insofar as their operation potentially impacts a multitude of persons. The term 'affected person' also covers situations where the ADM makes (or supports) decisions affecting a group of persons, a category of potential beneficiaries, or a multitude of possible recipients (eg the calculation of benefits that affects a decision to grant or reject public aid and its amount, a published credit rating of investment funds that impairs the investment decisions of a multitude of unsophisticated investors).

Other than the affected person, **third parties** can also be affected by the ADM. Should the ADM rank, list, rate, recommend, or classify, third parties are positively, or negatively affected by the process and the output. By being downrated, unrecommended, demoted, or unlisted, third parties may undergo undesired consequences in their competitive positions, market share, or customer retention capabilities.

The **providers**,[13] the **importers** or the **distributors** of the ADM, as defined by the AI Act as those that develop, place on the market, import, or distribute, may also be relevant for some of the Guiding Principles formulated below. If so, they will be explicitly referred thereto. Otherwise, the Guiding Principles are addressed to operators in relation to the ADM that they implement, use and employ in their activities and for their intended purposes.

[13] Providers include software developers, designers, and other providers participating in the development, design, and provision stages of ADM.

IV. ELI Guiding Principles for Automated Decision-Making in the EU

Guiding Principle 1: Law-compliant ADM

An operator that decides to use ADM for a particular purpose shall ensure that the design and the operation of the ADM are compliant with the laws applicable to an equivalent non-automated decision-making system.

Compliance with law is the first Guiding Principle in establishing an enabling legal regime for the use of ADM in any social or economy activity and for any particular purpose. Requiring that the design and the functioning of the ADM comply with applicable laws seems a basic and implied principle. Nevertheless, the acknowledgement of the law-compliance Principle is vital for fostering the use of ADM without compromising the protection of interests and rights at stake and, therefore, instrumental in providing a flexible and future-proof legal regime for ADM.

The law-compliance Principle plays a two-fold role.

On the one hand, it plays a limiting or negative role in deciding when the use of ADM is permitted and to which extent, and whether additional measures or safeguards have to be adopted. If an ADM cannot be designed, or cannot operate, in full compliance with applicable law, its use should be prohibited, limited or subject to certain conditions. This Principle provides guidance both to the legislator, where a general rule in a specific sector is to be adopted, and to the operator, who conducts a case-by-case assessment to decide when and whether to use ADM.

> Illustration A. In an enforcement procedure for monetary claims, once a judicial enforcement order is generated, an ADM might easily identify the debtor's bank accounts and liquid assets and automatically take steps to complete the seizure, the transfer of the funds from the bank accounts, or the transfer of receivables or digital assets. The system will complete the actions on an automated basis. However, that would disregard the existence of exceptions, limitations, rights of preference or prior attachments likely to reduce the amount, or the right of the debtor to challenge that order as per applicable law. A law-compliant ADM has to integrate and properly assess these exceptions, otherwise, the enforcement might be excessive, abusive or unjustified. If the ADM cannot be designed in a way that can take into consideration all possible exceptions, as some require a case-by-case evaluation, additional measures limiting automation have to be implemented. Thus, a stay mechanism would be necessary to suspend the automated execution until the period for the debtor to challenge the order or to claim exceptions or limitations apply elapses.

> Illustration B. An app for travel insurance covers the risks of flight delays and cancellations. The app operates on the basis of an ADM that receives inputs from airports, processes flight information, calculates compensation, and completes the payment. However, in the insurance terms and conditions, a limitation clause includes a list of circumstances exempting the insurer from paying compensation or entitling it to reduce the amount paid. The ADM has to be designed in such a way as to assess the occurrence of relevant exonerating/limiting circumstances. The intervention of oracles (meteorology agency) may solve the problem by inputting relevant data into the ADM (eg storm alerts, snow, low visibility). However,

if the occurrence of the exonerating/limiting circumstances requires a judicial assessment or an expert appraisal, the ADM has to be designed and operated in a manner that ensures the decision is not definitively made until account is taken of any exonerating/limiting circumstances. Otherwise, the ADM would not be compliant with the terms and conditions of the insurance policy stipulated.

On the other hand, this Principle plays a positive or enabling role in promoting the use of ADM for any purpose. Provided that the ADM is designed and functions in full compliance with applicable law, the ADM shall not be denied legal effect, validity or enforceability solely on the grounds that the decision has been reached by automated means. The law-compliance Principle enables the use of ADM without the need for specific recognition by the law.

Illustration C. A platform for delivery services implements an ADM that assigns work. Without prejudice to any algorithmic-specific rule, such as the provisions governing algorithmic management in the Directive on Platform Work, the ADM can be validly implemented by the platform, provided that the ADM complies with the applicable labour laws, respects working times as well as abides by relevant collective bargaining agreements and individual contractual conditions.

Illustration D. Regulation (EU) 2020/1503 on European crowd funding platforms[14] incidentally mentions 'automatic investing (auto-investing)' (recital 20), merely stating that it should be considered individual portfolio management of loans. However, there is no further reference in the provisions in relation to the use of automation, algorithms, or AI-driven functionalities. Should the ADM abide by the rules governing portfolio management, the use of ADM is permitted and is law-compliant. There is thus no need to specifically acknowledge the validity of algorithmic investing.

The Principle of non-discrimination (Guiding Principle 2), complemented by Guiding Principle 1 on law compliance, lay the foundations for the valid and enforceable use of ADM for automatic investing.

This Principle has to be complemented with Guiding Principle 11 (Responsible ADM) that goes beyond law compliance and incorporates other fundamental values, goals and interests into the development, the provision, and the use of ADM.

Guiding Principle 2: Non-discrimination against ADM

As a general rule, ADM shall not be denied legal effect, validity or enforceability solely on the grounds that it is automated.

The principle of non-discrimination is widely recognised in international legal harmonisation instruments on the use of electronic communications in international contracts. The United Nations Commission on International Trade Law (UNCITRAL) Model Law on Electronic Commerce (1996), on Electronic Signatures (2001), and on Electronic Transferable Records (2017) are all based on the principles of non-discrimination, technological neutrality and functional equivalence. More precisely, the United Nations Convention on the Use of Electronic Communications in International Contracts (2005)[15] extends the principle of non-discrimination to the use of automated systems whose actions are not reviewed or triggered by natural persons.[16] Thus, in the absence of human intervention, the action performed by an automated system shall not be denied legal effect, validity or enforceability solely on the grounds that it is performed by automated means.

[14] Regulation (EU) 2020/1503 of 7 October 2020 on European crowdfunding service providers for business, and amending Regulation (EU) 2017/1129 and Directive (EU) 2019/1937 [2020] OJ L347/1.

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

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

This Principle should inspire an enabling legal regime for ADM in the EU. In conjunction with Guiding Principle 1, the non-discrimination Principle endorses the use of law-compliant ADM, unleashing the full potential of automation without compromising the protection of rights and liberties.

> Illustration A. In a procurement system, bidders eligible for subsequent evaluation and interviews are shortlisted by automated means. The ADM takes into consideration, in conformity with the law and the relevant terms of reference, all relevant, objective, and quantifiable criteria requested in the tender. The shortlisting by automated means has the same legal effects as a human selection procedure.

> Illustration B. An electronic auction platform for improving the realisation of value of collateral is based on an algorithmic pricing mechanism. The ADM initiates the auction, sets the price, and adjudicates the sale. The ADM, in compliance with the law, has the same legal effects as a human-driven adjudication.

A non-discrimination rule does not necessarily mean that algorithmic-specific rules cannot be adopted. Certain rules, such as those providing for duties of transparency, explainability or human review, are indeed technology-dependent. They apply precisely where an algorithmic process exists. However, these duties, unless so provided by law, do not result in a questioning or challenging of the legal effects attributed to the ADM, its validity or its enforceability. If other legal consequences derive from failing to comply with such algorithmic-specific duties, the non-discrimination Principle is still preserved.

> Illustration C. In the DSA, implied references to automation and algorithmic decisions are scattered throughout the text without specifying the applicable regime. Guiding Principles 1 and 2 both provide guidance on implementing and developing ADM for a variety of purposes, even if there is no explicit recognition by the law:
>
> Measures against misuse (article 20 DSA):[17] does an evaluation 'on a case-by-case basis' exclude any form of automation? Or, on the contrary, is automation allowed? Automation is allowed and ADM systems are designed to detect the reiterative submission of manifestly unfounded notices or complaints. Provided that the ADM is designed to assess all the relevant circumstances as listed in article 20(3) DSA on a case-by-case basis, there is no objection to this use.
>
> Traceability of traders (article 22 DSA): is the use of automated means for collecting, detecting errors, and/or verification of information provided by traders permitted? Online platforms can employ automated systems to assess the reliability, completeness and accuracy of the information provided by traders or obtained by the platform to identify them. ADMs for such purposes are valid and enforceable, provided that they are designed to comply with article 22(2) and (3) DSA.

The non-discrimination Principle can be limited, exempted, or subject to conditions. Decisions that are likely to significantly affect fundamental rights or liberties (eg deprivation of rights, imprisonment, loss of the right to vote, interference with freedom of expression) or have other relevant legal effect (eg dismissal, closure of an account, loss of access to a service) on the user might be subject to a human-review procedure in accordance with Guiding Principle 10.

> Illustration D. Article 17(5) DSA prevents online platforms from making decisions by internal complaint-handling systems solely on the basis of automated means. Consequently, the ADM can assist or support the decision, but the decision-making cannot be entirely and exclusively automated.

[17] Art 20(3) DSA: '… 3. Online platforms shall assess, on a case-by-case basis and in a timely, diligent and objective manner, whether a recipient, individual, entity or complainant engages in the misuse referred to in paragraphs 1 and 2, taking into account all relevant facts and circumstances apparent from the information available to the online platform. Those circumstances shall include at least the following: (a) the absolute numbers of items of manifestly illegal content or manifestly unfounded notices or complaints, submitted in the past year; (b) the relative proportion thereof in relation to the total number of items of information provided or notices submitted in the past year; (c) the gravity of the misuses and its consequences; (d) the intention of the recipient, individual, entity or complainant …'.

Guiding Principle 3: Attribution of decisions adopted by ADM

The decision adopted by ADM shall be attributed to the operator. The operator shall not deny the attribution of a decision solely on the grounds that it was made by automated means.

The decision taken by ADM is deemed to be the decision of the operator implementing, employing or using that ADM for making or supporting its decisions. Accordingly, the legal consequences of such a decision are to be attributed to the operator, regardless of the fact that the decision was arrived at by automated means.[18]

This Principle has a twofold impact:

> On the one hand, the operator has to assume the legal effects and bear the consequences of the ADM's decision. That refers to the attribution of the legal effects and consequences to the operator as the decision maker: ie the contracting party (if the ADM concludes a contract vis-à-vis the affected person), the party of a declaration of will or a pre-contractual action, the promisor, or the party adopting a unilateral decision (ranking, downrating, demoting, removing). This Principle is complemented by Guiding Principle 7 that provides a specific rule for the allocation of liability.
>
> On the other hand, the operator cannot excuse itself from complying with the ADM's decision or bearing the legal consequences arising therefrom, solely on the grounds that the decision was made by automated means. The operator can also not deny that the decision can be attributed to it on the grounds that the ADM was developed by a third-party provider, or that data was collected from third-party data providers. The decision is not attributed to the programmer, the provider, the distributor or data providers. The operator is responsible for ensuring that the ADM is fit for its intended purpose and operates as it should.

Neither the autonomous nature of the process nor the malfunctioning of the ADM justifies non-attribution per se. Of course, the operator is entitled to prove that the decision was erroneous due to a system failure, inaccurate data or third-party interference, despite the operator having acted with due diligence to prevent such circumstances. In such cases, the burden of proof is on the operator. By providing evidence of the malfunctioning of the ADM, the operator is not denying that the decision can be attributed to it but is rather proving that an invalidating, excusable mistake that might render the decision null and void or annullable, or excuse liability under the applicable liability framework (as per Guiding Principle 7) has occurred.

> Illustration A. An e-recruiting programme implemented by a service company ranks applicants, shortlists eligible candidates, and finally selects the chosen candidate who automatically receives an offer of employment. The recruitment decision is attributed to the company, as it is the operator.
>
> Illustration B. A health insurance app implemented by an insurer enables the user to fill out a health questionnaire and to put forward an insurance proposal. The app assesses the eligibility conditions, calculates the premium, and accepts or rejects the insurance request. The decision to refuse the proposal or to conclude the insurance contract is attributed to the insurer operating the app. The insurer is the operator of the ADM and becomes the contracting party vis-à-vis the insured upon the acceptance of the insurance proposal.

[18] A similar principle was already enshrined in the UNCITRAL Model Law on Electronic Commerce:
Art 13: Attribution of data messages
'(1) A data message is that of the originator if it was sent by the originator itself.
(2) As between the originator and the addressee, a data message is deemed to be that of the originator if it was sent:
(a) by a person who had the authority to act on behalf of the originator in respect of that data message; or
(b) by an information system programmed by, or on behalf of, the originator to operate automatically …'.

This Guiding Principle is complemented by the risk-allocation Principle (Guiding Principle 7) that holds the operator liable. Both Principles aim to allocate legal effects and liability risks but they tackle two different legal issues. This Principle addresses the attribution of legal effects arising from the decisions stemming from the ADM; whereas the risk-allocation Principle (Guiding Principle 7) provides guidance in allocating liability for any damage caused by the operation of the ADM, including its non-functioning, the fact that no decision is adopted, or the damaging consequences of the intended operation.

Relevant contracts, or operating agreements related to the operation of the ADM, can help to identify or to confirm who the operator is by supplementing or contractually clarifying the control and benefit test as described in Guiding Principle 7.

Guiding Principle 4: Disclosure that the decision-making is automated

Unless it is obvious or unnecessary from the circumstances and the context of use or exempted by law, it shall be disclosed that the decision is being made by automated means.

Disclosing the fact that the system is automated (decision-making, facial recognition, content generation, profiling, credit scoring, etc) would allow parties to make informed decisions, minimise the manipulative or misleading effects of such a system, and enable objections to be subjected to such automated processes, where applicable. Therefore, it is particularly critical when the ADM makes a decision that can have legal effects on the rights, or the legal or contractual status of the affected person. The rationale underlying this Principle is similar to the rule requiring commercial communications (advertising) to be clearly identifiable as such,[19] so that the addressee is alerted and prepared to make informed free decisions.

The AI Act relies on transparency with such purpose for systems that interact with humans, detect emotions, or generate or manipulate content ('deep fakes') – article 52. Likewise, disclosure of the automated nature of the decision is also presumed in article 22 GDPR as a prerequisite to exercising the right to object. As, in fact, articles 13(2)(f), 14(2)(g) and 15(1)(h) GDPR explicitly include the information about the existence of ADM, including profiling (referred to in article 22 GDPR), in the list of further information that the controller shall provide the data subject with to ensure fair and transparent processing.[20]

This Guiding Principle exclusively refers to disclosing the fact that the decision is made by automated means. Beyond that, transparency and explainability of the parameters, the conditions, and the criteria an algorithm-driven system works on are policy solutions already contemplated in the GDPR and today commonly shared by the DSA, the P2B Regulation and the DMA. Thus, the information that the controller has to provide in respect of ADM, pursuant to articles 13, 14 and 15 GDPR, is not limited to the existence of ADM, but also has to include the logic behind the decision-making, as well as the significant and envisaged consequences of such automated processing for the data subject. However, the transparency obligations in the AI Act for high-risk systems go further and refer to clear instructions for use and other relevant information for users (article 13). The aim of this Guiding Principle is simply to ensure that the affected person becomes aware that they are interacting with (and possibly affected by a decision made by) an ADM.

How the information is effectively disclosed depends upon the context of use and the circumstances surrounding the operation of the ADM. Should the use of ADM be continuous and recurrent throughout

[19] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on Electronic Commerce') [2000] OJ L178 /1, Art 6(a).

[20] Same paragraph in arts 13(2)(f), 14(2)(g) and 15(1)(h): 'the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.'

the contractual relationship cycle, as in the case of algorithmic management in platforms, the information[21] is to be provided to a worker, say, in a document and in advance (at the latest on the first working day) – article 6(3) Directive on Platform Work. Equally, that information can be stated in the terms of reference of a tender, among the conditions for an award, or in a competition call. However, in other situations, the information needs to be disclosed at the moment that the affected person begins to interact with (or be affected by) the ADM if there was no prior contact between the parties or a long-term relationship that may render a previous disclosure feasible.

> Illustration A. A bank has implemented an ADM to assess the creditworthiness of credit applicants and screen eligible loan requests. The applicant has to fill out a questionnaire and submit their request via the bank app. The entire process is fully automated. The applicant is informed by way of a final decision of the refusal of the loan request or its acceptance. Only if loan conditions need to be discussed further will the applicant deal with bank staff. The applicant may not be aware that the submission and the screening process are fully automated. This process might not be sufficiently evident. Therefore, the bank has to ensure that the app alerts the applicant that the decision is made by ADM.

The duty to disclose should be linked to any ADM used by an operator, even if it does not produce the final decision but merely provides input for the decision-making process. For instance, if the credit scoring is based on automated profiling and then inputted into the ADM deciding on the approval or the rejection of the loan request, disclosure should also cover the scoring.

Insofar as the aim of this Principle is to protect the affected person from unknowingly interacting with ADM and ensuring they are aware of such process, a duty to disclose may not arise if it is obvious (or unnecessary) from the context of use or the circumstances under which the ADM operates that the decision is made by automated means. Such duty may be unnecessary if there is a previous commercial agreement establishing the use of the ADM, or the parties have a long-term relationship where the use of ADM is recurrent and the parties are aware of this. The absence of a need for disclosure should be carefully assessed in the case of standard terms and consumer transactions. A case-by-case assessment would determine how evident it should be for an average user that the decision is automatically taken. Several criteria may be taken into consideration: where the ADM is located, interface design, means of interaction between the affected persons and the ADM, types of decisions, context, transactional environment, etc. The targeted audience of a specific ADM should also be taken into account (minors, elderly people, etc).

> Illustration B. An electronic appliances store has installed a customer service robot at the main entrance of their premises. The robot processes basic complaints and settles claims at a primary stage. As the complaints are handled in situ, it is obvious for any affected person that it is interacting with an automated system and that a decision will be made by ADM.

Guiding Principle 5: Traceable decisions

ADM shall be designed and operate in a manner that enables the traceability of any decision.

The Principle of traceability is a prerequisite for the effective implementation of other Principles. Traceable decisions enable the human review of significant decisions, substantiate the statement of reasons when it has to be issued, and complement the risk-allocation rule on the operator insofar as the

[21] Art 6(2) Directive on Platform Work: 'The information referred to in paragraph 1 shall concern:
… (b) as regards automated decision-making systems:
(i) the fact that such systems are in use or are in the process of being introduced;
(ii) the categories of decisions that are taken or supported by such systems;
…'.

causes provoking malfunction, bias or failure can be investigated.

The ADM has to be designed and operated in such a way as to allow the traceability of decisions when so required. Traceability is a technical possibility but it is not an obligation for each delivered output. Otherwise, all the benefits of automation in cost reduction, time saving, and effectiveness would be reduced. Each decision is not expected to be traceable for the person affected by the ADM. The operator should be in a position to trace any decision either for internal purposes, auditing or monitoring, or upon the request of the affected person in certain circumstances.

The traceability Principle is not equivalent to the Principle of transparency. Informing affected persons of the criteria, parameters, and correlations of the ADM provides a generic image of the decision-making process, without explaining the concrete decision-making path for a specific decision. Traceability is a decision-specific exercise, upon request or under certain circumstances (ie for evidentiary purposes), that manifests itself in a concrete exploration of causes, steps, and concurring factors leading to a specific output.

> Illustration. For the purposes of minimising tax fraud, a tax authority has implemented an algorithmic process that assesses the risk of tax evasion on the basis of a set of pre-defined criteria revealing or evidencing fraudulent or suspicious behavioural patterns. After several months of functioning, it is alleged by a group of affected persons that the ADM's operation is discriminatory. The affected persons argue that the ADM systematically raises the risk of tax evasion by non-nationals. A legal action for the authority to review the process and reassess the sanctions imposed is initiated by a group of victims. The tax authority must be in a position to trace all the challenged decisions and provide evidence for the proceedings.

Rendering traceability feasible depends upon the implementation of auxiliary measures, functionalities, and procedures: logging, event record retention, event log management. The ADM has to be equipped with these features by design and, therefore, from the production stage and before being placed on the market. The producer, the provider, the distributor in the EU, or the importer are responsible for guaranteeing that the system put into circulation is duly equipped with such features. However, the operator should ensure that the ADM employed in its activity allows for the traceability of decisions. In that regard, the operator should guarantee that decisions are traceable for the user. Traceability standards and methodologies would help to develop best practices for the implementation of this Principle in industry. Following the same logic proposed by the *Report on Liability for Artificial Intelligence and Other Emerging Technologies* by the Expert Group on Liability and New Technologies – New Technologies Formation, in order to establish the liability of the producer and to entitle the operator to file a recourse claim against a producer who failed to equip a system with these required features, the operator would inform the affected person of the traceability of decisions and would have a recourse claim against the producer (provider, importer, distributor) if these logging functions were not present. [22]

[22] Expert Group on Liability and New Technologies – New Technologies Formation (n 11):
'9. Logging by design ([20]–[23])
[20] There should be a duty on producers to equip technology with means of recording information about the operation of the technology (logging by design), if such information is typically essential for establishing whether a risk of the technology materialised, and if logging is appropriate and proportionate, taking into account, in particular, the technical feasibility and the costs of logging, the availability of alternative means of gathering such information, the type and magnitude of the risks posed by the technology, and any adverse implications logging may have on the rights of others.

> [21] Logging must be done in accordance with otherwise applicable law, in particular data protection law and the rules concerning the protection of trade secrets.
>
> [22] The absence of logged information or failure to give the victim reasonable access to the information should trigger a rebuttable presumption that the condition of liability to be proven by the missing information is fulfilled.
>
> [23] If and to the extent that, as a result of the presumption under [22], the operator were obliged to compensate the damage, the operator should have a recourse claim against the producer who failed to equip the technology with logging facilities'.

Guiding Principle 6: Reasoned decisions

The complexity, the opacity or the unpredictability of ADM is not a valid ground for rendering an unreasoned, unfounded or arbitrary decision.

The complexity of the algorithms driving the ADM, the multitude of inputs and concurring factors throughout the process, or the unpredictability instilled by machine-learning/deep-learning techniques should not per se render the decision unreasoned, arbitrary, or unfounded.

The operator should ensure that any decision made or assisted by the ADM and employed for the intended purpose can be explained. Complex, opaque or unpredictable features should not be brought forward as excuses not to give reasons or to refuse to account for the decisions made by the ADM. Otherwise, the affected person is unprotected and defenceless against unfounded decisions. In the absence of reasons underlying a decision, the affected person is not in a position to assess the correctness of the decision and is deprived of the right to challenge a decision affecting them.

As mentioned above, articles 13(2)(f), 14(2)(g) and 15(1)(h) GDPR expressly require that the controller provide the data subject with 'meaningful information about the logic' involved in automated decision-making, and 'the significance and envisaged consequences of such' automated processing. Mere transparency of parameters, criteria, and factors upon which the ADM is based guarantees neither that the decision is properly reasoned nor that the data subject understands the basis upon which the decision has been made. The principle of reasoned decisions goes beyond transparency.

The DSA, in its article 15, provides for the duty of hosting services providers to provide a clear and specific statement of reasons for each decision to remove or disable access to specific pieces of information provided by recipients of the service. The statement of reasons has to at least contain the information listed in article 15(2) DSA. One of the aims of the statement of reasons is precisely to allow the affected person (recipient of the service in the DSA terminology) to effectively exercise the redress possibilities available in respect of the decision, ie internal complaint-handling mechanisms, out-of-court dispute settlement and judicial redress.

The Directive on Platform Work also stipulates, in its article 8(1), that:

> '(d)igital labour platforms shall provide the platform worker with a written statement of the reasons for any decision taken or supported by an automated decision-making system to restrict, suspend or terminate the platform worker's account, any decision to refuse the remuneration for work performed by the platform worker, any decision on the platform worker's contractual status or any decision with similar effects.'

Both legal provisions above are evidence and expressions of the principle of reasoned decisions. Their respective scopes of application reveal that those decisions likely to significantly affect a person's legal or contractual status, to impact their rights, or restrict, suspend or terminate the affected person's account – insofar as that entails the limitation or the termination of the exercise of rights – should be reasoned. Even if the decisions are taken by automated means, an explanation for them and the underlying reasoning behind them should be provided in the accompanying statement of reasons.

Not every output of an ADM process will have to be accompanied by a statement of reasons. In the case of ratings, rankings or recommendations, the mere transparency of the relevant criteria fulfils all the needs of the person who relies upon them.

The statement of reasons has to be proportionate in terms of costs and complexity for the operator, and formulated in such a way that the affected person can easily comprehend the specific decision under the given circumstances. Thus, a generic, standardised statement may not be sufficient. The statement of reasons has to be 'as precise and specific as reasonably possible' (article 15(3) DSA) and this requires that consideration be taken of the type of decision, the potentially affected rights or the complexity of the case. A decision dismissing an employee, resolving a complaint or settling a dispute should be more extensively reasoned in order to allow the affected person to protect and defend their rights, whereas a decision to remove illegal or inappropriate content can simply be explained by reference to the illicit character or the violated provision of the platform's internal policy (eg prohibition on hate speech, illegal advertising, infringement of IP rights).

> Illustration. An insurer employs an app to screen eligible insureds for health insurance. The applicants fill in a questionnaire provided by the insurer via the app. Additionally, the app tracks the behaviour of the applicants in social media and predicts future health risks on the basis of certain patterns inferred from their digital activity. A group of applicants whose requests have been refused by the app alleges that the denial of their requests is unfounded and arbitrary and ask for an explanation for the decision. The operator (the insurer) should be in a position to trace the decisions refusing the requests and to provide a statement of reasons. The variety of data collected, the inaccuracy of such information taken from social media or the fact that the app is fully automated cannot lead to arbitrary decisions or to these being put forward as excuses not to provide reasons to the affected persons.

Guiding Principle 7: Allocation of risks to the operator

The risks that the ADM may cause any harm or damage shall be allocated to the operator.

The decision made by ADM, particularly if it governs the operation of a physical device (eg a care or a surgical robot, autonomous vehicle, drone, access control machines, smart home system), can cause material damage, personal injuries or other economic losses. The autonomous vehicle may be involved in an accident, a care robot may harm the user or cause material damage at home, robotic surgery may aggravate the health of the patient undergoing the procedure, or a delivery drone may crash into a window.

Assuming that the operator is defined on the basis of the control-and-benefit binomial, the risks of such damage should be borne by the operator – not necessarily solely. As the operator controls the ADM governing the device, it is in the best position to assess, prevent and manage the risks. The operator has incentives to adopt the most effective preventive measures to minimise the risk of causing damage. Besides, as the operator benefits from the advantages of implementing and using the ADM in the course of its business or professional activity, it is reasonable that the risks inherent in the automated activity are borne by the operator. This risk-allocation Principle works coherently with Guiding Principles 3 and 5, in particular.

Guiding Principle 7 is inspired by the conceptualisation of an operator ('frontend operator'/'backend operator') proposed in the Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies23 by the Expert Group on Liability and New Technologies – New Technologies Formation, and the dual liability model (strict liability and fault-based liability) of the operator provided for by the cited EP Resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence.24 As the latter proposed, the strict liability of the operator should be the exception and be applied to high-risk ADM, whereas fault-based liability is the default regime for any ADM not listed as high risk.

The proposed liability model neither interferes with nor replaces liability for defective products[25] or the applicable contractual liability. The liability of the operator is triggered by its control of the ADM and the fact that it benefits from its operation. Additionally, the operator (specially, the backend operator) may be the producer as well. Then, provided that the damage was caused by a defect of the product, and as long as the system of liability for defective products can be applied to the ADM or to 'smart devices' operated by the ADM, the operator may also be, alternatively, liable as a producer. Besides, the operator may concurrently be the vendor of the device and/or the licensor of the ADM causing the damage. In such a case, the operator may be liable on grounds of lack of conformity or breach of contract but the basis of

[23] Expert Group on Liability and New Technologies – New Technologies Formation (n 11).

[24] European Parliament (n 8).

[25] ELI, *European Commission's Public Consultation on Civil Liability Adapting Liability Rules to the Digital Age and Artificial Intelligence, Response of the European Law Institute*, <https://europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/Public_Consultation_on_Civil_Liability.pdf> accessed 3 May 2022.

the liability, as well as the available remedies and the persons entitled to claim would be different. The risk-allocation model proposed by this Guiding Principle is exclusively based on the operation of the ADM under the parameters of control and benefit and has to be complemented by other liability regimes.

Illustration A. A basic smart home system is installed in a house to control the heating, the shutters, the sunshades, and the sprinklers in the garden. The operator – provider of the smart home system and producer of all interconnected devices – provides the ADM controlling the entire system pre-installed in the central smart home hub. The ADM is based on weather predictions and data relating to temperature, sun hours, light, and rain provided by the sensors connected to the smart home system. For unknown reasons, the ADM instructs the system to unfurl the sunshades, open the shutters, and activate the sprinkles at full power on a very rainy day. At the end of the day, the garden, the porch, and the garage are completely flooded, the sunshades collapse due to the weight of the water, and the water starts to seep through the windows of the living room and the hall. As the operator is the party who controls the entire system and benefits from selling the smart home system and licensing the software operating the system including the ADM, the risks are to be borne by the operator vis-à-vis the affected person. Once the risks are placed on the operator, the applicable liability regime will depend upon other elements. The applicable liability rules may differ (and concur) if the operator is also the producer (liability for defective products), the vendor and/or the licensor (contractual liability).

Illustration B. A university has implemented an algorithm-driven system for the delivery of mail and parcels on campus by a fleet of drones. The central system automatically classifies the received mail, assigns deliveries, and remotely activates and operates the drones up to destination. Should damage or personal injuries be caused by an accident, a collision with buildings or windows, or by a drone crashing in a garden, the university, as the operator, bears the risk, without prejudice to the liability of the producer, if damage is caused by a defect of the product.

Guiding Principle 8: No limitations to the exercise of rights and access to justice

Automation shall not prevent, limit, or render unfeasible the exercise of rights and access to justice by affected persons. An alternative human-based route to exercise rights should be available.

As a specific application of Guiding Principle 1, this Principle focuses on the risk that the exercise of rights by the affected person and effective access to justice may be prevented, hampered or limited by the inadequate use of automation. The situations that this Principle aims to avoid are grouped into two categories. First, where the affected person can only exercise a right by resorting to an algorithmic process. Second, where the affected person is deprived of the possibility of exercising a right or access to justice solely on the grounds that the contested decision was made by ADM.

In the first group of situations, the procedure required to exercise a right (the right to rectification of personal data, the right of withdrawal, the right to communicate risk-diminishing circumstances to an insurer to improve insurance conditions, the right of complaint, etc) is fully and exclusively automated. Therefore, the affected person is entirely dependent upon the correct, appropriate, and user-friendly functioning of the automated process. However, the process can be complex, poorly designed, user-unfriendly, or simply unfamiliar. Then, the exercise of a right is discouraged or prevented. The automated procedure may become an insurmountable obstacle for the affected person to exercise a right. If there is no alternative route to exercise the right, in practice, the affected person is effectively deprived of that right. In such cases, the operator should ensure that a human-based alternative is available.

Illustration A. An insurer employs an app that manages the entire insurance cycle. Notification of events, declaration of risks or changes of relevant circumstances by the insured can solely be communicated to the insurer via the app and on a fully automated basis. The insured attempts to notify the insurer of a relevant change in risk-reducing circumstances in order to benefit from an improvement of the economic conditions of the insurance policy by a corresponding reduction of the insurance premium. The app provides a form to communicate any change in the relevant circumstances. However, the form is very simple and merely offers the possibility for the insured to select one change among a list of pre-defined situations. The insured cannot communicate the change and there is no other way to contact the insurer.

In the second group of situations, the affected person has been affected by a decision taken by ADM and wishes to challenge it. The decisions made by an ADM system cannot be considered final or non-appealable solely on the grounds that they were taken by automated means. Furthermore, if the automated decision risks causing irreversible damage or irrecoverable effects (destruction of data, irreversible loss of digital content, etc), a mechanism to challenge the decision before execution should be implemented. Otherwise, the affected person would, in practice, be deprived of any possibility to challenge the decision other than claiming compensation. Article 22 GDPR, the classical centrepiece of the EU's legal approach to ADM, provides for suitable measures to safeguard the affected person's rights, freedoms and legitimate interests that stem from the right of the affected person to object, to express their point of view and to contest a decision. Therefore, the availability of suitable mechanisms for the affected person to contest, raise objections concerning, and seek an explanation about a decision is crucial. As this Guiding Principle states, an alternative human-driven route to exercise these rights to object, contest or express the affected person's views should be available. Article 22 GDPR refers to the right to 'human intervention'.

Illustration B. A platform for generating and sharing artistic digital content has developed a strict algorithm-driven content system to remove illegal content and content violating the platform's policies. The content system detects inappropriate/illegal content and removes it automatically. If the system identifies more than two infringements in one week, it immediately closes the content-publishing account and deletes all the user-generated content. Once the account of the user (affected person) is removed, there is no access to the internal complaint-handling process, dispute resolution mechanisms or the affected digital content. Thus, the user is totally defenceless, as there is no alternative way to challenge the decision other than through judicial action for restitution or compensation. Meanwhile, the user has lost all their posted digital content and has no reasonable means at their disposal to exercise their rights. The effects of closing the account and deleting all the digital content may be considered excessive and undesired if the consequence is that the user has means of challenging the decision and defending their rights before the closure and the removal are executed.

Guiding Principle 9: Human oversight/action

The operator shall ensure reasonable and proportionate human oversight over the operation of ADM taking into consideration the risks involved and the rights and legitimate interests potentially affected by the decision.

A delicate, but fundamental, balance has to be struck between the benefits associated with full automation, minimising human intervention, and the proper protection of rights and interests at stake, ensuring reasonable and proportionate human oversight.

Human oversight should not compromise the benefits in cost reduction, effectiveness, and economies of scale gained by introducing automation. However, human monitoring of the operation of the implemented ADM to assess its intended functioning and to evaluate its impact on the affected person's rights and legitimate interests, as well as on the overall socio-economic context in which the ADM is used, is vital. The positive effects of automated and increasingly autonomous systems in decision-making should not lead to uncontrollable, unsupervised ADM.

The scope, the intensity, and the extent of human oversight should be decided on the basis of the characteristics of the ADM, the potential risks involved, and the rights and interests affected thereby. Hence, human oversight should be reasonable and proportionate. The ultimate aim is to guarantee that human control is always preserved in any decision-making that may affect the rights and legitimate interests of affected persons and third parties. Implementing ADM for any intended purpose should not mean that an operator 'outsources' its business risks by placing them in 'externalised' automated systems. As any ADM operates within the operator's sphere of risk, its functioning should also fall under the operator's sphere of control.

In assessing the reasonableness and the proportionality of the required human oversight, the operator needs to calibrate costs, risks, and legal implications of each ADM. For instance, the Directive on Platform Work (article 7(2)) provides for rules on human monitoring to 'evaluate the risks of automated monitoring and decision-making systems to the safety and health of platform workers, in particular as regards possible risks of work-related accidents, psychosocial and ergonomic risks'. Reasonable human oversight is not expected to involve monitoring on a decision-by-decision basis – this is beyond dispute as, otherwise, the rationale for automating any process vanishes. Requiring human oversight should also not imply that the automatic functioning of the ADM to make a decision without human intervention is disregarded. On the contrary, automation means that certain tasks, activities, or decision-making are indeed performed by automated means without being actioned, reviewed or ratified by a natural person.[26] Hence, human oversight should be aimed at ensuring effective mechanisms, adequate resources, and fit-for-purpose procedures to carry out an overall supervision of the operation of the ADM in order to retain an element of human control in the decision-making process.

Human monitoring would enable the operator to detect malfunctions, to prevent failures, to identify unexpected outputs or deviations from the expected functioning or abnormal functioning, to discover systematic biases, or to assess the overall impact of the ADM considering its intended purpose. Thus, human oversight should be exercised as required in order for the operator to remain in control.

The frequency and the scope of human oversight or the resources allocated to it would be set in accordance with the risk assessment of each ADM. The legislator may wish to provide for specific rules on frequency, scope or human involvement in certain sectors. Otherwise, the combination of Guiding Principles 3 (attribution of legal effects) and 7 (risk allocation) should serve as effective incentives for the operator to devise an appropriate strategy to implement reasonable and proportionate human oversight, insofar as the operator assumes the legal effects of the decision taken by the ADM and wishes to mitigate the risks arising from its operation.

> Illustration A. A consulting firm implements an algorithm-driven recruiting programme to hire junior consultants. The e-recruiting system screens the curricula of the candidates, uploaded in a standardised form, rates them on the basis of pre-defined parameters, adjusted by the performance level of employees recruited in previous rounds in the past, and selects the eligible candidates to be recruited. There are at least two risks to manage. First, the e-recruiting system self-learns from the performance of current employees, so it may perpetuate 'past decisions'. Second, the curricula submitted by the candidates are strictly standardised in the application forms, with the result that the professional profiles of the eligible candidates are determined by the design and contents of the standard form. Considering the risk of bias in the selection and the relevant impact of the decision on the individual candidates and on the conditions of the labour market, regular human oversight over the e-recruiting system would be advisable to supervise the correct operation of the ADM, identify biased or discriminatory

[26] See art 12ff UN Convention on the Use of Electronic Communications in International Contracts on the *Use of automated message systems for contract formation*:

> 'A contract formed by the interaction of an automated message system and a natural person, or by the interaction of automated message systems, shall not be denied validity or enforceability on the sole ground that *no natural person reviewed or intervened in each of the individual actions carried out by the automated message systems or the resulting contract*'. Emphasis added.

choices, and assess the impact of the automated recruiting on the diversity and quality of staff.

Illustration B. A private business school employs an algorithm-driven programme to effectively assign classroom space to different courses, depending on the number of attendants, technical needs, and timetable. In this case, neither the risks nor the affected rights justify regular human oversight. Sporadically, the business school can verify if the ADM is fulfilling the expected goals, if the assignment of space is optimised, or if any error occurs in the assignment of classroom space. Human action would be available to solve unexpected false assignments (clashes of activities) but regular human oversight is neither expected nor needed.

Guiding Principle 10: Human review of significant decisions

Human review of selected significant decisions on the grounds of the relevance of the legal effects, the irreversibility of their consequences, or the seriousness of the impact on rights and legitimate interests shall be made available by the operator.

Among the most evident benefits of ADM, efficiency in adopting recurrent, mass decisions and effectiveness in executing them at reasonable cost and speed can be mentioned. Unlike human-based decision-making, ADM can optimise resourses and processes by reducing delays, minimising errors, and ensuring immediate enforcement. The decision to remove digital content, to close an infringing user account, to reduce the credit scoring of a debtor, or to dismiss a platform worker by disabling an account, taken by automated means, are enforced immediately and with full effectiveness. These benefits also hold significant risks. Should certain decisions be enforced and executed immediately, without the opportunity to challenge or review them, the resultant effects can be excessive, irreversible or irreparable.

Actions and remedies provided for by applicable law are available and, in many cases, can protect the person affected by the decision. Judicial or extra-judicial actions will entitle the affected person to challenge the decision, or claim compensation or restitution following the ordinary routes to protect and defend their rights.

This notwithstanding, certain significant decisions made by ADM, due to their relevance and the importance of their impact on the affected person, may be subject to human review upon request and prior to any other available means to challenge the decision as provided for by the law being taken. This is the rationale behind article 8(2) Directive on Platform Work[27] that entitles the affected person (the affected platform worker) to request that the operator review the decision and rectify it, if applicable. Equally, pursuant to article 17 DSA, online platforms must implement an internal complaint-handling system for users to lodge complaints against decisions to remove or disable access to information, decisions to suspend or terminate the provision of services, or decisions to suspend or terminate an account, on the ground that the user has provided illegal content or content incompatible with the platform's terms and conditions. The complaint-handling systems can be partially automated but human intervention has to be guaranteed (article 17(5) DSA).[28] Thus, the DSA provides for human review in the form of complaint-handling of certain decisions, as listed in article 17(1) DSA. However, article 22 GDPR, a pivotal provision in the EU legal framework for ADM, requires the controller to implement, as suitable measures to safeguard the affected persons' rights, freedoms, and legitimate interests, mechanisms that enable the affected person to contest the decision, provided that it is based solely on automated processing (and under the conditions of article 22(2) GDPR). Thus, the right to have human intervention and to contest the decision also converge in this Guiding Principle.

[27] Art 8 Directive on Platform Work: '… 2. Where platform workers are not satisfied with the explanation or the written statement of reasons obtained or consider that the decision referred to in paragraph 1 infringes their rights, they shall have the right to request the digital labour platform to review that decision. The digital labour platform shall respond to such request by providing the platform worker with a substantiated reply without undue delay and in any event within one week of receipt of the request. … 3. Where the decision referred to in paragraph 1 infringes the platform worker's rights, the digital labour platform shall rectify that decision without delay or, where such rectification is not possible, offer adequate compensation. …'

[28] Art 17(5) DSA: 'Online platforms shall ensure that the decisions, referred to in paragraph 4, are not solely taken on the basis of automated means'.

The mass character of decisions made by automated means, the immediacy in their execution, and the great effectiveness of the technology-enabled enforcement do not properly tally with the longer time periods and the additional costs and complexities of ordinary judicial ways to challenge a decision. This results in a 'bottleneck' in the judicial satisfaction of affected persons' demands.

The human review of certain significant decisions would be the corollary of Guiding Principles 5 and 6 and can serve as an effective complementary measure to ensure full compliance with applicable law as per Guiding Principle 1. Once the affected person is aware of the statement of reasons supporting a decision, they are in a position to assess the (un-)reasonableness of the decision and, accordingly, to challenge it immediately by requesting human review. Even if the operator has made all efforts to ensure that the ADM is law-complaint, the inherent limitations of an automated system may lead to specific decisions being made which infringe upon the rights of the affected person. Therefore, human review guarantees full compliance with applicable law without relinquishing the benefits of automation.

Human review represents an exceptional, additional safeguard complementing human oversight. But, unlike human oversight, whose operating conditions are determined by the operator as part of internal systems and controls, human review begins upon the request of the affected person. Therefore, the human review of each and every decision is neither proposed nor expected.

The operator will inform the affected persons of which decisions can be subject to human review and specify the conditions under which human review will be conducted. The operator may decide to devise a specific human-review procedure (article 8 Directive on Platform Work), or design it as a complaint-handling mechanism (article 17 DSA).

Human review mechanisms can be required by law in specific sectors or as regards certain types of decisions, and/or can be implemented by operators on a voluntary basis and, therefore, under the conditions and to the extent established by the operator.

> Illustration. A mobility-as-a-service platform monitors the performance of drivers on the basis of data provided automatically by phones installed in their cars and upon being connected to an app. Pursuant to the platform's internal policy, if a driver fails to complete at least five rides per week, the driver's account is immediately disabled and this constitutes a cause of dismissal. The decision to terminate the account and dismiss the worker should be subject to human review upon request of the driver. Justified reasons (malfunction of the phone, error, inaccurate data, failed connection, work leave) for not completing the required minimum number of rides per week might be duly put forward by the affected driver to stop or reverse the dismissal.

Guiding Principle 11: Responsible ADM

> Operators should acknowledge the potential impact of the ADM systems they employ on the socio-economic context (democratic values, fundamental rights and liberties, human dignity, social cohesion, etc), and ensure that they use ADM systems responsibly.

The recurrent, repetitive, automatic, and mass operation of ADM result in amplifying and multiplying its intended (or unintended) effects. These effects can be positive, and intended, but also negative, being either intended or unintended as a result of its abnormal functioning. The viral potential of algorithm-driven automation calls for a responsible use of ADM systems, taking into consideration the possible impact on democratic values, fundamental rights and civil liberties, market stability, environmental, social and governance goals, sustainability or social cohesion. In sum, the decision to employ and implement ADM for any intended purpose should take into account the potential impact on individual and collective rights and on the socio-economic environment.

Beyond ensuring that the implemented ADM systems comply with the law, the operator should do their utmost to use responsible ADM systems. To that end, the operator should be aware of the potential risks of the mass operation of the ADM for the socio-economic context where it impacts and mitigate such risks to the maximum extent possible.

Non-responsible ADM systems stoke social alarm by enhancing the visibility of 'controversial posts', radicalise public opinion by polarising debate, contribute to mis-/disinformation by igniting the virality of 'fake news', encourage discrimination by disregarding subtle algorithmic biases, undermine democracy with targeted advertising before elections, or interfere in human autonomy with misleading dark patterns or vulnerability-exploiting personalised recommendations. At the same time, the legislator may decide to prohibit or limit some uses of ADM in certain sectors or for specific purposes (ie prohibited AI systems under the AI Act) that are deemed 'non responsible'. Thus, this Principle, to some extent, becomes a law-compliant Principle (Guiding Principle 1).

Responsible ADM systems contribute to strengthening social cohesion, promoting diversity, facilitating fact-checking to counter 'fake news', mindful of the discriminatory effects of algorithmic bias, containing the virality potential of hate speech, and to improving personalising techniques to reduce the risk of echo chambers, or exclude from their recommender systems criteria that might be associated with vulnerabilities or target vulnerable groups.

> Illustration. A home delivery platform implements an algorithm-driven model to calculate salaries. The ADM system penalises drivers who are not able to reduce the delivery time by at least 2% every week compared to the average delivery time of all workers in the platform the previous week with a salary reduction of 30%. The number of accidents the drivers have increases dramatically as the drivers struggle to prevent a reduction in their salaries, putting their physical integrity at risk. Article 7(2) Directive on Platform Work specifically prohibits the use of automated monitoring and decision-making systems in any manner that puts undue pressure on platform workers or otherwise puts at risk their physical and mental health.

Guiding Principle 12: Risk-based approach to ADM

These Guiding Principles shall be applied on a risk-based approach.

These Guiding Principles apply to a wide variety of ADM as defined, for the purposes of this Innovation Paper, in Section II above. The proposed definition for ADM covers ratings, rankings, predictions or recommendations that the affected person may rely on and subsequently use as an input for further decision-making, and decisions that affect the affected person's rights, legal or contractual status, or legitimate interests. Given such a broad scope, the risks involved in the use and the operation of ADM largely vary depending on the type of decision, the context of use, the affected rights, or the social implications.

As noted throughout the text in the explanatory comments to each Guiding Principle, the application of these Principles has to be based on a risk approach. The intensity and the extent of the proposed Principles to be implemented by the legislator as policy goals, as well as the conditions under which the implementation has to be carried out by the operator fully depend upon the assessment of potential risks.

Insofar as automation pervades all social and economic activities, ADM systems are used in a multitude of contexts and with myriad purposes. 'One-size-fit-all' rules are unfeasible and unadvisable. Therefore, these Principles provide guidance for the formulation of rules and implementation of effective operator-driven solutions on the basis of a risk-based approach. Risks involved in the use and the operation of the ADM will be the measure to calibrate the crystallisation of the Principles in the form of legal rules and the implementation actions that the operator is expected to undertake.

ELI is an independent non-profit organisation established to initiate, conduct and facilitate research, make recommendations and provide practical guidance in the field of European legal development. Building on the wealth of diverse legal traditions, its mission is the quest for better law-making in Europe and the enhancement of European legal integration. By its endeavours, ELI seeks to contribute to the formation of a more vigorous European legal community, integrating the achievements of the various legal cultures, endorsing the value of comparative knowledge, and taking a genuinely pan-European perspective. As such, its work covers all branches of the law: substantive and procedural; private and public.

ELI

EUROPEAN
LAW
INSTITUTE

ISBN: 978-3-9505192-7-3